


RESEARCH ARTICLE

Random networks grown by fusing edges via urns

Kiran R. Bhutani¹, Ravi Kalpathy^{1*}  and Hosam Mahmoud²

¹Department of Mathematics, The Catholic University of America, Washington, DC, USA and ²Department of Statistics, The George Washington University, Washington, DC, USA

*Corresponding author. Email: kalpathy@cua.edu

Action Editor: Ulrik Brandes

Abstract

Many classic networks grow by hooking small components via vertices. We introduce a class of networks that grows by fusing the edges of a small graph to an edge chosen uniformly at random from the network. For this random edge-hooking network, we study the local degree profile, that is, the evolution of the average degree of a vertex over time. For a special subclass, we further determine the exact distribution and an asymptotic gamma-type distribution. We also study the “core,” which consists of the well-anchored edges that experience fusing. A central limit theorem emerges for the size of the core.

At the end, we look at an alternative model of randomness attained by preferential hooking, favoring edges that experience more fusing. Under preferential hooking, the core still follows a Gaussian law but with different parameters. Throughout, Pólya urns are systematically used as a method of proof.

Keywords: network; random graph; degree profile; Pólya urn; limit law; preferential attachment; phase transition; Stirling number; asymptotic analysis; polyacenes

1. Network growth by edge hooking

Constructing graphs by adding vertices and edges is in numerous classical studies. In the vogue these days is an area of research investigating networks grown by hooking more complex components (Bhutani et al., 2021; Chen & Mahmoud, 2016; Desmarais & Holmgren, 2018; Desmarais & Mahmoud, 2021; Drmota et al., 2008; Gopaladesikan et al., 2014; Mahmoud, 2019) via *vertices*.

In this article, we consider networks grown by adding complex components by fusing *edges*. We have a seed graph from which we build a dynamic network (a sequence of graphs G_n , for $n \geq 0$). The first graph in the network, G_0 , is the seed itself. A particular edge in the seed is designated as the *hooking edge* (or simply the hook). We call that hooking edge $e = \{u, v\}$, where u and v are two distinct vertices in the seed.

At step $n - 1$, we have grown the network G_{n-1} . The next graph in the sequence is obtained by choosing *uniformly at random* an edge, $e' = \{u', v'\}$, where u' and v' are two distinct vertices in the network G_{n-1} , to which a copy of the seed is attached by fusing the edges e and e' . This fusing can be made in two different ways—we can identify u and u' together and identify v and v' together, or we can do the opposite, that is, we identify u and v' together and identify v and u' together. We call the edge chosen in the network for fusing the *latching edge* (or simply the latch), and the two possible hookings correspond to two *orientations*. We take the two orientations to be equally likely. The fused edges merge into one. Different orientations can give rise to nonisomorphic graphs.

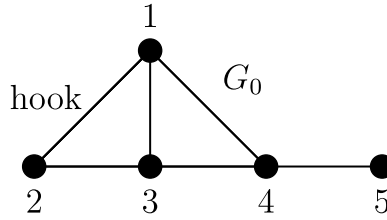


Figure 1. A seed labeled canonically.

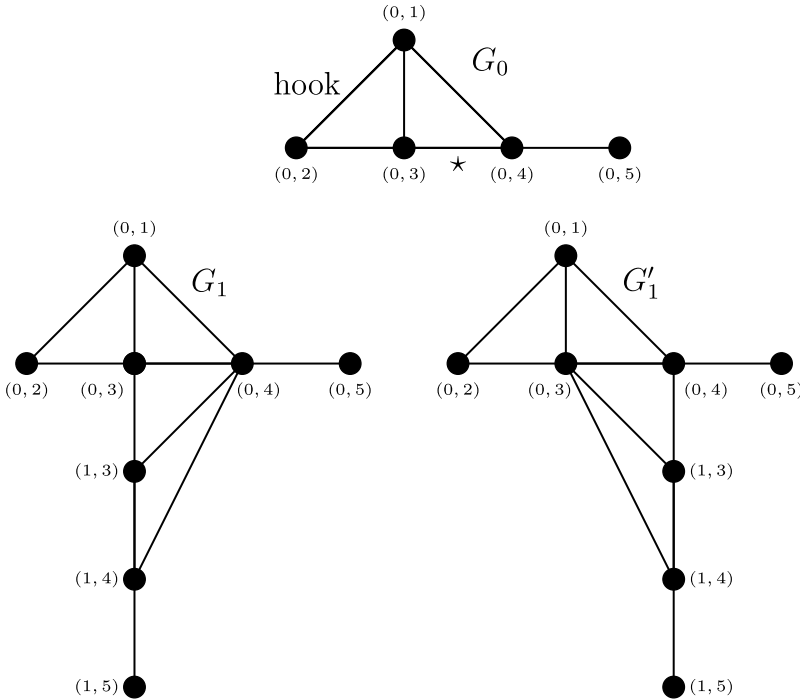


Figure 2. A dynamically labeled seed (top) and two dynamically labeled networks grown from it (second row), each corresponding to one hooking orientation. The edge chosen for hooking is marked with a star.

Figure 1 illustrates a seed with a hooking edge. The seed nodes carry numbers according to a canonical labeling that will be explained in Subsection 3.1.

Figure 2 shows networks grown from the seed by the two different orientations, while fusing the hook and the latch. The graphs in the growing network carry labels derived from the canonical numbering of the seed nodes and additional timestamps indicating the entry point in time of a seed copy (more on this in Section 3.1).

The graph G_1 is constructed by obtaining a new copy of the seed and hooking vertex 1 in the new seed to vertex 4 in the graph G_0 and vertex 2 in the new seed to vertex 3 in G_0 . The graph G'_1 corresponds to the alternative orientation, in which vertex 1 in the new seed is hooked to vertex 3 in the graph G_0 and vertex 2 in the new seed is hooked to vertex 4 in G_0 . Each of the two graphs G_1 and G'_1 occurs with probability $\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$. The nodes in the graphs of Figure 2 carry labels that will be explained later. Note that the two orientations give rise to the two *nonisomorphic* graphs G_1 and G'_1 .

