Structural Information in Graphs: Symmetries and Admissible Relabelings*

November 2, 2017

Tomasz Łuczak
Faculty of Math. & Comp. Sci.
Adam Mickiewicz University
61-614 Poznań
Poland
tomasz@amu.edu.pl

Abram Magner
Coordinated Science Lab
U. Illinois at Urbana-Champaign
Urbana, IL 618015
U.S.A.
anmagner@illinois.edu

Wojciech Szpankowski
Dept.Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

We study structural properties of *preferential attachment* graphs (with parameter $m \geq 1$ giving the number of attachment choices that each new vertex makes) which intervene in two complementary algorithmic/statistical/information-theoretic problems involving the information shared between a random graph's labels and its structure: in structural compression, we seek to compactly describe a graph's structure by a bit string, throwing away its label information; in the node arrival order recovery, we seek to recover node labels, given only a graph structure.

In particular, we study the typical size of the automorphism group, as well as some shape parameters (such as the number of linear extensions and height) of the directed version of the graph, which in turn allows us to estimate the typical number of *admissible labeled representatives* of a given graph structure. Our result on the automorphism group positively settles a conjecture to the effect that, provided that $m \geq 3$, preferential attachment graphs are asymmetric with high probability, and completes the characterization of the number of symmetries for a broad range of parameters of the model (i.e., for all fixed m). These results allow us to give an algorithmically efficient, asymptotically optimal algorithm for compression of unlabeled preferential attachment graphs. To show the optimality of our scheme, we also derive new, precise estimates of the Shannon entropy of both the unlabeled and labeled version of the model. Our results also imply inapproximability results for the problem of node arrival order recovery. Finally, we give several new results on the degree sequence of preferential attachment graphs, which may be of independent interest.

Index Terms: graph compression, symmetry, preferential attachment, random graphs

*This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01, and NCN grants 2012/06/A/ST1/00261 and 2013/09/B/ST6/02258.

1 Introduction

The purpose of this paper is to present mathematical results on structural parameters which are fundamental to statistical and information-theoretic problems involving the information shared between the labels and the structure of a random graph. We first describe two such problems, which are in a sense complementary – compression of graph structures, wherein the goal is to *remove* label information to produce a compact description of a graph structure, and recovery of node arrival order in dynamic networks, wherein the goal is to *recover* label information by examining a graph structure – and then explain the structural parameters involved in their analysis, which form the focus of this work. In a nutshell, we study the question, *how much information about the labels of a (random) graph is contained in its structure*?

Removing label information – structural compression: Formally, the labeled graph compression problem is as follows: fix a distribution $\mathbb{G}_n$ on (multi)graphs on n vertices. We would like to exhibit an efficiently computable *source code* [7] $(\mathcal{C}_n, \mathcal{D}_n)$ for $\mathbb{G}_n$, where $\mathcal{C}_n$ is a function mapping graphs in the support of $\mathbb{G}_n$ to bit strings, in such a way as to minimize the expected length of the output bit string when the input is a graph distributed according to $\mathbb{G}_n$, and $\mathcal{D}_n$ inverts $\mathcal{C}_n$ and is efficiently computable. A related problem, and one focus of our paper, seeks to compress graph structures: here, the encoding function $\mathcal{C}_n$ is presented with a multigraph G isomorphic to a sample from $\mathbb{G}_n$, and $\mathcal{D}_n(\mathcal{C}_n(G))$ is only required to be a labeled multigraph isomorphic to G (that is, the labels are "discarded", leaving only the structural information referred to in the title). We again insist on a source code with the minimum possible expected code length (which is given by the Shannon entropy of the distribution on unlabeled graphs induced by $\mathbb{G}_n$, an often non-trivial quantity to estimate; we call this the *structural entropy* of the model).

The structural compression problem is motivated by scenarios in which one only cares to transmit or store information about the isomorphism type of a graph – e.g., its degree sequence, number of occurrences of certain subgraphs, etc. In such scenarios, one *does not* care about labeled graph information, such as the fact that, say, vertex 2 connects to vertex 7. Taking advantage of this fact allows for a more compact description of the relevant information than would result if we naively encoded the labeled graph. More philosophically, structural compression allows to quantify and encode the information contained in the *shape* of graph-structured data.

