Kinetic Theory of Random Graphs: from Paths to Cycles

E. Ben-Naim[1,*] and P. L. Krapivsky[2,†]

[1] *Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*

[2] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*

Structural properties of evolving random graphs are investigated. Treating linking as a dynamic aggregation process, rate equations for the distribution of node to node distances (paths) and of cycles are formulated and solved analytically. At the gelation point, the typical length of paths and cycles, l, scales with the component size k as $l \sim k^{1/2}$. Dynamic and finite-size scaling laws for the behavior at and near the gelation point are obtained. Finite-size scaling laws are verified using numerical simulations.

PACS numbers: 05.20.Dd, 02.10.Ox, 64.60.-i, 89.75.Hc

I. INTRODUCTION

A random graph is a set of nodes that are randomly joined by links. When there are sufficiently many links, a connected component containing a finite fraction of all nodes, the so-called giant component, emerges. Random graphs, with varying flavors, arise naturally in statistical physics, chemical physics, combinatorics, probability theory, and computer science [1–5].

Several physical processes and algorithmic problems are essentially equivalent to random graphs. In gelation, monomers form polymers via chemical bonds until a giant polymer network, a "gel", emerges. Identifying monomers with nodes and chemical bonds with links shows that gelation is equivalent to the emergence a giant component [6–8]. A random graph is also the most natural mean-field model of percolation [9, 10]. In computer science, satisfiability, in its simplest form, maps onto a random graph [11]. Additionally, random graphs are used to model social networks [12, 13].

Random graphs have been analyzed largely using combinatorial and probabilistic methods [3–5]. An alternative statistical physics methodology is kinetic theory, or equivalently, the rate equation approach. The formation of connected components from disconnected nodes can be treated as a dynamic aggregation process [14–17]. This kinetic approach was used to derive primarily the size distribution of components [18–20].

Recently, we have shown that structural characteristics of random graphs can be analyzed using the rate equation approach [21]. In this study, we present a comprehensive treatment of paths and cycles in evolving random graphs. The rate equation approach is formulated by treating linking as a dynamic aggregation process. This approach allows an analytic calculation of the path length distribution. Since a cycle is formed when two connected nodes are linked, the path length distribution yields the cycle length distribution. More subtle statistical properties of cycles in random graphs can be calculated as well. In particular, the probability that the system contains no cycles and the size distribution of the first, second, etc. cycles are obtained analytically.

We focus on the behavior near and at the phase transition point, namely, when the gel forms. We show that the path and the cycle length distribution approach self-similar distributions near the gelation transition. At the gelation point, these distributions develop algebraic tails.

The exact results obtained for an infinite system allow us to deduce scaling laws for finite systems. Using heuristic and extreme statistics arguments, the size of the giant component at the gelation point is obtained. This size scale characterizes the size distribution of components and it leads to a number of scaling laws for the typical path size and cycle size. Extensive numerical simulations validate these scaling laws for finite systems.

The rest of the paper is organized as follows. First, the evolving random graph process is introduced (Sec. II), and then the size distribution of all components is analyzed in Sec. III. Statistical properties of paths are derived in Sec. IV and then used to obtain statistical properties of all cycles (Sec. V) and of the first cycle (Sec. VI). We conclude in Sec. VII. Finally, in an appendix, some details of contour integration used in the body of the paper are presented.

II. EVOLVING RANDOM GRAPHS

A graph is a collection of nodes joined by links. In a random graph, links are placed randomly. Random graphs may be realized in a number of ways. The links may be generated instantaneously (static graph) or sequentially (evolving graph); additionally a given pair of nodes may be connected by at most a single link (simple graph) or by multiple links (multi-graph).

We consider the following version of the random graph model. Initially, there are N disconnected nodes. Then, a pair of nodes is selected at random and a link is placed between them (Fig. 1). This linking process continues

*Electronic address: ebn@lanl.gov
†Electronic address: paulk@bu.edu

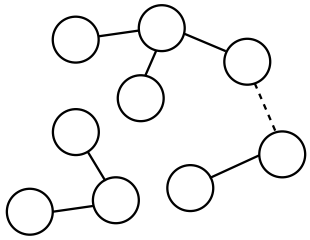

FIG. 1: An evolving random graph. Links are indicated by solid lines and the newly added link by a dashed line.

ad infinitum and it creates an evolving random graph. The process is realized dynamically. Links are generated with a constant rate in time, set equal to $(2N)^{-1}$ without loss of generality. There are no restrictions associated with the identity of the two nodes. A pair of nodes may be selected multiple times, i.e., a multi-graph is created. Additionally, the two nodes need not be different, so self-connections are allowed.

At time t, the total number of links is on average $Nt/2$, the average number of links per node (the degree) is t, and the average number of self-connections per node is $N^{-1}t/2$. Therefore, whether or not self-connections are allowed is a secondary issue. Since the linking process is completely random, the degree distribution is Poissonian with a mean equal to t.

III. COMPONENTS

The evolving random graph model has several virtues that simplify the analysis. First, the linking process is completely random as there is no memory of previous links. Second, having at hand a continuous variable (time) allows us to use continuum methods, particularly the rate equation approach. This is best demonstrated by determination of the size distribution of connected components.

As linking proceeds, connected components form. When a link is placed between two distinct components, the two components join. For example, the latest link in Fig. 1 joins two components of size $i = 2$ and $j = 4$ into a component of size $k = i + j = 6$. Generally, there are $i \times j$ ways to join disconnected components. Hence, components undergo the following aggregation process

$$(i, j) \xrightarrow{ij/2N} i + j. \tag{1}$$

Two components aggregate with a rate proportional to the product of their sizes.

A. Infinite Random Graph

Let $c_k(t)$ be the density of components containing k nodes at time t. In terms of $N_k(t)$, the total number of components with k nodes, then $c_k(t) = N_k(t)/N$. For finite random graphs, both $N_k(t)$ and $c_k(t)$ are random variables, but in the $N \to \infty$ limit the density $c_k(t)$ becomes a deterministic quantity. It evolves according to the *nonlinear* rate equation (the explicit time dependence is dropped for simplicity)

$$\frac{dc_k}{dt} = \frac{1}{2}\sum_{i+j=k}(ic_i)(jc_j) - k\,c_k. \tag{2}$$

The initial condition is $c_k(0) = \delta_{k,1}$. The gain term accounts for components generated by joining two smaller components whose sizes sum up to k. The second term on the right-hand side of Eq. (2) represents loss due to linking of components of size k to other components. The corresponding gain and loss rates follow from the aggregation rule (1).