In the context of edge-fused networks, a *seed* is a connected graph $S = (V, \mathcal{E})$ with a set of vertices V and a set of edges \mathcal{E} containing at least one edge. A particular edge in \mathcal{E} is distinguished

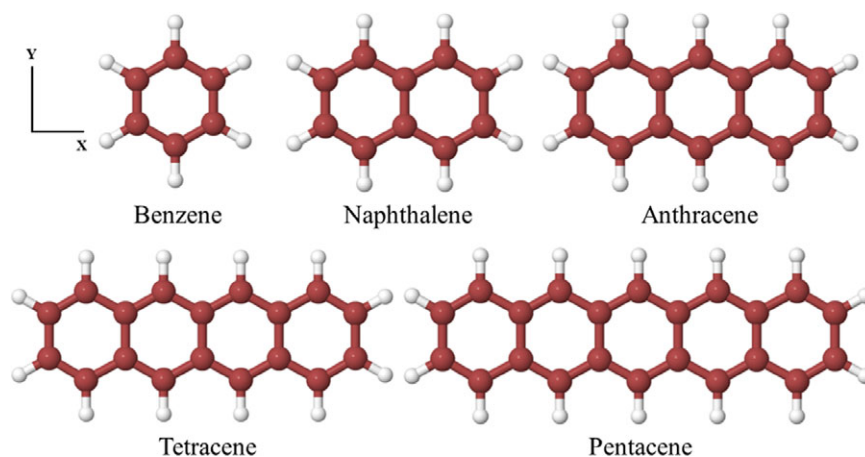


Figure 3. The topology of benzene ring fusion in polyacenes.

as the hook, the edge designated for fusing into larger graphs. The seed is used as a building block in a bigger network in discrete time steps. We use n for discrete time. Consequently, we can think of the graph G_n as the network at age n . Technically speaking, we generate a sequence of seeds S_0, S_1, S_2, \dots that are identical to one prototype seed S . Each copy then has its own hook. At time 0, the network starts as S_0 . At time $n \geq 1$, the copy S_n is adjoined to the network and its hook is used for edge hooking.

In the sequel, when we speak of a property of a seed without specifying its timestamp, we refer to the prototype seed S . We call the class of graphs so built *edge-hooking networks*.

The study is motivated by the recent interest in graphs with “community structure,” see van der Hofstad et al. (2018), for example. Graphs with community structure have been suggested as models for epidemiology (Bhamidi et al., 2021; Trapman, 2007). The proposed edge-hooking network has a tree as a backbone, where a tree edge may develop into a “local community,” that is, blown out to be an entire graph (copy of the seed).

As an application of the edge-hooking mechanism, we present an example from organic chemistry. Acenes or polyacenes are polycyclic aromatic hydrocarbons that consist of linearly fused benzene rings. This threadlike network of benzene ring fusion is shown in Figure 3.¹

Polyacenes follow the molecular formula $C_{4k+2}H_{2k+4}$, where $k \geq 1$. As a particular example of edge-hooking networks, G_0 will be benzene having a molecular formula with $k = 1$. One can think of G_1 as naphthalene having a molecular formula with $k = 2$. Then, G_2 will be anthracene having a molecular formula with $k = 3$ and so on. Our edge fusing network model exhibits the structure and geometry that happens in these organic compounds.

Acenes are an important and unique class of organic compounds (Tönshoff & Bettinger, 2021). The smaller acenes like benzene, naphthalene, and anthracene are among the well-studied organic compounds that can be produced from coal or petroleum products. The larger acenes and their derivatives form organic semiconductor materials. They possess unique and versatile electronic properties, and scientists find them to be attractive candidates for use in organic electronic devices. They are currently the subject of great interest in materials science and engineering (Akter et al., 2021; Anthony, 2008; Zamoshchik et al., 2013).

2. Scope

A paramount property of graphs is the distribution of node degrees within. For example, the degree of a node in a social network, namely the number of followers, may be taken as a measure of the influence of a person in the network.

We study the local degree profile of a network, which tracks the evolution of a specific vertex in the network over time. We also study the size of the core, which is the number of edges that experience fusing. Via fusing, edges in the core are more anchored in the graph and can be “thicker” than other edges. The notion of edge fusing is studied in a different context in Bhutani & Khan (2003), where the authors looked at the multiplicity of an edge with respect to a given set of paths. Multiplicity there is defined as the number of times an edge is used to construct the named set of paths. The concept in (Bhutani & Khan, 2003) is introduced to define a distance between two graphs.

At the end of the study, we look into an alternative model of preferential edge hooking. To see the contrast to the uniform model, we develop a result for the size of the core under preferential attachment.

3. Notation

We denote the cardinality of a set B by $|B|$. The prototype seed $S = (V, \mathcal{E})$ has structural properties, such as the degrees of its nodes. It is expected that network characteristics (such as a degree profile) will depend on these properties. For instance, distributions associated with the network may be parameterized by some counts in S . The size of the seed $\epsilon_0 := |\mathcal{E}|$ appears in the results. In the case $\epsilon_0 = 1$, the seed is the K_2 ,² the complete graph on two vertices. Such a case is degenerate, as fusing copies of the seed produces K_2 . The network does not progress beyond the initial K_2 , and there is nothing to study. In what follows we assume $\epsilon_0 \geq 2$.

Let ϵ_n be the size (the number of edges) of the network at age n . As we add a copy of the seed at each step, and the hook is subsumed in the graph, i.e., does not add to the edges of the network, we have

$$\epsilon_n = (\epsilon_0 - 1)n + \epsilon_0.$$

Some exact results will be represented in terms of Pochhammer’s symbol for the rising factorial. The m -times rising factorial of a real number x is

$$\langle x \rangle_m = x(x + 1) \dots (x + m - 1),$$

with the interpretation that $\langle x \rangle_0 = 1$.

We also have occasion to use $\left\{ \begin{matrix} m \\ i \end{matrix} \right\}$, the i th Stirling number of order m of the second kind, which is the number of ways to divide a nonempty set of size m into i nonempty subsets. For properties of Stirling numbers, we refer the reader to David & Barton (1962) and Graham et al. (1994).