Inferring label information – node arrival order recovery: A complementary problem, *node arrival order recovery* in a dynamic graph, seeks to recover the labels of nodes of a graph drawn from some distribution, given its structure. The motivation is as follows: many networks in the real world, such as protein interaction and social networks, are constructed dynamically, and it is potentially useful to be able to discover node and edge attributes which correlate with time. Formally, the setting is as follows: a labeled graph G (where node j is thought of as the jth node to be added to the graph) is drawn from a known (generally non-vertex-exchangeable) distribution (such as preferential attachment or duplication-divergence), an unknown permutation π is drawn uniformly at random from $\mathbb{S}_n$ (the symmetric group on n letters), and we are shown $H = \pi(G)$. The task of an estimator is to infer (to the extent that it is possible) π^{-1} from H. For more details on the motivation and on lower bounds, see [11]. In that work, the focus was on lower bounds on the probability of error and on the typical number of inversion errors of any estimator, for a broad class of random graph models. These were phrased in terms of structural parameters which we study in the present work.

Structural properties: A few structural quantities arise in both of the above problems: as we will see, the structural entropy for a broad class of graph models involves the size of the automorphism group of a sampled graph, as well as the typical number of positive-probability *labeled representatives* of a given structure, and the number of positive-probability re-labelings (i.e., permutations) of a sampled graph. The same quantities also give lower bounds on the probability of error and the expected number of inversion errors in the node arrival order recovery problem.

We will focus on the analysis of these quantities for *preferential attachment* graphs[1]. Additional structural properties will arise in the analysis of an asymptotically optimal structural compression algorithm which we will give below.

Our contributions: Succinctly, our contributions in this work are threefold: (i) in the setting of preferential attachment graphs, we analyze several structural parameters (explained more precisely below) which arise in both of the motivating problems above and which may be of independent interest; (ii) we use our structural results to precisely determine the entropies of the preferential attachment distributions on both labeled graphs and their structures, giving the fundamental limits of labeled and structural compression; (iii) we give an efficient, asymptotically optimal structural compression algorithm whose analysis relies on our structural results (an easy optimal labeled graph compression algorithm can be devised, using arithmetic coding).

The structural properties include the typical size of the automorphism group, as well as some structural characteristics of the *directed* version of the graph (e.g., the number of *admissible labeled representatives* of a given graph structure, which is related to the number of *linear extensions* of the directed version, viewed as a partial order). Our result on the automorphism group positively settles a conjecture in [13] to the effect that preferential attachment graphs in which each node makes a sufficiently large number of choices are asymmetric with high probability. This completes the characterization of the number of symmetries for a broad range of parameters of the model: when the number of attachment choices m of each vertex is 1, with high probability, there are many symmetries; when $m = 2$, the probability of symmetry is asymptotically positive; and we show in this work that the only symmetry when $m \geq 3$ is the identity with high probability (see Theorem 2).

Regarding structural characteristics of the directed version of the graph (wherein edges are directed from younger nodes to those older nodes that they choose), we analyze a natural partitioning of the vertices into *layers*, which intervenes in the depth-first search process on the directed graph and in the estimation of the number of admissible labeled representatives of the graph (i.e., the number of isomorphic graphs which could have arisen by preferential attachment): in particular, we show that the order of growth of the number of layers is $\Theta(\log n)$ with high probability (see Theorem 4), and almost all vertices occur within the first few layers (Theorem 3). The result on the number of layers is important for our structural compression algorithm (summarized in Theorem 8). The concentration result allows us to prove that the number of admissible representatives is typically $\exp(n \log n - O(n \log\log n))$ (which should be compared with $n! = e^{n \log n - n + o(n)}$), which intervenes in our derivation of the structural entropy.

We use the above results to provide new, precise estimates of the Shannon entropy of both the labeled and unlabeled models (Theorems 5 and 6).

Finally, in order to obtain our main results, we prove a number of results on the degrees of nodes, as well as on the degree sequence, which may be of interest in other applications.

We provide details of proofs in the appendix. Full proofs can also be found in the journal version [10] of this work. In this conference version, we also present new results on compression algorithms and structural parameters relevant to their analyses.