The rate equations can be solved using a number of techniques. Throughout this investigation, we use a convenient method in which the time dependence is eliminated first. Solving the rate equations recursively yields $c_1 = e^{-t}$, $c_2 = \frac{1}{2}te^{-2t}$, $c_3 = \frac{1}{2}t^2e^{-3t}$, etc. These explicit results suggest that $c_k(t) = C_k\,t^{k-1}\,e^{-kt}$. Substituting this form into (2), we find that the coefficients C_k satisfy the recursion relation

$$(k-1)\,C_k = \frac{1}{2}\sum_{i+j=k}(iC_i)\,(jC_j) \tag{3}$$

subject to $C_1 = 1$. This recursion is solved using the generating function approach. The form of the right-hand side of Eq. (3) suggests to utilize the generating function of the sequence kC_k rather than C_k, i.e., $G(z) = \sum_k k\,C_k\,e^{kz}$. Multiplying Eq. (3) by $k\,e^{kz}$ and summing over all k, we find that the generating function satisfies the nonlinear ordinary differential equation

$$(1-G)\,\frac{dG}{dz} = G. \tag{4}$$

Integrating this equation, $z = \ln G - G + A$ and using the asymptotics $G \to e^z$ as $z \to -\infty$ fixes the constant $A = 0$. Thus, we arrive at an implicit solution for the generating function

$$G\,e^{-G} = e^z. \tag{5}$$

The coefficients C_k can be extracted from (5) via the Lagrange inversion formula, or using contour integration as detailed in Appendix A. Substituting $r = 1$ in Eq. (A1) yields $C_k = \frac{k^{k-2}}{k!}$ reproducing the well-known result for the size distribution [18, 19]

$$c_k(t) = \frac{k^{k-2}}{k!}\,t^{k-1}\,e^{-kt}. \tag{6}$$

In the following, we shall often use the generating function for the size distribution $c(z,t) = \sum_k k\,c_k(t)e^{kz}$. This generating function is readily expressed via the auxiliary generating function $G(z) = \sum_k k\,C_k\,e^{kz}$:

$$c(z,t) = t^{-1}G(z + \ln t - t). \tag{7}$$

Let us consider the fraction of nodes in finite components, $M_1 = \sum_k k\, c_k(t)$. This quantity is merely the first moment of the size distribution (hence the notation). Equivalently $M_1 = c(z=0,t)$. From (7) we find $M_1 = \tau/t$ with $\tau = G(\ln t - t)$. Using (5), we express τ through t:

$$\tau e^{-\tau} = t e^{-t}. \tag{8}$$

For $t<1$, there is a single root $\tau = t$, and all nodes reside in finite components, $M_1 = 1$. For $t > 1$ the physical root satisfies $\tau < t$ and only a fraction of the nodes resides in finite components, $M_1 < 1$. Thus, at time $t = 1$, the system undergoes a gelation transition with a finite fraction of the nodes contained in infinite components. We term this time the gelation time, $t_g = 1$. In the late stages of the evolution $t \gg 1$, one has $\tau \simeq te^{-t}$ and $M_1 \simeq c_1 = e^{-t}$, so the system consists of a single giant component and a small number of isolated nodes.

The behavior at and near the transition point are of special interest. The critical behavior of the component size distribution is echoed by other quantities as will be shown below. Size distributions become algebraic near the critical point. Moreover, there is a self-similar behavior as a function of time (dynamical scaling) and as a function of the system size (finite-size scaling).

At the gelation point, the component size distribution has an algebraic large-size tail, obtained using the Stirling formula,

$$\boldsymbol{c}_k \simeq C\, k^{-5/2}. \tag{9}$$

with $C = (2\pi)^{-1/2}$. [Throughout this paper, bold letters are used for critical distributions, so $\boldsymbol{c}_k \equiv c_k(t=1)$.] In the vicinity of the gelation time, the size distribution is self-similar, $c_k(t) \to (1-t)^5 \Phi_c\big(k(1-t)^2\big)$ with the scaling function

$$\Phi_c(\xi) = (2\pi)^{-1/2} \xi^{-5/2} \exp(-\xi/2). \tag{10}$$

Thus, the characteristic component size diverges near the gelation point, $k \sim (1-t)^{-2}$.

B. Finite Random Graphs

In the previous subsection, we applied kinetic theory to an infinite system. This approach can be extended to finite systems. Unfortunately, such treatments are very cumbersome [22, 23]. Since the number of components is finite, the fluctuations are no longer negligible, and instead of a deterministic rate equation approach, a stochastic approach is needed. Here we follow an alternative path, employing the exact infinite system results in conjunction with scaling and extreme statistics arguments.

The characteristic size of components at the gelation point exhibits nontrivial dependence on the system size. This is conveniently seen via the cumulative size distribution. The size of the largest component in the system, k_g, is estimated from the extreme statistics criterion, $N \sum_{k \geq k_g} \boldsymbol{c}_k \sim 1$, to be

$$k_g \sim N^{2/3}. \tag{11}$$

The largest component in the system grows sub-linearly with the system size [3]. The time by which this component emerges approaches unity for large enough systems as follows from the diverging characteristic size scale $k_g \sim (1-t_g)^{-2}$,

$$1 - t_g \sim N^{-1/3}. \tag{12}$$

The maximal component size (11) underlies the entire size distribution. Let $c_k(N,t)$ be the size distribution in a system of size N at time t. At the gelation point, the size distribution $\boldsymbol{c}_k(N) \equiv c_k(N, t=1)$ obeys the finite-size scaling form (Figs. 2 and 3)

$$\boldsymbol{c}_k(N) \sim N^{-5/3} \Psi_c(kN^{-2/3}). \tag{13}$$

The scaling function has the following extremal behaviors

$$\Psi_c(\xi) \simeq \begin{cases} (2\pi)^{-1/2}\xi^{-5/2} & \xi \ll 1; \\ \exp(-\xi^\gamma) & \xi \gg 1. \end{cases} \tag{14}$$

The small-ξ behavior corresponds to sizes well below the characteristic size and thus reflects the infinite system behavior (9). The large-ξ behavior was obtained numerically with $\gamma \cong 3$. To appreciate the large-ξ asymptotic, let us estimate the probability that the system managed to generate the largest possible component of size $N/2$ at time $t = 1$. The lower bound for this probability can be established via a "greedy" evolution which assumes that after k linking events the graph is composed of a tree of size $k+1$ and $N-k-1$ disconnected nodes. Such evolution occurs with probability

$$\frac{2}{N}\cdot\frac{N-2}{N} \times \frac{3}{N}\cdot\frac{N-3}{N} \times \ldots \times \frac{N-N/2}{N}\cdot\frac{N/2}{N} \sim \frac{N!}{N^N},$$

that scales as e^{-N}. While this lower bound is not necessarily optimal, it suggests that the actual probability is exponentially small. The scaling variable $\xi = kN^{-2/3}$ becomes $\xi \sim N^{1/3}$ for $k = N/2$, so $\exp(-N^{\gamma/3})$ matches the probability $\exp(-N)$ when $\gamma = 3$.

To check the critical behavior in finite systems, we performed numerical simulations. In the simulations, $N/2$ links are placed randomly and sequentially among the N nodes as follows. A node is drawn randomly, and then another node is drawn randomly. Last, these two nodes are linked. Self-connections are therefore allowed. The simulations differ slightly from the above random graph model in that the number of links is not a stochastic variable. For large N, this simulation is faithful to the evolving random graph model because the number of links is self-averaging.

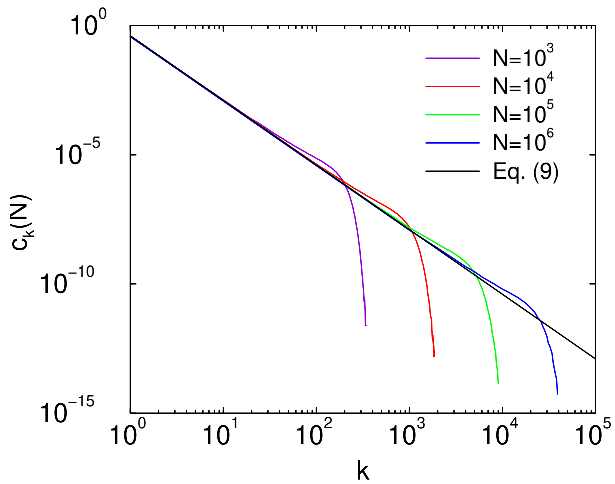

FIG. 2: The size distribution for a finite system at the gelation point. Shown is $\boldsymbol{c}_k(N)$ versus k for various N. The infinite system behavior is shown for reference. The data represents an average over 10^6 independent realizations.