3.1 Canonical labeling

Assume the nodes of the seed are numbered $1, \dots, |V|$, a numbering in which the two nodes of the hook are numbered 1 and 2, the rest are arbitrarily labeled distinctly with the numbers $3, 4, \dots, |V|$. We take this numbering as canonical to be preserved in every copy of S . Since the canonical labels in the seed repeat in the network, a tiebreaking mechanism is needed for unique identification of network nodes. The tiebreaker is the timestamp—the point in time at which a node joins the graph. In such a “dynamic labeling,” a node’s tag carries two components: the canonical labeling and the timestamp.

Figure 2 shows one such canonical numbering. A node appearing in the network is thus adequately specified by a pair (j, r) , if it appears for the first time at stage j and is the r th in the canonical numbering of S_j .

Let d_r be the degree of node r in the seed, for $1 \leq r \leq |V|$. According to the canonical labeling, in the seed, the two ends of the hook $\{1, 2\}$ have degrees d_1 and d_2 . We use $\Delta_{j,n}^{(r)}$ to denote the degree of node (j, r) in a network of age $n \geq j$. For instance, in the running example of Figure 2, we

have $\Delta_{0,0}^{(3)} = 3$, and if the stochastic path follows G_1 we would have $\Delta_{0,1}^{(3)} = 4$, but if the stochastic path follows G'_1 , we would have $\Delta_{0,1}^{(3)} = 5$.

Note that, for $r = 3, 4, \dots, |V|$, the node (j, r) appears at time $j \geq 0$, whereas the nodes $(j, 1)$ and $(j, 2)$ appear only when $j = 0$; this is because the fusing of edges subsumes the seed edge $\{1, 2\}$ of the seed in older edges, and the end nodes of these receiving edges have already been labeled at earlier stages.

In the sequel, we use the symbol $\xrightarrow{a.s.}$ to denote almost sure convergence and the symbol $\xrightarrow{\mathcal{D}}$ to denote convergence in distribution. The notation $\mathcal{N}(0, \sigma^2)$ stands for a centered normal random variable with variance σ^2 .

4. Pólya urns

An instrument we use in this investigation is Pólya urn (Mahmoud, 2008). A c -color Pólya urn scheme is comprised of an initial nonempty urn containing balls of up to c -colors and rules to operate on the urn at discrete time steps. The colors are numbered $1, 2, \dots, c$. At each time step, a ball is drawn at random from the urn and its color is observed. If the color of the ball withdrawn is i , we put it back in the urn and add $a_{i,j}$ (possibly negative or even random) balls of color j , for $j = 1, \dots, c$, and the drawing is continued. These dynamics are captured in a $c \times c$ replacement matrix:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,c} \\ a_{2,1} & a_{2,2} & \dots & a_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{c,1} & a_{c,2} & \dots & a_{c,c} \end{pmatrix}.$$

When some entries of the replacement matrix are negative, a tenability issue (of indefinite drawing) may arise. These cases do not appear in our study—all the replacement matrices in this article contain only nonnegative elements.

5. Modeling the local profile by a Pólya urn

The strategy to track the degree $\Delta_{j,n}^{(r)}$ is to map the edges of the network onto balls in a Pólya urn. The edges incident with the vertex we are tracking correspond to the white balls in the urn and all the other edges in the network correspond to the blue balls in the urn. Thus, the count of the white balls at any stage is the degree of the vertex under tracking at that stage.

We wait till the node $y = (j, r)$ appears in the network to start a Pólya urn with d_r white balls and $\epsilon_j - d_r = (\epsilon_0 - 1)j + \epsilon_0 - d_r$ blue balls. We use the terms “edges” and “balls” interchangeably, so we can think directly of an urn of white and blue edges.

When we pick a white edge at step n (one of the $\Delta_{j,n-1}^{(r)}$ edges incident with y), node y and one of its neighbors are involved in the edge hooking. Upon hooking a copy of the seed, we fuse the seed edge joining the vertices with labels 1 and 2 in the canonical numbering. We increase the degree of y either by $d_1 - 1$ or $d_2 - 1$, (with equal probability) depending on the orientation of the hooking; the -1 accounts for the loss of the seed fused edge upon hooking. The orientation chosen is determined as in the outcome of flipping a fair coin. Let $\mathbb{I}_{\mathcal{H}}$ be the indicator of the event that the coin toss yields Heads. The number of white edges increases by a variable distributed like

$$d_1 \mathbb{I}_{\mathcal{H}} + d_2(1 - \mathbb{I}_{\mathcal{H}}) - 1.$$

Except for the edges incident with y in G_{n-1} , hooking at any other edge does not alter the degree of y , we add $\epsilon_0 - 1$ blue edges.

Letting \mathbb{F}_n be the sigma field generated by the first n steps, the conditional distribution of the degree of node (j, r) is

$$\Delta_{j,n}^{(r)} \mid \mathbb{F}_{n-1} = \begin{cases} \Delta_{j,n-1}^{(r)} + d_1 \mathbb{I}_{\mathcal{H}} + d_2(1 - \mathbb{I}_{\mathcal{H}}) - 1, & \text{with probability } \frac{\Delta_{j,n-1}^{(r)}}{\epsilon_{n-1}}; \\ \Delta_{j,n-1}^{(r)}, & \text{with probability } 1 - \frac{\Delta_{j,n-1}^{(r)}}{\epsilon_{n-1}}. \end{cases}$$

These dynamics are captured by the Pólya replacement matrix

$$\begin{pmatrix} d_1 \mathbb{I}_{\mathcal{H}} + d_2(1 - \mathbb{I}_{\mathcal{H}}) - 1 & \epsilon_0 - d_1 \mathbb{I}_{\mathcal{H}} - d_2(1 - \mathbb{I}_{\mathcal{H}}) \\ 0 & \epsilon_0 - 1 \end{pmatrix}.$$

Traditionally, a two-color urn scheme with a replacement matrix with one off-diagonal element being 0 and the other off-diagonal element being positive is called *triangular*.

We analyze the average behavior of balanced urns of this type; the balance means the row sum is constant. Take an integer $\delta \geq 1$ and construct a triangular Pólya urn scheme on white and blue balls with the replacement matrix

$$\begin{pmatrix} X & \delta - X \\ 0 & \delta \end{pmatrix}, \tag{1}$$

with X having a distribution on the integers $0, 1, \dots, \delta - 1$, with mean μ .