Prior work: The general connection between structural compression and the automorphism group of a random graph was pointed out in [5] in the case of unlabeled Erdős-Rényi graphs. The relation between the node arrival order recovery problem, automorphisms, and feasible labeled representatives was pointed out in [11] (but we connect the latter quantity to graph compression in the present work).

[1]Preferential attachment models, though they have limitations, were initially devised to produce graphs with power law degree distributions (frequently observed in the aforementioned applications) via a natural mechanism [3] and continue to be well studied.

There has been significant work on compression of labeled graph and tree models in recent years in both the information theory and computer science communities [16, 2, 1, 5, 8, 4]. In the computer science community, the focus has been on algorithmic complexity, and no attempt seems to have been made to compare with or derive fundamental information- theoretic limits. Work in both communities has largely been restricted to labeled graphs or graphs with strong edge independence assumptions. As we show, additional complications arise when the goal is graph structure compression.

We also remark that there have been many extensions of the preferential attachment model (as well as models which adopt completely different mechanisms) to provide better fits for certain aspects of real networks: see, e.g., the web graph model [6]. It is likely that many of our techniques and results adapt to certain parameter ranges of models extending preferential attachment; we restrict to the plain preferential attachment model (which, in any case, continues to be studied), since the analysis in even this case is quite involved, making it a natural first step in the direction of a more comprehensive study of models with more parameters.

2 Main results

We now introduce the model that we consider and formulate the main results.

We say that a multigraph G on vertex set $[n] = \{1, 2, \ldots, n\}$ is *m-left regular* if the only loop of G is at the vertex 1, and each vertex v, $2 \leq v \leq n$, has precisely m neighbours in the set $[v-1]$. The *preferential attachment model* $\mathcal{PA}(m; n)$ is a dynamic model of network growth which gives a probability measure on the set of all m-left regular graphs on n vertices, proposed in [3]. More precisely, for an integer parameter $m \geq 1$ we define the graph $\mathcal{PA}(m; n)$ with vertex set $[n] = \{1, 2, \ldots, n\}$ using recursion on n in the following way: the graph $G_1 \sim \mathcal{PA}(m; 1)$ is a single node with label 1 with m self-edges (these will be the only self-edges in the graph, and we will only count each such edge once in the degree of vertex 1).

Inductively, to obtain a graph $G_{n+1} \sim \mathcal{PA}(m; n+1)$ from G_n, we add vertex $n+1$ and make m random choices (with replacement) $v_1, ..., v_m$ of neighbors in G_n as follows: for each vertex $w \leq n$ (i.e., vertices in G_n),

$$\Pr(v_i = w | G_n, v_1, ..., v_{i-1}) = \frac{\deg_n(w)}{2mn},$$

where throughout the paper we denote by $\deg_n(w)$ the degree of vertex $w \in [n]$ in the graph G_n (in other words, the degree of w after vertex n has made all of its choices). Our proof techniques adapt to tweaks of the model in which multiple edges are not allowed.

For any graph G, we denote by $S(G)$ its unlabeled version (i.e., the equivalence class consisting of all labeled graphs isomorphic to G). Our structural compression/entropy results will be concerned with the unlabeled preferential attachment model, defined by first generating $G \sim \mathcal{PA}(m; n)$, then taking $S(G)$.

2.1 Entropy estimates and structural results

Our first concern will be to derive the fundamental lower bound on the expected code length for compression of unlabeled preferential attachment graphs, as described above. As usual, this is given by the *Shannon entropy* of the distribution on unlabeled graphs induced by $\mathcal{PA}(m; n)$. Recall that for a discrete random variable X with probability mass function $p(\cdot)$, its entropy $H(X)$ is given by $H(X) = -\mathbb{E}_X[\log p(X)]$. We are thus interested in $H(S(G))$, where $G \sim \mathcal{PA}(m; n)$.