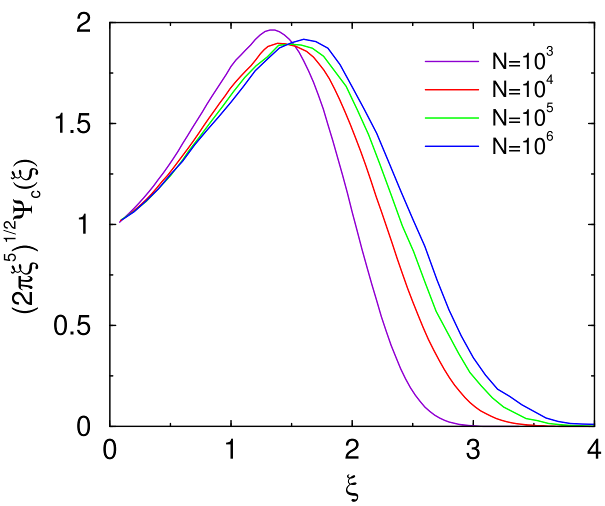

FIG. 3: Finite-size scaling of the size distribution. Shown is $(2\pi\xi^5)^{1/2}\Psi_c(\xi)$ versus ξ, obtained from simulations with various N.

The simulation results are consistent with the postulated finite-size scaling form (13). We note that the scaling function $\Psi_c(\xi)$ converges slowly as a function of N. The simulations reveal an interesting behavior of the finite-size scaling function. The function $\boldsymbol{c}_k(N)$ has a "shoulder" — a non-monotonic behavior compared with the pure algebraic behavior (9) characterizing infinite systems (Fig. 2). The properly normalized scaling function $(2\pi\xi^5)^{1/2}\Psi_c(\xi)$ is a non-monotonic function of ξ (Fig. 3). Obtaining the full functional form of the scaling function $\Psi_c(\xi)$ remains a challenge. A very similar shoulder has been observed for the degree distribution of finite random networks generated by preferential attachment [24–27].

IV. PATHS

Structural characteristics of components can be investigated in a similar fashion. By definition, every two nodes in a component are connected. In other words, there is a *path* consisting of adjacent links between two such nodes. We investigate statistical properties of paths in components. Characterization of paths yields useful information regarding the connectivity of components as well as internal structures such as cycles.

For every node in the graph, there are (generally) multiple paths that connect it with all other nodes in the respective component. With new links, new paths are formed. For every pair of paths of lengths n and m originating at two separate nodes, a new path is formed as follows

$$n, m \longrightarrow n + m + 1. \tag{15}$$

In Fig. 1, linking two paths of respective lengths $n = 1$ and $m = 2$ generates a path of length $n + m + 1 = 4$. Thus, paths also undergo an aggregation process. However, this aggregation process is simpler than (1) because the aggregation rate is independent of the path length.

Let $q_l(t)$ be the density of *distinct* paths containing l links at time t. By distinct we mean that the two paths connecting two nodes are counted separately. By definition, $q_0(t) = 1$. The rest of the densities grow according to the rate equation

$$\frac{dq_l}{dt} = \sum_{n+m=l-1} q_n q_m \tag{16}$$

for $l > 0$. The initial condition is $q_l(0) = \delta_{l,0}$. This rate equation reflects the uniform aggregation rate. Another notable feature is the lack of a loss term — once a path is created, it remains forever. Solving recursively gives $q_1 = t$, $q_2 = t^2$, etc. By induction, the path length density is

$$q_l(t) = t^l. \tag{17}$$

Indeed, this expression satisfies both the rate equation and the initial condition. The first quantity $q_1 = t$ is consistent with the facts that the link density is equal to $t/2$ and that every link corresponds to two distinct paths of length one.

The above path density represents an aggregate over all nodes and all components. Characterization of path statistics in a component of a given size is achieved via $p_{l,k}$, the density of paths of length l in components of size k. Note the obvious length bounds $0 \le l \le k-1$ and the sum rule $\sum_l p_{l,k} = k^2 c_k$ reflecting that there are k^2 distinct paths in a component of size k (every pair of nodes is connected). The density of the linkless paths is $p_{0,k} = kc_k$, because kc_k is the probability that a node belongs to a component of size k.

We have seen that components and paths form via the aggregation processes (1) and (15), respectively.

The joint distribution $p_{l,k}$ therefore undergoes a bi-aggregation process [28]. In the present case,

$$(n,i)+(m,j) \longrightarrow (n+m+1, i+j) \tag{18}$$

where the first index corresponds to the path length and the second to the component size. The joint distribution evolves according to the rate equation

$$\frac{dp_{l,k}}{dt} = \sum_{\substack{i+j=k\\ n+m=l-1}} p_{n,i}p_{m,j} + \sum_{i+j=k} (ip_{l,i})(jc_j) - kp_{l,k}. \tag{19}$$

The initial conditions are $p_{l,k}(0) = \delta_{k,1}\delta_{l,0}$. The first term on the right-hand side of Eq. (19) describes newly formed paths due to linking. The last two terms correspond to paths that do not contain the newly placed link.

We now repeat the steps used to determine the size distribution. The time dependence is eliminated using the ansatz $p_{l,k} = P_{l,k}t^{k-1}e^{-kt}$. The corresponding coefficients $P_{l,k}$ satisfy the recursion

$$(k-1)P_{l,k} = \sum_{\substack{i+j=k\\ n+m=l-1}} P_{n,i}P_{m,j} + \sum_{i+j=k} (iP_{l,i})(jC_j). \tag{20}$$

The generating function $P_l(z) = \sum_k P_{l,k}e^{kz}$ satisfies the recursion relation $(1-G)\frac{dP_l}{dz} = \sum_{n+m=l-1} P_nP_m + P_l$ for $l>0$. Dividing this equation by (4) yields

$$G\,\frac{dP_l}{dG} = \sum_{n+m=l-1} P_nP_m + P_l \tag{21}$$

for $l>0$. As noted above $P_{0,k} = kC_k$, so $P_0(z) = G(z)$. Solving Eq. (21) recursively gives $P_1 = G^2$, $P_2 = G^3$, etc. In general,

$$P_l(z) = G^{l+1}(z). \tag{22}$$

This solution can be validated directly. The time dependent generating function $p_l(z) = \sum_k p_{l,k}e^{kz}$ is therefore $p_l(z) = t^{-1}G^{l+1}(z+\ln t - t)$. The total density of paths of length l, $p_l(z=0) = t^l$, coincides with (17) prior to the gelation transition ($t<1$) because all components are finite. However, the total number of paths is reduced, $p_l(z=0) = t^{-1}\tau^{l+1}$, past the gelation time ($t>1$).