Let W_m be the number of white balls after m draws from the urn, and let Q_m be the total number of balls in the urn at time m . Thus, Q_0 is the initial number of balls in the urn and we have

$$Q_m = \delta m + Q_0.$$

More precisely, at the m th drawing, X_m is the m th realization of X . So, X_1, X_2, \dots is a sequence of independent identically distributed random variables, and in the m th drawing we use the replacement matrix

$$\begin{pmatrix} X_m & \delta - X_m \\ 0 & \delta \end{pmatrix}.$$

After m draws, the number of white balls satisfies the stochastic recurrence

$$W_m = W_{m-1} + X_m \mathbb{I}_m^{(W)},$$

where $\mathbb{I}_m^{(W)}$ is the indicator of the event of picking a white ball in the m th drawing. Conditioning on \mathbb{F}_{m-1} (the sigma field generated by the first $m - 1$ draws), we have the expectation

$$\mathbb{E}[W_m \mid \mathbb{F}_{m-1}] = W_{m-1} + \mathbb{E}[X_m] \mathbb{E}\left[\mathbb{I}_m^{(W)} \mid \mathbb{F}_{m-1}\right],$$

following from the independence of X_m of the history. We then have

$$\mathbb{E}[W_m \mid \mathbb{F}_{m-1}] = W_{m-1} + \mu \frac{W_{m-1}}{Q_{m-1}}. \tag{2}$$

An iterated expectation gives us the unconditional expectation

$$\mathbb{E}[W_m] = \left(1 + \frac{\mu}{Q_{m-1}}\right) \mathbb{E}[W_{m-1}].$$

Unwinding this recurrence, we obtain

$$\begin{aligned} \mathbb{E}[W_m] &= \left(1 + \frac{\mu}{Q_{m-1}}\right) \dots \left(1 + \frac{\mu}{Q_0}\right) W_0, \\ &= \left(\frac{(m-1)\delta + Q_0 + \mu}{(m-1)\delta + Q_0}\right) \dots \left(\frac{Q_0 + \mu}{Q_0}\right) W_0 \\ &= \left(\frac{m-1 + \frac{Q_0 + \mu}{\delta}}{m-1 + \frac{Q_0}{\delta}}\right) \dots \left(\frac{\frac{Q_0 + \mu}{\delta}}{\frac{Q_0}{\delta}}\right) W_0 \\ &= \frac{\Gamma\left(m + \frac{Q_0 + \mu}{\delta}\right) \Gamma\left(\frac{Q_0}{\delta}\right)}{\Gamma\left(m + \frac{Q_0}{\delta}\right) \Gamma\left(\frac{Q_0 + \mu}{\delta}\right)} W_0. \end{aligned} \tag{3}$$

This derivation furnishes a proof for the average degree of a node in the network. The proof concerns a node (j, r) , and we keep in mind that j may be a function of n . For example, we may specify some large n and decide to look at the degree of a node that appeared at time $j = j(n) = \lceil \sqrt{n} \rceil$.

In the context of exact probability calculation, we just use j and n , as they are both fixed numbers. In the context of asymptotics, it is preferable to use the notion j_n to capture any dependence therein between the index j and the age n .

Theorem 5.1. *Let $y = (j, r) = (j_n, r)$ be the r th node in the canonical representation of the j_n th seed copy in the network. Suppose the two ends of the hook $\{1, 2\}$ have degrees d_1 and d_2 . At time n , the average degree of y is*

$$\mathbb{E}\left[\Delta_{j,n}^{(r)}\right] = \frac{\Gamma\left(j + \frac{\epsilon_0}{\epsilon_0 - 1}\right) \Gamma\left(n + \frac{d_1 + d_2 + 2\epsilon_0 - 2}{2(\epsilon_0 - 1)}\right)}{\Gamma\left(j + \frac{d_1 + d_2 + 2\epsilon_0 - 2}{2(\epsilon_0 - 1)}\right) \Gamma\left(n + \frac{\epsilon_0}{\epsilon_0 - 1}\right)} d_r.$$

Asymptotically, as $n \rightarrow \infty$, we have

$$\mathbb{E}\left[\Delta_{j,n}^{(r)}\right] = \begin{cases} d_r \frac{\Gamma\left(j + \frac{\epsilon_0}{\epsilon_0 - 1}\right)}{\Gamma\left(j + \frac{d_1 + d_2 + 2\epsilon_0 - 2}{2(\epsilon_0 - 1)}\right)} n^{\frac{d_1 + d_2 - 2}{2(\epsilon_0 - 1)}}, & \text{if } j_n = O(1); \\ d_r \left(\frac{n}{j}\right)^{\frac{d_1 + d_2 - 2}{2(\epsilon_0 - 1)}}, & \text{if } j_n \rightarrow \infty. \end{cases}$$

Proof. The replacement matrix of the urn scheme associated with the local degree profile of the node (j, r) is triangular of the type (1), with

$$X = d_1 \mathbb{I}_{\mathcal{H}} + d_2(1 - \mathbb{I}_{\mathcal{H}}) - 1, \quad \mu = \frac{d_1 + d_2}{2} - 1, \quad \delta = \epsilon_0 - 1;$$

The scheme has the initial conditions

$$W_0 = d_r, \quad Q_0 = \epsilon_j = (\epsilon_0 - 1)j + \epsilon_0.$$

The exact average of $\Delta_{j,n}^{(r)}$ follows upon plugging these values in (3), and evaluating at time $m = n - j$.

The asymptotic equivalent follows from Stirling approximation of the ratio of gamma function (see article 6.1.47 in Abramowitz & Stegun (1972)):

$$\frac{\Gamma(x + a)}{\Gamma(x + b)} \sim x^{a-b}, \quad \text{as } x \rightarrow \infty. \quad \square$$

Remark 5.1. Reference (Janson, 2006) deals only with urn schemes with fixed entries. Edge fusing calls for urn schemes with random entries.

Remark 5.2. From the affine transformation of W_m in (2), we can quickly show that

$$S_m = \left(\prod_{i=0}^{m-1} \frac{Q_i}{Q_i + \mu} \right) W_m$$

is a martingale. The top line in the chain of derivations (3) can be obtained from $\mathbb{E}[S_m] = \mathbb{E}[S_0]$. The gamma asymptotics used in the proof of Theorem 5.1 also tell us (via the martingale convergence theorem) that $W_m/m^{\mu/\delta}$ converges (as $m \rightarrow \infty$) almost surely and in L^1 to “some” limit.