By the chain rule for conditional entropy, $H(G) = H(S(G)) + H(G|S(G))$. The second term, $H(G|S(G))$, measures our uncertainty about the labeled graph if we are given its structure. We will give a formula for $H(G|S(G))$ in terms of the automorphism group $|\mathrm{Aut}(G)|$ and another quantity, defined as follows: suppose that, after generating G, we relabel G by drawing a permutation π uniformly at random from $\mathbb{S}_n$, the symmetric group on n letters, and computing $\pi(G)$. Then conditioning on $\pi(G)$ yields a probability distribution for possible values of $\pi^{-1} = \sigma$. We can write $H(G|S(G))$ in terms of $H(\sigma|\sigma^{-1}(G)) = H(\sigma|\sigma(G))$ (intuitively, the amount of uncertainty about the value of the random permutation σ upon seeing the result of its application to G) and $\mathbb{E}[\log|\mathrm{Aut}(G)|]$ using the chain rule for entropy, resulting in the following lemma (which, in fact, is not specific to preferential attachment models).

Lemma 1 (Structural entropy for preferential attachment graphs)**.** *Let $G \sim \mathcal{PA}(m;n)$ for fixed $m \geq 1$, and let σ be a uniformly random permutation from $\mathbb{S}_n$. Then we have*

$$H(G) - H(S(G)) = H(\sigma|\sigma(G)) - \mathbb{E}[\log|\mathrm{Aut}(G)|]. \tag{1}$$

To evaluate $H(S(G))$ and to analyze our compression algorithms, we are thus led to evaluate $\mathbb{E}[\log|\mathrm{Aut}(G)|]$, $H(\sigma|\sigma(G))$, and $H(G)$. The next few results give the structural properties that we need for this. The term $H(\sigma|\sigma(G))$ has multiple interpretations: defining $\Gamma(G)$ to be the set of relabelings of G which produce positive-probability graphs under preferential attachment, we have (at least asymptotically) $H(\sigma|\sigma(G)) = \mathbb{E}[\log|\Gamma(G)|]$. This, in turn, is related to the number of linear extensions of the directed version of G, viewed as a partial order.

Structural results: The proof of Theorem 6 (our expansion of $H(S(G))$) below and the analyses of our algorithms depend on the following structural results.

The next theorem (whose proof we sketch in Section 3 and which we fully prove in the appendix) says that with high probability G has no symmetries, provided that $m \geq 3$. As mentioned in the introduction, this essentially completes the analysis of the precise behavior of the number of symmetries of $\mathcal{PA}(m;n)$ for constant m. For most of this paper, we will focus on the case $m \geq 3$, since the behaviors for $m = 1, 2$ are qualitatively different (for $m = 1, 2$, there are many symmetries with high probability and with asymptotically positive probability, respectively).

Theorem 2 (Asymmetry for preferential attachment model)**.** *Let $G \sim \mathcal{PA}(m;n)$ for fixed $m \geq 3$. Then, with high probability as $n \to \infty$, $|\mathrm{Aut}(G)| = 1$. More precisely, for $m \geq 3$, $\Pr(|\mathrm{Aut}(G)| > 1) = O(n^{-\delta})$, for some fixed $\delta > 0$ and large n.*

We will also state some results on the *directed* version of G (denoted by $\mathrm{DAG}(G)$). This is the directed multigraph defined on $[n]$, with an edge from w to the older node $v < w$ for each edge between v and w in G. We can partition the vertices of $\mathrm{DAG}(G)$ into *levels* inductively as follows: L_1 consists of the vertices with in-degree 0 (i.e., with total degree m). Inductively, L_j is the set of vertices incident on edges coming from vertices in L_{j-1}. Equivalently, a vertex w is an element of some level $\geq j$ if and only if there exist vertices $v_1 < \cdots < v_j$ such with $v_1 > w$ and the path $v_j v_{j-1} \cdots v_1 w$ exists in G. The *height* of $\mathrm{DAG}(G)$ is then defined to be the number of levels in this partition.

The next result (proven in Section 6.2) says that almost all of the vertices are concentrated within the first few levels. This will be instrumental in the proof of Theorem 6.