One may also obtain the bivariate generating function $p(z,w) = \sum_{l,k} p_{l,k}w^le^{kz}$. Using (22) one gets

$$p(z,w) = t^{-1}\frac{G(z+\ln t - t)}{1 - wG(z+\ln t - t)}. \tag{23}$$

The total density of paths in finite components is of course $g = \sum_{l,k} p_{l,k}$, so $g \equiv p(z=0, w=1)$. Generally, $g = \frac{\tau}{t(1-\tau)}$; for $t<1$ the total density of paths is $g(t) = (1-t)^{-1}$.

The coefficients are found via the contour integration $P_{l,k} = (2\pi i)^{-1}\oint dy\, P_l\, y^{-k-1}$ (see Appendix A). Substituting $r = l+1$ in Eq. (A1) yields $P_{l,k} = (l+1)\,\frac{k^{k-l-2}}{(k-l-1)!}$. As a result, the density of paths of length l in components of size k is

$$p_{l,k} = (l+1)\,\frac{k^{k-l-2}}{(k-l-1)!}\,t^{k-1}\,e^{-kt}. \tag{24}$$

Comparing (24) and (6) we notice that the densities of the two shortest paths satisfy $p_{0,k} = kc_k$ and $p_{1,k} = 2(k-1)c_k$. The latter reflects that there are $k-1$ links in a tree of size k and that with unit probability all components are trees (as discussed in the next section).

Note also that the longest possible path, $l = k-1$, corresponds to linear (chain-like) components. According to Eq. (24), the density of such paths is $p_{k-1,k} = t^{k-1}e^{-kt}$. This density decays exponentially with length, so these components are typically small, their length being of the order one.

The path length density can be simplified in the large k-limit by considering the properly normalized ratio of factorials

$$\begin{aligned}\frac{k!}{k^l\,(k-l)!} &= \prod_{j=1}^{l-1}\left(1-\frac{j}{k}\right)\\ &= \exp\left(-\sum_{j=1}^{l}\frac{j}{k} + \frac{1}{2}\sum_{j=1}^{l}\frac{j^2}{k^2} - \ldots\right)\\ &\simeq \exp(-l^2/2k)\,.\end{aligned}$$

Using the Stirling formula, in the limits $k \gg 1$ and $l \gg 1$, the path density becomes

$$p_{l,k} \simeq l\,(2\pi k^3)^{-1/2}\,t^{k-1}\,e^{k(1-t)}\,e^{-l^2/2k}. \tag{25}$$

As was the case for the component size distribution, the path length density is self-similar in the vicinity of the gelation point, $p_{l,k} \to (1-t)^2\Phi_p\left(k(1-t)^2, l(1-t)\right)$, with the scaling function

$$\Phi_p(\xi,\eta) = \eta\,(2\pi\xi^3)^{-1/2}\exp(-\eta^2/2\xi). \tag{26}$$

Thus, the characteristic path length diverges near the gelation point, $l \sim (1-t)^{-1}$.

At the critical point, the path length density becomes

$$\boldsymbol{p}_{l,k} \simeq l(2\pi k^3)^{-1/2}\,\exp(-l^2/2k). \tag{27}$$

It is evident that the typical path length scales as square root of the component size

$$l \sim k^{1/2}. \tag{28}$$

For finite systems, the scaling law for the typical path length (28) combined with the characteristic component size (11) leads to the following characteristic path length

$$l \sim N^{1/3}. \tag{29}$$

One can deduce several other scaling laws and finite-size scaling functions underlying the path density. For example, substituting the gelation time $1-t_g \sim N^{-1/3}$ into the total number of paths $g = (1-t)^{-1}$ yields $g \sim N^{1/3}$.

V. CYCLES

Each component has a certain number of nodes and links. The complexity of a component is defined as the number of links minus the number of nodes. Components with complexity -1 are trees; components with complexity 0 and 1 are termed unicyclic and bicyclic correspondingly. Finite components are predominantly trees. We have seen that the overall number of links is proportional to N and that the overall the number of self-links is of the order unity. The overall numbers of trees and of unicyclic components mirror this behavior. Generally, the number of components of complexity R is proportional to N^{-R} (this result is well-known, see e.g. [5, 21] and especially [29]). Therefore, it suffices to characterize trees and unicyclic components only.

Each unicyclic component contains a single cycle. Cycles are an important characteristic of a graph [30, 31]. In this section, we analyze cycles and unicyclic components using the rate equation approach. We first note that cycles in random graphs were also studied using various other approaches: Janson [32, 33] employs probabilistic and combinatorial techniques; Marinari and Monasson [31] assign an Ising spin to each node and deduce certain properties of loops from the partition function of the Ising model; Burda *et al* [34] modify a random graph model to favor the creation of short cycles, and examine the model using a diagrammatic technique. A number of authors also studied cycles on information networks like the Internet (see [35] and references therein).

A. Infinite System

There is a significant difference between the distribution of trees and unicyclic components. In the thermodynamic limit, the number of trees is extensive and as a result, it is a deterministic, or a self-averaging quantity. The number of unicyclic components is not extensive, but rather of the order unity; as a result it is a random quantity with a nontrivial distribution even for infinite random graphs. In what follows, we study the *average* number of unicyclic components of a given size or cycle length.

The average number of cycles follows directly from the path length density. Quite simply, when the two extremal nodes in a path are linked, a cycle is born. Let the number of cycles of size l at time t be $w_l(t)$. It grows according to the rate equation

$$\frac{dw_l}{dt} = \frac{1}{2}\, q_{l-1}. \tag{30}$$

The right-hand side equals the link creation rate $1/(2N)$ times the total number of paths Nq_{l-1}; indeed, the total number of cycles of a given length is of the order one. The cycle length distribution is

$$w_l = \frac{t^l}{2l}. \tag{31}$$

In particular, at the gelation point, the cycle length distribution is inversely proportional to the cycle length [5]

$$\boldsymbol{w}_l = (2l)^{-1}. \tag{32}$$

This result can alternatively be obtained using combinatorics.

To characterize cycles in a given component size, we consider the joint distribution $u_{l,k}$, the average number of unicyclic components of size k containing a cycle of length l with $1 \leq l \leq k$. This joint distribution evolves according to the *linear* rate equation

$$\frac{du_{l,k}}{dt} = \frac{1}{2} p_{l-1,k} + \sum_{i+j=k} (iu_{l,i})\,(jc_j) - k\, u_{l,k} \tag{33}$$

for $l \geq 1$. Initially there are no cycles, and therefore $u_{l,k}(0) = 0$. Eliminating the time dependence via the substitution $u_{l,k} = U_{l,k} t^k e^{-kt}$, the coefficients satisfy the recursion

$$k\, U_{l,k} = \frac{1}{2}\, P_{l-1,k} + \sum_{i+j=k} (iU_{l,i})\,(jC_j). \tag{34}$$

Using the generating function $U_l(z) = \sum_k e^{kz} U_{l,k}$ this recursion is recast into the differential equation $(1-G)\,\frac{dU_l}{dz} = \frac{1}{2}\, P_{l-1}$. Dividing by (4), we obtain

$$\frac{dU_l}{dG} = \frac{1}{2} G^{l-1}. \tag{35}$$

Integrating this equation yields the generating function

$$U_l(z) = \frac{1}{2l}\, G^l(z). \tag{36}$$

Consequently, the cycle length distribution (in finite components only) is $p_l = \frac{\tau^l}{2l}$, in agreement with (31) prior to the gelation time ($t < 1$).