Remark 5.3. Reference (Aguech, 2009) gives strong laws for a broader class of triangular urns, but in less explicit terms. The entries of the replacement matrix in Aguech (2009) are bounded in L^2 , and the author gets convergence almost surely. The strong convergences in Aguech (2009) are to unknown limits that come out of the martingale convergence theorem, which is only an existential statement that does not specify the limits. To guarantee tenability, the entries of the triangular urn we are considering are uniformly bounded by δ . So, they are bounded in L^2 and in the almost-sure sense, as well. Results in Aguech (2009) apply.

5.1 Interpreting the phases in the local degree profile

Note that in Theorem 5.1 the power of n is less than 1—this can be seen from the fact that

$$d_1 + d_2 - 2 < \sum_{r=1}^{|V|} d_r - 2 = 2\epsilon_0 - 2,$$

with K_1 and K_2 forbidden as seeds, the inequality is strict.

Theorem 5.1 portrays a spectrum covering all the nodes by time n . We can think of j as $j_n = an + h(n)$, with $h(n) = o(n)$. In the very early sublinear phase in which $j = j_n = O(1)$, such as the cases $j = 5$ or $j = 13 + (-1)^n$, the average degree of the node (j, r) is of the form $\mathbb{E}[\Delta_{j,n}^{(r)}] \sim c_j n^{\frac{d_1+d_2-2}{2(\epsilon_0-1)}}$, for a $O(1)$ coefficient c_j that depends on j and the seed parameters. Note that in the instance $j_n = 13 + (-1)^n$, the coefficient c_j oscillates.

In the later sublinear phase, we still have $a = 0$, and $j_n \rightarrow \infty$, such that $j_n = 0 + h(n) = o(n)$. For instance j_n may be $\lfloor 6 \ln n + \pi \rfloor$, or $\lceil 4 \frac{n}{\ln n} + \frac{7}{6} + \frac{1}{n} \rceil$. In this phase, we have $\mathbb{E}[\Delta_{j_n,n}^{(r)}] \sim d_r (n/h(n))^{\frac{d_1+d_2-2}{2(\epsilon_0-1)}}$.

Next comes the linear phase, with $a \in (0, 1)$. In this phase, we have

$$\mathbb{E}[\Delta_{j_n,n}^{(r)}] \sim d_r \left(\frac{1}{a} \right)^{\frac{d_1+d_2-2}{2(\epsilon_0-1)}}. \tag{4}$$

Finally, we reach the very late linear phase, with $a = 1$, such as the case $j_n = \lceil n - \ln n \rceil$ or $j_n = n - 7$. In this phase, we have $\mathbb{E}[\Delta_{j_n,n}^{(r)}] \sim d_r$. This is to be anticipated, since the edges incident with these very late nodes compete with an overwhelming number of edges in the graph to be the latch,

and by time n they have not had enough time to succeed. So, they retain the degree with which they joined the network.

The index j_n may have a complex dependence on n . For instance, one could conceive of the choice $j_n = a_n n + h(n)$, where the coefficient a_n is a function of n . This may lead to multiple asymptotic subsequences of asymptotic equivalents, as for example the case $a_n = \frac{1}{3}$ for even n , and $a_n = \frac{2}{3}$ for odd n . In such an instance, formula (4) is applicable with a alternating appropriately with the parity of n .

5.2 The subclass with both ends of the hook having the same degree

With $d_1 \neq d_2$, the underlying urn is unbalanced. As mentioned, an appeal to the known theory of urns can only produce nonconstructive almost-sure convergence (to an unspecified limit variable). In the case of $d_1 = d_2$, we can go further and develop a distributional result eliciting the limit.

If both ends of the hook $\{1, 2\}$ are of the same degree d_1 , the Pólya urn scheme is simplified to a balanced scheme with a triangular deterministic replacement matrix, which is namely

$$\begin{pmatrix} \alpha & \delta - \alpha \\ 0 & \delta \end{pmatrix}.$$

This is a well-studied urn and we can say much more beyond the average degree of a node. This scheme has been analyzed in Flajolet et al. (2006), Janson (2006), and Zhang et al. (2015). The exact moments have been characterized in Zhang et al. (2015). In this case, the moments of the number of white balls after m draws are given by

$$\mathbb{E}[W_m^k] = \frac{\alpha^k}{\left\langle \frac{Q_0}{\delta} \right\rangle_m} \sum_{i=1}^k (-1)^{k-i} \begin{Bmatrix} k \\ i \end{Bmatrix} \left\langle \frac{W_0}{\alpha} \right\rangle_i \left\langle \frac{Q_0 + i\alpha}{\delta} \right\rangle_m.$$

For an edge-hooking network, where the two ends of the hooking edge have the same degree d_1 , we have a correspondence with an urn with

$$\alpha = d_1 - 1, \quad \delta = \epsilon_0 - 1.$$

The scheme has the initial conditions

$$W_0 = d_r, \quad Q_0 = \epsilon_j = (\epsilon_0 - 1)j + \epsilon_0,$$

and the degree of node (j, r) by time n is the number of white balls in the urn at time $m = n - j$. The moments equation in Zhang et al. (2015) readily yields

$$\begin{aligned} \mathbb{E}\left[\left(\Delta_{j,n}^{(r)}\right)^k\right] &= \frac{(d_1 - 1)^k}{\left\langle \frac{(\epsilon_0 - 1)j + \epsilon_0}{\epsilon_0 - 1} \right\rangle_{n-j}} \sum_{i=1}^k (-1)^{k-i} \begin{Bmatrix} k \\ i \end{Bmatrix} \left\langle \frac{d_r}{d_1 - 1} \right\rangle_i \\ &\quad \times \left\langle \frac{(\epsilon_0 - 1)j + \epsilon_0 + i(d_1 - 1)}{\epsilon_0 - 1} \right\rangle_{n-j}. \end{aligned}$$

The asymptotic moments are given in Flajolet et al. (2006) and Janson (2006), see also Zhang et al. (2015). Translated to the urn associated with node degrees, as $j_n \rightarrow \infty$, such that $j_n = o(n)$, we have

$$\mathbb{E}\left[\left(\frac{\Delta_{j_n,n}^{(r)}}{(n/j_n)^{\frac{d_1-1}{\epsilon_0-1}}}\right)^k\right] \rightarrow (d_1 - 1)^k \frac{\Gamma\left(k + \frac{d_r}{d_1 - 1}\right)}{\Gamma\left(\frac{d_r}{d_1 - 1}\right)}.$$

In this convergence relation, the right-hand side is the k th moment of the so-called gamma-type distribution (Janson, 2010), which is uniquely determined by its moments. Hence, $\Delta_{j_n,n}^{(r)} / (n/j_n)^{\frac{d_1-1}{\epsilon_0-1}}$ converges to the so-characterized gamma-type distribution.