Theorem 3. *For any $\delta = \delta(n) > 0$, there exists $\ell = \ell(\delta)$ for which the number of vertices that are not in the first ℓ layers of $\mathrm{DAG}(G)$ is at most δn, with high probability. In particular, we can take $\ell \geq \frac{15m}{2\delta^4}\log(3/(2\delta^2))$.*

Next, we find the order of growth of the typical height of $\mathrm{DAG}(G)$, which will be useful in the analysis of our structural compression algorithm. We give the proof in Section 4.

Theorem 4 (Height of $\mathrm{DAG}(G)$)**.** *Consider $G_n \sim \mathcal{PA}(m;n)$ for fixed $m \geq 1$. Then, with probability at least $1 - o(n^{-1})$, the height of $\mathrm{DAG}(G_n)$ is at most $Cm\log n$, for some absolute positive constant C.*

It is simple to show that with high probability the height is also lower bounded by $\Omega(\log n)$.

Using these results, we will be able to connect $H(\sigma|\sigma(G))$ in (1) to a combinatorial parameter of $\mathrm{DAG}(G)$ (the number of *linear extensions* of $\mathrm{DAG}(G)$, viewed as a partial order), which we will be able to show is estimated by $n \log n + R(n)$, where $C_1 n \leq |R(n)| \leq C_2 n \log \log n$.

Entropy results: The final quantity to evaluate in (1) is $H(G)$. Since, in many real applications, n is small enough that $n \log n$ is comparable to n, it is worthwhile (and theoretically interesting) to provide a few terms in the asymptotic expansion of $H(G)$. We give the proof of the following theorem in the appendix.

Theorem 5 (Entropy of preferential attachment graphs)**.** *Consider $G \sim \mathcal{PA}(m;n)$ for fixed $m \geq 1$. We have*

$$H(G) = mn \log n + m\left(\log 2m - 1 - \log m! - A\right) n + o(n), \tag{2}$$

where $A = A(m) = \sum_{d=m}^{\infty} \frac{\log d}{(d+1)(d+2)}$.

This entropy should be compared with the naive method of encoding these graphs, which takes $mn \log(mn) = mn \log n + mn \log m$ space. As $m \to \infty$, compressing to the entropy saves $nm^2 \log m(1 + o_{n,m}(1))$ bits over the naive encoding. For even moderate m (say, $m = 5$ and $n = 10^8$), this is an appreciable difference. This is a more precise analysis than the one given in [14], which only recovers the first term and the order of the second.

Using the above results, we finally have the following expression for $H(S(G))$.

Theorem 6 (Structural entropy of preferential attachment graphs)**.** *Let $m \geq 3$ be fixed. Consider $G \sim \mathcal{PA}(m;n)$. We have*

$$H(S(G)) = (m-1) n \log n + R(n), \tag{3}$$

where $R(n)$ satisfies $Cn \leq |R(n)| \leq O(n \log \log n)$ for some nonzero constant $C = C(m)$.

We sketch the proof of this in Section 5 and complete it in the appendix. Compared with the naive encoding method which simply stores a labeled representative of the structure using $mn \log(mn)$ bits, the structural entropy is smaller by $n \log n(1 + o(1))$ bits.

2.2 Optimal compression algorithms

We established (via a variant of Shannon's source coding theorem) in the previous section that *there exist* source codes for unlabeled and labeled graph compression for $\mathcal{PA}(m;n)$ with expected length within one bit of the entropies (3) and (2), respectively. In this section, we give our results on *efficient algorithms* for compression and decompression of unlabeled/labeled samples from $\mathcal{PA}(m;n)$ which asymptotically achieve these bounds.

First, we give an asymptotically optimal algorithm for compression of unlabeled graphs (see Theorem 8 below): that is, given an arbitrary labeled representative G isomorphic to $G' \sim \mathcal{PA}(m;n)$, we construct a code from which $S(G')$ can be efficiently recovered. We note that the algorithm can be run on general undirected graphs; our optimality guarantee is under the assumption that the input is generated by preferential attachment.

Structural compression algorithm. We first state our algorithm and analyze it in the case where the model is preferential attachment with m self-loops on the oldest vertex. In the appendix (Section 6.5), we explain the (simple) tweaks needed to generalize to the case where there are no self-loops (and hence where one cannot necessarily uniquely identify the oldest vertex).