Additionally, the joint generating function defined as $u(z,w) = \sum_{l,k} e^{kz}\, w^l\, u_{l,k}$ is given by

$$u(z,w) = \frac{1}{2} \ln \frac{1}{1 - wG(z + \ln t - t)}. \tag{37}$$

As for paths, statistics of cycles are directly coupled to statistics of components via the generating function $G(z)$. The total number of unicyclic components of finite-size $h = \sum_{l,k} u_{l,k}$ is therefore

$$h(t) = \frac{1}{2}\, \ln \frac{1}{1-\tau}. \tag{38}$$

Below the gelation point, $h(t) = \frac{1}{2} \ln \frac{1}{1-t}$, for $t < 1$. The total number of unicyclic components can alternatively be obtained by noting that (i) it satisfies the rate equation $dh/dt = \frac{1}{2} \sum_k k^2 c_k = \frac{1}{2}\, M_2$, and (ii) the second moment of the size distribution is $M_2 = (1-t)^{-1}$ for $t < 1$ as follows from (7).

The coefficients underlying the cycle distribution are found using contour integration. Indeed, writing

$U_{l,k} = (2\pi i)^{-1} \oint U_l \, y^{-k-1} \, dy$ and substituting $r = l$ in (A1) gives $U_{l,k} = \frac{1}{2} \frac{k^{k-l-1}}{(k-l)!}$ [4]. The cycle length-size distribution is therefore

$$u_{l,k}(t) = \frac{1}{2} \frac{k^{k-l-1}}{(k-l)!} t^k e^{-kt}. \tag{39}$$

The smallest cycle, $l = 1$, is a self-connection, and the average number of such cycles is $u_{1,k} = \frac{t}{2} \, k c_k$. The largest cycles are rings, $l = k$, and their total number is on average $u_{k,k} = \frac{1}{2k} \, t^k \, e^{-kt}$.

The large-k behavior of the cycle length distribution is found following the same steps leading to (25)

$$u_{l,k}(t) \simeq (8\pi k^3)^{-1/2} \, t^k \, e^{k(1-t)} \, e^{-l^2/2k}. \tag{40}$$

This distribution is self-similar in the vicinity of the gelation transition, $u_{l,k}(t) \to (1-t)^3 \Phi_u \left(k(1-t)^2, l(1-t)\right)$, with the scaling function

$$\Phi_u(\xi, \eta) = (8\pi\xi^3)^{-1/2} \exp(-\eta^2/2\xi). \tag{41}$$

We see that the cycle length is characterized by the same scale as the path length, $l \sim (1-t)^{-1}$. At the gelation point, the distribution is

$$\boldsymbol{u}_{l,k} \simeq (8\pi k^3)^{-1/2} \exp(-l^2/2k). \tag{42}$$

Fixing the component size, the typical cycle length behaves as the typical path length, $l \sim k^{1/2}$.

The size distribution of unicyclic components is found from the joint distribution $v_k = \sum_l u_{l,k}$. Using (39) we get [21]

$$v_k(t) = \frac{1}{2} \left(\sum_{n=0}^{k-1} \frac{k^{n-1}}{n!} \right) t^k \, e^{-kt}. \tag{43}$$

This distribution can alternatively be derived from the *linear* rate equation

$$\frac{dv_k}{dt} = \frac{1}{2} k^2 c_k + \sum_{i+j=k} (iv_i)(jc_j) - k \, v_k. \tag{44}$$

This equation is obtained from (33) using the equality $k^2 c_k = \sum_l p_{l,k}$. It reflects that linking a pair of nodes in a component generates a unicyclic component. Integrating (42) over the cycle length, the critical size distribution of unicyclic components has an algebraic tail

$$\boldsymbol{v}_k \simeq (4k)^{-1}. \tag{45}$$

B. Finite Systems

We turn now to finite systems, restricting our attention to the gelation point. The total number of unicyclic components is obtained by estimating $h(N, t_g)$. Substituting (12) into (38) shows that the average number of

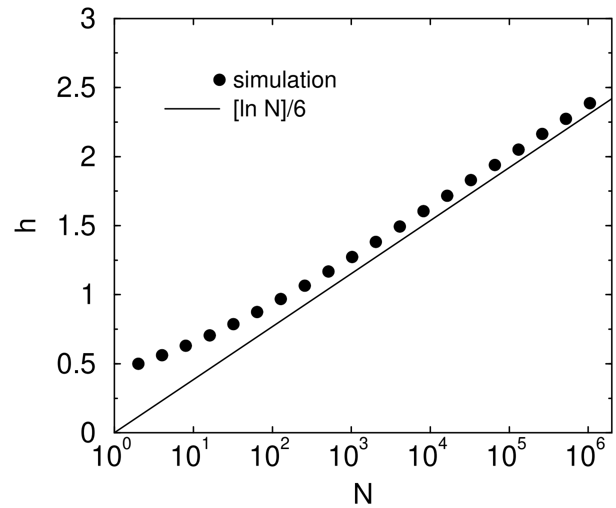

FIG. 4: The total number of unicyclic components versus the system size at the gelation point. Shown is h versus N. Each data point represents an average over 10^6 independent realizations.

unicyclic components (and hence, cycles) grows logarithmically with the system size (Fig. 4)

$$h(N) \simeq \frac{1}{6} \ln N. \tag{46}$$

Comparing the path length distribution (27) and the cycle length distribution (42), we conclude that the characteristic cycle length and the characteristic path length obey the same scaling law, $l \sim N^{1/3}$. This implies that the cycle length distribution in a finite system of size N, $w_l(N)$, obeys the finite-size scaling law

$$\boldsymbol{w}_l(N) \sim N^{-1/3} \Psi_w \left(l N^{-1/3} \right). \tag{47}$$

Numerical simulations confirm this behavior (Fig. 5).

In the simulations, analysis of cycle statistics requires us to keep track of all links. Cycles are conveniently identified using the standard "shaving" algorithm. Dangling links, i.e., links involving a single-link node are removed from the system sequentially. The link removal procedure is carried until no dangling links remain. At this stage, the system contains no trees. Simple cycles are those components with an equal number of links and nodes.

The extremal behaviors of the finite-size scaling function are as follows

$$\Psi_w(\eta) \simeq \begin{cases} (2\eta)^{-1} & \eta \to 0, \\ \exp(-C \, \eta^{3/2}) & \eta \to \infty. \end{cases} \tag{48}$$

The small-η behavior follows from (32). Statistics of extremely large cycles can be understood by considering the largest possible cycles. When there are $n = N/2$ links, the largest possible cycle has length $l = N/2$. Its likelihood $w(n, 2n)$ is obtained using combinatorics

$$w(n, 2n) = \binom{2n}{n} \times \frac{n!}{2n} \times (2n)^{-n}. \tag{49}$$

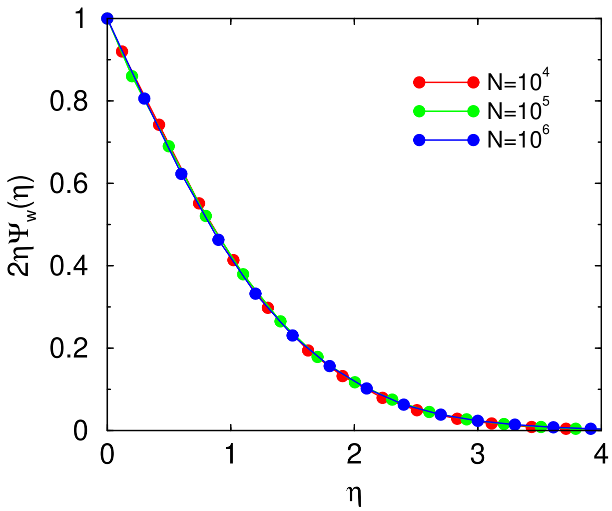

FIG. 5: Finite-size scaling of the cycle-length distribution. Shown is $2\eta\Psi_w(\eta)$ versus η obtained using systems with size $N = 10^4$, 10^5, and 10^6. The data represents an average over 10^6 independent realizations.