The special case in which both ends of the hooking edge have the same degree covers, for example, hooking regular graphs. As an instantiation, consider the graph obtained from triangles by fusing edges. Here, we have $\epsilon_0 = 3$ and $d_1 - 1 = 1$, and we have the convergence

$$\mathbb{E} \left[\left(\frac{\Delta_{j_n,n}^{(r)}}{(n/j_n)^{\frac{1}{2}}} \right)^k \right] \rightarrow \Gamma(k + 2),$$

as $j_n \rightarrow \infty$, with $j_n = o(n)$. Note that the right-hand side of the above convergence relation is the k th moment of a gamma (2, 1) random variable.

6. The core

Edges of the network fall in two categories. There are edges that are used as hooks and there are other edges that are yet to be used in hooking. We call the edges fused in hooking operations the *core* as they are “soldered” with other edges and their position is thicker and sturdier in the network.

Let C_n be the size of the core (i.e., the number of edges in it) at age n . For example, the starred edge in Figure 2 in G_0 is not in the core of the network at time 0, but becomes in the core of G_1 . We always have $C_0 = 0, C_1 = 1$, and randomness appears at age 2 and beyond. Note that when an edge is selected as a latch and the seed is hooked, both orientations of the hooking give the same number of edges in the core. If the edge (0, 1) and (0, 3) in G_1 is the latch to produce G_2 , both hooking orientations give $C_2 = 2$.

We compute the exact mean and ultimately an asymptotic Gaussian distribution via Pólya urns.

6.1 Exact mean

Let C_{n-1} be the event of picking a latch from the core of the graph G_{n-1} for hooking, and let $\mathbb{I}_{C_{n-1}}$ be its indicator. The size of the core increases (by 1) only if the latch chosen is not in the core. We have a stochastic recurrence:

$$C_n = C_{n-1} + 1 - \mathbb{I}_{C_{n-1}}. \tag{5}$$

From this relation, we can get the exact mean.

Theorem 6.1. *Let C_n be the size of the core in an edge-hooking network of age n . We have*

$$\mathbb{E}[C_n] = 1 + \sum_{i=1}^{n-1} \frac{\langle i + 1 \rangle_{n-i}}{\left\langle i + \frac{\epsilon_0}{\epsilon_0 - 1} \right\rangle_{n-i}}.$$

Proof. Let \mathbb{F}_n be the sigma field generated by the first n hookings. The conditional expectation (on \mathbb{F}_{n-1}) arising from the stochastic recurrence (5) is

$$\mathbb{E}[C_n | \mathbb{F}_{n-1}] = C_{n-1} + 1 - \mathbb{E}[\mathbb{I}_{C_{n-1}} | \mathbb{F}_{n-1}] = C_{n-1} + 1 - \frac{C_{n-1}}{\epsilon_{n-1}}. \tag{6}$$

Collecting similar terms and taking an iterated expectation, we get the unconditional expectation

$$\mathbb{E}[C_n] = \left(1 - \frac{1}{\epsilon_{n-1}} \right) \mathbb{E}[C_{n-1}] + 1.$$

This recurrence equation is of the standard linear form

$$y_n = g_n y_{n-1} + h_n,$$

with solution

$$y_n = \sum_{i=1}^n h_i \prod_{j=i+1}^n g_j + y_0 \prod_{j=1}^n g_j.$$

In view of the initial condition $C_0 = 0$, the solution for the average core size (for $n \geq 1$) is

$$\mathbb{E}[C_n] = \sum_{i=1}^n \prod_{j=i+1}^n \frac{\epsilon_{j-1} - 1}{\epsilon_{j-1}},$$

with the interpretation of an empty product as 1. So, we have

$$\begin{aligned} \mathbb{E}[C_n] &= 1 + \sum_{i=1}^{n-1} \prod_{j=i+1}^n \frac{(\epsilon_0 - 1)(j - 1) + \epsilon_0 - 1}{(\epsilon_0 - 1)(j - 1) + \epsilon_0} \\ &= 1 + \sum_{i=1}^{n-1} \prod_{j=i+1}^n \frac{j}{j - 1 + \frac{\epsilon_0}{\epsilon_0 - 1}} \\ &= 1 + \sum_{i=1}^{n-1} \frac{\langle i + 1 \rangle_{n-i}}{\left\langle i + \frac{\epsilon_0}{\epsilon_0 - 1} \right\rangle_{n-i}}. \end{aligned}$$

□

6.2 Core asymptotics

We derive a strong law and an asymptotic Gaussian law for C_n via a connection to Bagchi–Pal urn (Bagchi & Pal, 1985).

Theorem 6.2. *Let C_n be the size of the core in an edge-hooking network of age n . We have*

$$\frac{C_n}{n} \xrightarrow{a.s.} \frac{\epsilon_0 - 1}{\epsilon_0},$$

and

$$\frac{C_n - \frac{\epsilon_0 - 1}{\epsilon_0} n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{(\epsilon_0 - 1)^2}{\epsilon_0^2(\epsilon_0 + 1)}\right).$$

Proof. Think of the edges in the graph as colored balls in an urn—edges in the core are white and edges not in the core are blue. For the next graph, we choose an edge at random (pick an edge (ball) at random from the urn). The chosen edge is the latch. If the latch is white, this edge has already participated in hooking and is in the core. If we hook a copy of the seed at this latch, the status of the latch does not change, and it remains in the core; however, we add $\epsilon_0 - 1$ fresh edges that have not yet participated in any recruiting. Consequently, we leave the number of white balls unchanged and add $\epsilon_0 - 1$ blue balls to the urn.