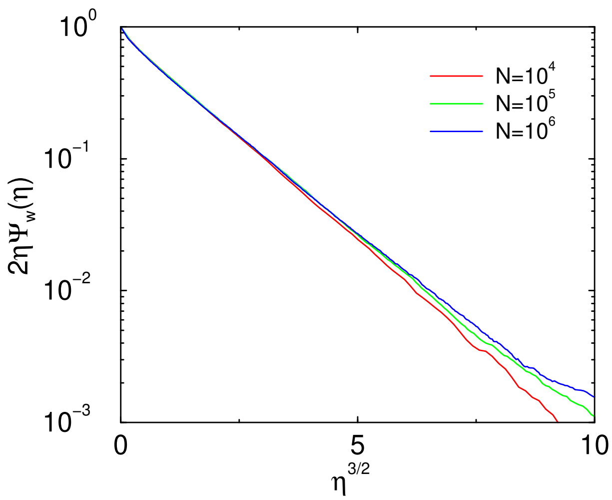

FIG. 6: The tail of the scaling function. Shown is $2\eta\Psi_w(\eta)$ versus $\eta^{3/2}$.

There are $\binom{2n}{n}$ ways to choose the nodes participating in the cycle and the next term is the number of ways to arrange them in a cycle. The corrective factor $2n$ accounts for rotation and reflection symmetries. The last term is the probability that each pair of consecutive nodes are linked. The large-n asymptotic behavior is

$$w(n, 2n) \simeq \frac{1}{\sqrt{2}n}\left(\frac{2}{e}\right)^n. \tag{50}$$

Therefore, $w(n, 2n) \sim \exp(-C\,N)$. Substituting $l \sim N$ into the scaling form (47) leads to the super-exponential behavior $\Psi_w(\eta) \sim \exp(-C\eta^{3/2})$, see Fig. 6.

Typically, cycles are of size $N^{1/3}$. The average moments $\langle l(N)\rangle = \sum_l l\, w_l(N)/\sum_l w_l(N)$ reflect this law.

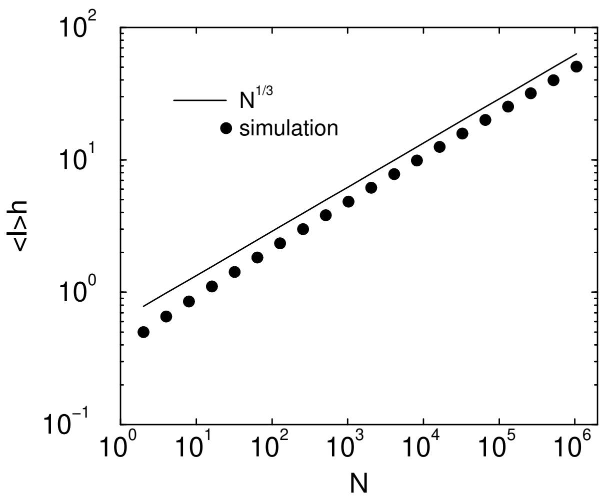

FIG. 7: The average cycle size at the gelation point. Shown is $\langle l(N)\rangle h(N)$ versus N. Each data point represents an average over 10^6 independent realizations.

However, the algebraic divergence, $w_l \sim l^{-1}$, leads to a logarithmic correction as follows from (46)–(48):

$$\langle l^n(N)\rangle \sim N^{n/3}[\ln N]^{-1}. \tag{51}$$

The behavior of the average cycle length is verified numerically (Fig. 7).

Finite-size scaling of other cycle statistics such as the joint distribution can be constructed following the same procedure. For example, the size distribution of unicyclic components should follow the scaling form

$$v_k(N) \sim N^{-2/3}\Psi_v\left(kN^{-2/3}\right). \tag{52}$$

The scaling function diverges $\Psi_v(\xi) \simeq (4\xi)^{-1}$ for $\xi \to 0$.

VI. THE FIRST CYCLE

The above statistical analysis of cycles characterizes the average behavior but not necessarily the typical one because the number of cycles is a fluctuating quantity. There are numerous interesting features concerning cycles that are not captured by the average number of cycles. For instance, what is the probability that the system does not contain a cycle up to time t? It suffices to answer this question in the pre-gel regime as the giant component certainly contains cycles.

Let $s_0(t)$ be the (survival) probability that the system does not contain a cycle at time t. The cycle production rate is $J = \frac{dh}{dt} = \frac{1}{2(1-t)}$. The number of cycles is finite in the pre-gel regime, since cycles are independent of each other in the $N \to \infty$ limit. This assertion (supported by numerical simulations, see Fig. 8) implies that the cycle production process is completely random. The cycle production rate characterizes the survival probability s_0 as

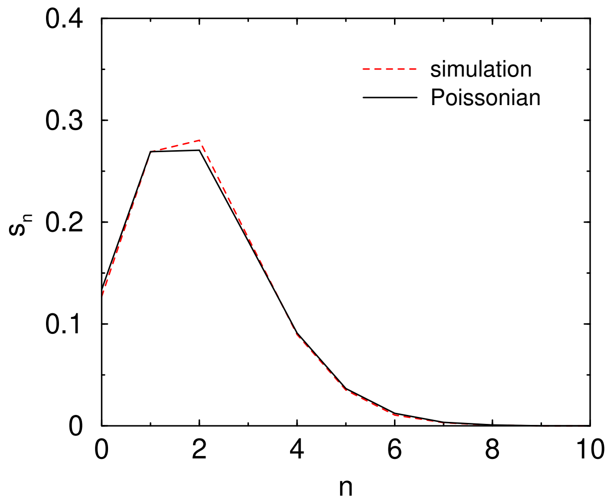

FIG. 8: The distribution of the number of cycles. Shown is s_n versus n at the gelation point. The system size is $N = 10^5$ and an average over 10^5 realizations has been performed. A Poissonian distribution with an identical average is also shown for reference.

follows

$$\frac{ds_0}{dt} = -J s_0. \tag{53}$$

The initial condition is $s_0(0) = 1$. As a result, the survival probability is

$$s_0(t) = (1-t)^{1/2} \tag{54}$$

for $t \leq 1$. The survival probability vanishes beyond the gelation point, $s_0(t) = 0$ for $t > 1$. This reiterates that in the thermodynamic limit, a cycle is certain to form prior to the gelation transition [5].