Alternatively, the latch is a blue edge that has not participated in hooking before. After hooking the seed to the latch, the fused edge has now moved into the core, a new white edge at the expense of an exiting blue edge, and $\epsilon_0 - 1$ blue edges come with the new seed, a net gain of $\epsilon_0 - 2$ blue edges. The corresponding replacement matrix is

$$\begin{pmatrix} 0 & \epsilon_0 - 1 \\ 1 & \epsilon_0 - 2 \end{pmatrix}.$$

This is an instance of Bagchi–Pal urn, and the stated result follows from a well-developed theory. The strong law has long been known (Athreya & Karlin, 1968), and a central limit theorem appears in Bagchi & Pal (1985), see also Section of Mahmoud (2008). □

Remark 6.1. The standard urn theory specifies an asymptotic average but does not give an exact value. We thought it is worth it to develop the exact average (cf. Theorem 6.1) as it can give us sharper asymptotic approximation. For instance, if $\epsilon_0 = 3$, the exact average formula in Theorem 6.1 reduces to

$$\mathbb{E}[C_n] = \frac{2}{3}n + 1 - \frac{\sqrt{\pi} \Gamma(n + 1)}{2 \Gamma(n + \frac{3}{2})},$$

and $n \rightarrow \infty$, we have the approximation

$$\mathbb{E}[C_n] = \frac{2}{3}n + 1 - \frac{1}{2}\sqrt{\frac{\pi}{n}} + \frac{3}{16}\sqrt{\frac{\pi}{n^3}} - \frac{25}{256}\sqrt{\frac{\pi}{n^5}} + O\left(\sqrt{\frac{1}{n^7}}\right),$$

following from well-known approximations of the ratio of gamma function (see article 6.1.47 in Abramowitz & Stegun (1972)).

7. Preferential hooking as an alternative model of randomness

So far, we have considered only a uniform model of randomness, in which all the edges of the network G_{n-1} are equally likely candidates to be the latch to progress into G_n . This model stems from the notion of fairness and equal opportunity.

Under a uniform model, each of the two graphs G_1 and G'_1 occurs with probability $\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$ (see Figure 2), continuing the stochastic path started from G_1 , the edge $\{3, 4\}$ has probability $1/11$, like any other edge in G_1 .

In the preferential attachment (hooking) model of randomness, priority is given to edges that are used as latches. The more an edge latches, the higher the probability is that it latches again. This model reflects experience. In an economics setting, preferential hooking represents the Matthew principle, according to which “the rich get richer,” also characterized as “success breeds success.”

To precisely describe this model, we introduce the notion of “thickness” of an edge. If the network edges are physical objects such as electrical wires or pipes, fusing edges increases the thickness of the resultant edge. We define the *thickness of an edge* as 1 plus the number of edges fused with it (the number of times the edge is used for latching). Any edge starts out in G_0 with thickness 1, and its thickness may grow over time. Each fusing into the edge adds 1 to the thickness. Let $T_{n-1}(e)$ be the thickness of an edge e in the graph G_{n-1} . In the *preferential hooking model*, the probability that e is the latch for the next graph G_n is proportional to its thickness. So, if \mathcal{E}_n is the set of edges of G_{n-1} , the probability that e is the latch to build G_n is

$$\frac{T_{n-1}(e)}{\sum_{e' \in \mathcal{E}_{n-1}} T_{n-1}(e')}.$$

Note that the total thickness in the entire graph G_n is $\sum_{e' \in \mathcal{E}_n} T_n(e') = \epsilon_0(n + 1)$.

Under preferential hooking, all the edges of G_0 have thickness 1; the graphs G_1 and G'_1 still occur with probability $\frac{1}{12}$ (see Figure 2). The edge connecting $(0, 3)$ and $(0, 4)$ in G_1 has twice as much probability as any other edge in G_1 to be the latch to build G_2 . Continuing the stochastic path from G_1 , the edge joining $(0, 3)$ to $(0, 4)$ has probability $2/12$, whereas any other edge in G_1 has probability $1/12$ to be the latch to build G_2 .

7.1 The core under preferential hooking

Represent an edge with thickness $t \geq 2$ inside the core with t white balls in an urn, and represent an edge outside the core (of thickness 1) with one blue ball in the urn.

After n draws, let W_n be the number of white balls in the urn to be the latch to build G_n (i.e., the sum of thickness of all edges in the core) and B_n be the number of blue balls in the urn. In the starting graph, all edges are not in the core—the starting urn has ϵ_0 blue balls (edges) and none white. If a white edge in G_{n-1} is used as a latch, we increase its thickness by 1 (we add a white ball, upgrading the edge thickness) and add $\epsilon_0 - 1$ noncore new edges coming from the added seed copy (we add $\epsilon_0 - 1$ blue edges). Alternatively, if a blue edge in G_{n-1} is used as a latch, we increase its thickness by 1 (we add two white balls; the edge joins the core with thickness 2) and add $\epsilon_0 - 2$ noncore new edges (we add $\epsilon_0 - 1$ blue edges coming from the seed, but we lose the latch as noncore edge, accounting for an extra -1). The ball addition matrix is

$$\begin{pmatrix} 1 & \epsilon_0 - 1 \\ 2 & \epsilon_0 - 2 \end{pmatrix}.$$

This again is a Bagchi–Pal urn scheme (Bagchi & Pal, 1985). Note that the total number of balls in the urn after n draws represents the sum of the thickness of all the edges in the graph G_n .

In the limit, we have

$$\frac{W_n - \frac{2\epsilon_0}{\epsilon_0+1} n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{2\epsilon_0(\epsilon_0 - 1)}{(\epsilon_0 + 1)^2(\epsilon_0 + 2)}\right).$$

Let \tilde{C}_n be the size of the core at age n . The number of blue balls is the number of edges outside the core. So, the size of the core is

$$\tilde{C}_n = \epsilon_n - B_n = (\epsilon_0 - 1)n + \epsilon_0 - \left(\sum_{e \in \mathcal{E}_n} T_n(e) - W_n\right).$$

Hence, we have $\tilde{C}_n = W_n - n$, satisfying the following central limit theorem.