Since the number of cycles produced is of the order of one in the pre-gel regime, one may expect that statistical properties of cycles strongly depend on their generation number or alternatively on their creation time. This is manifested by the first cycle. The quantity $dt\, s_0\, \frac{dw_l}{dt}$ is the probability that (i) the system contains no cycles at time t, (ii) a cycles is produced during the time interval $(t, t+dt)$, and (iii) its length is l. Summing these probabilities gives the probability that the first cycle produced sometimes during the pre-gel regime has length l:

$$f_l = \int_0^1 dt\, s_0\, \frac{dw_l}{dt} = \frac{1}{2}\int_0^1 dt\,(1-t)^{1/2}\, t^{l-1}. \tag{55}$$

Summing these quantities, we verify the normalization

$$\sum_{l\geq 1} f_l = \frac{1}{2}\int_0^1 dt\,(1-t)^{-1/2} = 1.$$

The length distribution of the first cycle can be expressed in terms of the beta function $f_l = \frac{1}{2}B(3/2, l)$ or alternatively

$$f_l = \frac{\sqrt{\pi}}{4}\,\frac{\Gamma(l)}{\Gamma(l+3/2)}. \tag{56}$$

The probability distribution f_l has an algebraic tail,

$$f_l \simeq C\, l^{-3/2}, \tag{57}$$

with $C = \frac{\sqrt{\pi}}{4}$ for $l \gg 1$. The tail exponent characterizing the distribution of the first cycle is larger compared with the exponent characterizing all cycles, reflecting the fact that the first cycle is created earlier.

Similarly, one can obtain additional properties of the first cycle. We mention the probability F_k that the first unicyclic component has size k,

$$F_k = \int_0^1 dt\, s_0\, \frac{1}{2}\, k^2 c_k = \frac{1}{2}\,\frac{k^k}{k!}\, I_k \tag{58}$$

with the integral $I_k = \int_0^1 dt (1-t)^{1/2} t^{k-1} e^{-kt}$. This integral can be expressed in terms of the confluent hypergeometric function. Its asymptotic behavior can be readily found by noting that the integrand has a sharp maximum in the region $1-t \sim k^{-1/2}$ leading to $I_k \simeq 2^{-1/4}\Gamma(3/4) k^{-3/4} e^{-k}$. Using this in conjunction with the Stirling's formula, the size distribution has the algebraic tail

$$F_k \simeq C\, k^{-5/4} \tag{59}$$

with $C = 2^{-7/4}\pi^{-1/2}\Gamma(3/4)$ for $k \gg 1$.

Under the assumption that cycle production is completely random, the number of cycles obeys Poisson statistics. The probability that there are n cycles, s_n, then satisfies the straightforward generalization of Eq. (53), viz. $\frac{ds_n}{dt} = J[s_{n-1} - s_n]$ with the initial condition $s_n(0) = \delta_{n,0}$. The solution is the Poisson distribution $s_n = \frac{h^n}{n!} e^{-h}$, see Fig. 8. Explicitly, the distribution reads

$$s_n = \frac{(1-t)^{1/2}}{n!}\left[\frac{1}{2}\ln\frac{1}{1-t}\right]^n. \tag{60}$$

The cumulative distribution $S_n(t) = s_0(t) + \ldots + s_n(t)$ is plotted in Fig. 9.

The Poisson distribution (60) can also be used to calculate $f_{n,l}$ the size distribution of the nth cycle. We merely quote the large-l tail behavior

$$f_{n,l} \sim \frac{1}{(n-1)!}\, l^{-3/2}\left[\frac{1}{2}\ln l\right]^{n-1} \tag{61}$$

Indeed, summation over the cycle generation reproduces the overall cycle distribution (32).

In finite systems, it is possible that no cycle are created at the gelation time. This probability decreases algebraically with the system size, as seen by substituting (12) into (54)

$$s_0 \sim N^{-1/6}. \tag{62}$$

This prediction agrees with simulations, see Fig. 10. In practice, this slow decay indicates that a relatively large

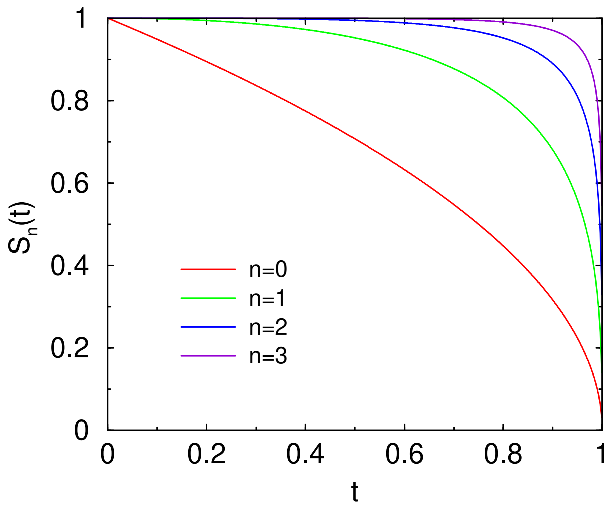

FIG. 9: The cumulative distribution $S_n(t) = \sum_{0\leq j\leq n} s_j(t)$ versus t for $n = 0, 1, 2, 3$.

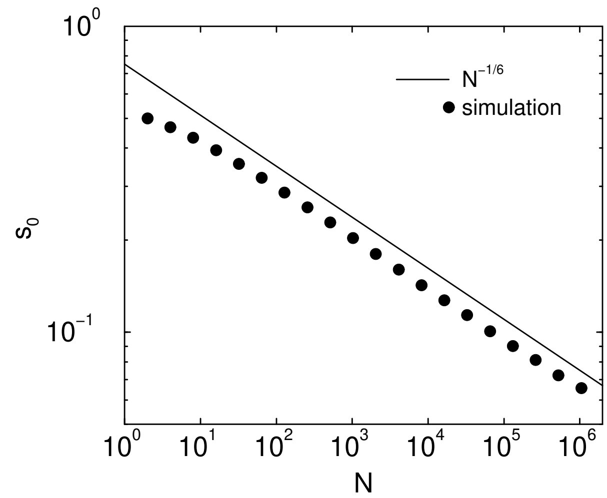

FIG. 10: The survival probability versus the system size. Shown is $s_0(N)$ versus N at the gelation point, i.e., when $N/2$ links are placed. Each data point represents an average over 10^6 realizations.

system may contain no cycles after $N/2$ links are placed. Generally, the probability that there is a finite number of cycles increases with the number of cycles

$$s_n \sim \frac{1}{n!}\,N^{-1/6}\left[\frac{1}{6}\ln N\right]^n. \tag{63}$$

The length distribution of the first cycle is characterized by the same $l \sim N^{1/3}$ size scale as does the overall cycle distribution. We focus on the behavior of the moments

$$\langle l^n\rangle \sim N^{n/3-1/6}. \tag{64}$$

This behavior is obtained from the distribution (57) that should be integrated up to the appropriate cutoff, i.e., $\langle l^n\rangle \sim \int_1^{N^{1/3}} dl\, l^n\, l^{-3/2}$. As a result, the average size of the first cycle is much smaller than the characteristic cycle size $\langle l\rangle \sim N^{1/6}$. Moments corresponding to the size of the first unicyclic component grow as follows

$$\langle k^n\rangle \sim N^{2n/3-1/6}, \tag{65}$$

as obtained from (59). Consequently, the average size of the first unicyclic component is smaller than the characteristic component size, $\langle k\rangle \sim N^{1/2}$.

VII. CONCLUSIONS

In summary, we have extended the kinetic theory description of random graphs to structures such as paths and cycles. Modeling the linking process dynamically leads to an aggregation process for both components and paths. The density of paths in finite components is coupled to the component size distribution via nonlinear rate equations while the average number of cycles is coupled to the path density via linear rate equations. Both path and cycle length distributions are coupled to the component size distribution.

Generally, size distributions decay exponentially away from the gelation point, but at the gelation time, algebraic tails emerge. As the system approaches this critical point, the size distributions follow a self-similar behavior characterized by diverging size scales.

The kinetic theory approach is well-suited for treating infinite systems. The complementary behavior for finite systems can be obtained from heuristic scaling arguments. This approach yields scaling laws for the typical component size, path length, and cycle length at the gelation point. These scaling laws can be formalized using finite-size scaling forms, i.e., self-similarity as a function of the system size, rather than time. Obtaining the exact form of these scaling functions is a nice challenge in particular for the most fundamental quantity, the component size distribution that is characterized by a non-monotonic scaling function.

The kinetic theory approach seems artificial at first sight. Indeed, graphs are discrete in nature and therefore combinatorial approaches appear more natural. Yet, once the rate equations are formulated, the analysis is straightforward. Utilizing the continuous time variable allows us to employ powerful analysis tools. Moreover, some of the kinetic theory results are less cumbersome compared with the combinatorial results.

The same methodology can be expanded to analyze other features of random graphs. For example, correlations between the node degree and the cluster size can be analyzed using bi-aggregation rate equations [36]. It is quite possible that structural properties in other aggregation processes, for example, polymerization with a sum kernel [17], and in other variants of random graphs such as small-world networks [37] can be analyzed using kinetic theory.

One could try to utilize kinetic theory to probe the distribution of various families of subgraphs. We have limited ourselves to cycles since they, alongside with trees, do appear in random graphs while more interconnected families of subgraphs are very rear [29]. Yet in biological and technological networks certain interconnected families of subgraphs do appear. Such populated families of subgraphs, motifs, are believed to carry information processing functions [38, 39]. It will be interesting to use kinetic theory to analyze motifs in special random graphs.

This research was supported by the DOE (W-7405-ENG-36).

APPENDIX A: CONTOUR INTEGRATION

Let $A(z) = \sum_k A_k e^{kz}$ be the generating function of the coefficients A_k. For the family of generating functions $A(z) = G^r(z)$ with $G(z)$ satisfying $Ge^{-G} = e^z$, the coefficients A_k can be obtained via contour integration in the complex y plane where $y = e^z$ as follows

$$\begin{aligned} A_k &= \frac{1}{2\pi i}\oint dy\, \frac{G^r}{y^{k+1}} \\ &= \frac{1}{2\pi i}\oint dG\, G^r\, \frac{e^{(k+1)G}}{G^{k+1}}\frac{dy}{dG} \\ &= \frac{1}{2\pi i}\oint dG\, G^r\, \frac{e^{(k+r)G}}{G^{k+1}}(1-G)e^{-G} \\ &= \frac{1}{2\pi i}\oint dG \sum_n \frac{k^n}{n!}(G^{n+r-k} - G^{n+r+1-k}) \\ &= r\frac{k^{k-r-1}}{(k-r)!}. \end{aligned} \tag{A1}$$

Since $Ge^{-G} = e^z$, it is convenient to perform the integration in the complex G plane. In writing the third line, we used $\frac{dy}{dG} = (1-G)e^{-G}$.

[1] R. Solomonoff and A. Rapaport, Bull. Math. Biophys. **13**, 107 (1959).
[2] P. Erdős and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17 (1960).
[3] B. Bollobás, *Random Graphs* (Academic Press, London, 1985).
[4] S. Janson, T. Łuczak, and A. Rucinski, *Random Graphs* (John Wiley & Sons, New York, 2000).
[5] S. Janson, D. E. Knuth, T. Łuczak, and B. Pittel, Rand. Struct. Alg. **3**, 233 (1993).
[6] P. J. Flory, J. Amer. Chem. Soc. **63**, 3083 (1941).
[7] W. H. Stockmayer, J. Chem. Phys. **11**, 45 (1943).
[8] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953).
[9] D. Stauffer, *Introduction to Percolation Theory* (Taylor & Francis, London, 1985).
[10] T. Kalisky, R. Cohen, D. ben-Avraham, and S. Havlin, Lect. Notes. Phys. **650**, 3 (2004).
[11] B. Bollobás, C. Borgs, J. T. Chayes, J. H. Kim, and D. B. Wilson, Rand. Struct. Alg. **18**, 201 (2001).
[12] M. E. J. Newman, S. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[13] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. **99**, 7821 (2002).
[14] M. V. Smoluchowski, Physik. Zeits. **17**, 557 (1916). Zeits. Phys. Chem. **92**, 129 (1917).
[15] S. Chandrasekhar, Rev. Mod. Phys. **15**, 1–89 (1943).
[16] D. J. Aldous, Bernoulli **5**, 3 (1999).
[17] F. Leyvraz, Phys. Rep. **383**, 95 (2003).
[18] J. B. McLeod, Quart. J. Math. Oxford **13**, 119 (1962); *ibid* **13**, 193 (1962); *ibid* **13**, 283 (1962).
[19] E. M. Hendriks, M. H. Ernst, and R. M. Ziff, J. Stat. Phys. **31**, 519 (1983).
[20] E. Ben-Naim and P. L. Krapivsky, Europhys. Lett. **65**, 151 (2004).
[21] E. Ben-Naim and P. L. Krapivsky, J. Phys. A **37**, L189 (2004).
[22] A. A. Lushnikov, J. Colloid. Inter. Sci. **65**, 276 (1977).
[23] P. G. J. van Dongen and M. H. Ernst, J. Stat. Phys. **49**, 879 (1987).
[24] D. H. Zanette and S. C. Manrubia, Physica A **295**, 1 (2001).
[25] S. N. Dorogovtsev J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **63**, 062101 (2001).
[26] Z. Burda, J. D. Correia, and A. Krzywicki, Phys. Rev. E **64**, 046118 (2001).
[27] P. L. Krapivsky and S. Redner, J. Phys. A **35**, 9517 (2003).
[28] P. L. Krapivsky and E. Ben-Naim, Phys. Rev. E **53**, 291 (1996).
[29] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, Phys. Rev. E **68**, 026127 (2003).
[30] H. D. Rozenfeld, J. E. Kirk, E. M. Bollt, and D. ben-Avraham, *cond-mat/0403536*.
[31] E. Marinari and R. Monasson, *cond-mat/0407253*.
[32] S. Janson, Rand. Struc. Alg. **17**, 343 (2000).
[33] S. Janson, Combin. Probab. Comput. **12**, 27 (2003).
[34] Z. Burda, J. Jurkiewicz, and A. Krzywicki, Phys. Rev. E **69**, 026106 (2004); *ibid* **70**, 026106 (2004).
[35] G. Bianconi and A. Capocci, Phys. Rev. Lett. **90**, 078701 (2003); G. Bianconi, G. Caldarelli, and A. Capocci, *cond-mat/0408349*.
[36] E. Ben-Naim and P. L. Krapivsky, unpublished.
[37] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[38] R. Milo *et al*, Science **298**, 824–827 (2002); *ibid* **303**, 1538–1542 (2004).
[39] V. Spirin and L. A. Mirny, Proc. Natl. Acad. Sci. **100**, 12123–12128 (2003).