Theorem 7.1. *In an edge-hooking network grown under preferential hooking, let \tilde{C}_n be the size of the core at age n . We have*

$$\frac{\tilde{C}_n}{n} \xrightarrow{a.s.} \frac{\epsilon_0(\epsilon_0 - 1)}{\epsilon_0 + 1},$$

and

$$\frac{\tilde{C}_n - \frac{\epsilon_0(\epsilon_0-1)}{\epsilon_0+1} n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\epsilon_0 - 1}{\epsilon_0^2(\epsilon_0 + 1)}\right).$$

Acknowledgment. We thank the anonymous reviewers for their insightful comments that improved the presentation of the paper.

Competing interests. None.

Funding statement. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Data availability statement. Our findings do not rely on any data or code.

Notes

1 Reproduced from Akter, S., Yamamoto, Y., Zope, R. and Baruah, T. (2021). Static dipole polarizabilities of polyacenes using self-interaction-corrected density functional approximations. *Journal of Chemical Physics* **154**, 114305, <https://doi.org/10.1063/5.0041265>, with the permission of AIP Publishing.

2 The graph K_n is the complete graph on n vertices in which any two vertices are joined by an edge. This graph has $\binom{n}{2}$ edges.

References

- Abramowitz, M., & Stegun, I. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Washington, DC: U.S. Department of Commerce, National Bureau of Standards.
- Ageuch, R. (2009). Limit theorems for random triangular urn schemes. *Journal of Applied Probability*, *46*(3), 827–843.
- Akter, S., Yamamoto, Y., Zope, R., & Baruah, T. (2021). Static dipole polarizabilities of polyacenes using self-interaction-corrected density functional approximations. *Journal of Chemical Physics*, *154*(11), 114305. doi: [10.1063/5.0041265](https://doi.org/10.1063/5.0041265).
- Anthony, J. (2008). The larger acenes: Versatile organic semiconductors. *Angewandte Chemie International Edition*, *47*(3), 452–483. doi: [10.1002/anie.200604045](https://doi.org/10.1002/anie.200604045).
- Athreya, K., & Karlin, S. (1968). Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *The Annals of Mathematical Statistics*, *39*(6), 1801–1817.
- Bagchi, A., & Pal, A. (1985). Asymptotic normality in the generalized Pólya-Eggenberger urn model with applications to computer data structures. *SIAM Journal on Algebraic and Discrete Methods*, *6*(3), 394–405.
- Bhamidi, S., Fan, R., Fraiman, N., & Nobel, A. (2021). Community modulated recursive trees and population dependent branching processes. *Random Structures & Algorithms*, *1-32*(2), 201–232. doi: [10.1002/rsa.21027](https://doi.org/10.1002/rsa.21027).
- Bhutani, K., & Khan, B. (2003). A metric on the set of connected simple graphs of given order. *Aequationes Mathematicae*, *66*(3), 232–240.
- Bhutani, K., Kalpathy, R., & Mahmoud, H. (2021). Average measures in polymer graphs. *International Journal of Computer Mathematics: Computer Systems Theory*, *6*(1), 37–53. doi: [10.1080/23799927.2020.1860134](https://doi.org/10.1080/23799927.2020.1860134).
- Chen, C., & Mahmoud, H. (2016). Degree profile of self-similar bipolar networks. *Journal of Applied Probability*, *53*(2), 434–447.
- David, F., & Barton, E. (1962). *Combinatorial chance*. London: Charles Griffin.
- Desmarais, C., & Holmgren, C. (2018). Degree distributions of generalized hooking networks. In *2019 Proceedings of the sixteenth workshop on analytic algorithmics and combinatorics (ANALCO)*.
- Desmarais, C., & Mahmoud, H. (2021). Depths in hooking networks. In *Probability in the Engineering and Informational Sciences* (pp. 1–9). doi: [10.1017/S0269964821000164](https://doi.org/10.1017/S0269964821000164).
- Drmota, M., Gittenberger, B., & Panholzer, A. (2008). The degree distribution of thickened trees. *Discrete Mathematics and Theoretical Computer Science*. In *Proceedings of the fifth colloquium on mathematics and computer science. Proceedings AI*, (pp. 149–162).
- Flajolet, P., Dumas, P., & Puyhaubert, V. (2006). Some exactly solvable models of urn process theory. In *Discrete Mathematics and Theoretical Computer Science AG* (pp. 59–118).
- Gopaladesikan, M., Mahmoud, H., & Ward, M. (2014). Building random trees from blocks. *Probability in the Engineering and Informational Sciences*, *28*(1), 67–81.
- Graham, R., Knuth, D., & Patashnik, O. (1994). *Concrete mathematics*. Reading, Massachusetts: Addison-Wesley.
- Janson, S. (2006). Limit theorems for triangular urn schemes. *Probability Theory and Related Fields*, *134*(3), 417–452.
- Janson, S. (2010). Moments of Gamma type and the Brownian supremum process area. *Probability Surveys*, *7*(none), 1–52.
- Mahmoud, H. (2008). *Pólya Urn models*. Orlando, Florida: Chapman-Hall.
- Mahmoud, H. (2019). Local and global degree profiles of randomly grown self-similar hooking networks under uniform and preferential attachment. *Advances in Applied Mathematics*, *111*, 101930.
- Tönshoff, C., & Bettinger, H. (2021). Pushing the limits of acene chemistry: The recent surge of large acenes. *Chemistry—A European Journal*, *27*(10), 3193–3212. doi: [10.1002/chem.202003112](https://doi.org/10.1002/chem.202003112).
- Trapman, P. (2007). On analytical approaches to epidemics on networks. *Theoretical Population Biology*, *7*(2), 160–173.
- van der Hofstad, R., van Leeuwen, J., & Stegehuis, C. (2018). Mesoscopic scales in hierarchical configuration models. *Stochastic Processes and their Applications*, *128*(12), 4246–4276.
- Zamoshchik, N., Zade, S., & Bendikov, M. (2013). Formation of acene-based polymers: Mechanistic computational study. *The Journal of Organic Chemistry*, *78*(20), 10058–10068. doi: [10.1021/jo4006415](https://doi.org/10.1021/jo4006415).
- Zhang, P., Chen, C., & Mahmoud, H. (2015). Characterization of the moments of balanced triangular urn schemes via an elementary approach. *Statistics and Probability Letters*, *96*, 149–153.