Our algorithm starts with finding a certain orientation of the edges of the input graph G to produce a directed, acyclic graph D. In the case where G is isomorphic to a sample G' from $\mathcal{PA}(m;n)$ (say, $G = \pi(G')$), we have $D = \pi(\mathrm{DAG}(G'))$, and D is m-left regular.

We accomplish this by a *peeling* procedure: at each step, consider the set D_{min} of minimum-degree nodes in the graph. We orient the edges incident on those nodes away from them, and then recurse on the subgraph excluding the nodes in D_{min}. This procedure terminates precisely when there are no remaining vertices. For a general input graph G, which might not have arisen by preferential attachment, there may be edges between vertices in D_{min}. We orient edges from nodes with larger labels to those with smaller ones. In general, this yields a directed, acyclic graph (aside from self-loops).

That this yields the directed graph $D = \pi(\mathrm{DAG}(G'))$ when the input is isomorphic to a preferential attachment graph is spelled out in detail in Lemma 2 of [12]. Hence, we are free to apply our structural results (such as Theorem 4) on $\mathrm{DAG}(G')$. We remark that it is not too hard to generalize our algorithm to tweaks of the model, since the only thing that is required is that the height of the resulting directed graph be at most $O(\log n)$; such an orientation of the edges of G exists with high probability, because of Theorem 4.

With this procedure in hand, the structural compression algorithm works as follows, on input G:

1. Construct the directed version $D = \mathrm{DAG}(G)$ by the procedure just described.

2. Starting from the "bottom" vertex (i.e., the vertex with no out-edges except for self-loops), we will do a depth-first search of D (following edges only from their destinations to their sources). To the jth vertex in this traversal, for $j = 1, ..., n$, we will associate a *backtracking number* B_j, which tells us how many steps to backtrack in the DFS process after visiting the jth node; e.g., when there is at least one in-edge leading to an unvisited node (so that we do not backtrack), $B_j = 0$.

   Upon visiting vertex w from vertex v in the DFS, we do the following:

   (a) Denote by k the maximum out-degree of D (which can be determined in a preprocessing step, and which is equal to m if the input arises from preferential attachment). Using $\lceil \log k \rceil$ bits, encode the out-degree d_w of w (for preferential attachment, $d_w = m$, but we encode it for the sake of generality).

   Encode the names of the $d_w - 1$ vertex choices made by w, excluding one choice to connect to vertex v. Here, the *name* of a vertex is the binary expansion of its index in the DFS, which we can represent using exactly $\lceil \log n \rceil$ bits. These can be determined in a preprocessing step, by doing an initial DFS to label the nodes with their names.

   (b) We need to know what happens after we visit vertex w: do we go forward in the search, or is there nowhere left to go along the current route (i.e., do we need to backtrack)? Suppose w is the jth vertex to be visited. Then we output an encoding of B_j. We need to more precisely examine how we encode these numbers, since it would be suboptimal to simply encode them in $\Theta(\log n)$ bits. Lemma 7 below tells us how to more efficiently perform this encoding.

3. For the purposes of decoding, we store (once, for the entire graph) the sequence of code words for the code used for the backtracking numbers. This can be done in at most $O(n \log \log n)$ extra bits, at the beginning of the code. We also store k (the maximum out-degree), which can be done with at most $O(\log n)$ bits.

Lemma 7. *The backtracking numbers $B_1, ..., B_n$ can be encoded using a total of $O(n \log \log n)$ bits on average.*

Proof. Consider a random variable X whose distribution is given by the empirical distribution of the collection $B = \{B_1, ..., B_n\}$. That is,

$$P_X(x) = \frac{|\{j \ : \ B_j = x\}|}{n} \tag{4}$$

for each x. Note that this empirical distribution is itself a random variable. We will show that $\mathbb{E}[n \cdot H(X)] = O(n \log\log n)$.

Denote by W the event that the number of levels in D is upper bounded by $O(\log n)$. Under conditioning on this event, X can take on at most $O(\log n)$ values, which implies that $H(X) = O(\log\log n)$. Then we have

$$\begin{aligned}\mathbb{E}[H(X)] &\leq \mathbb{E}[H(X)|W] + (1 - \Pr(W))\mathbb{E}[H(X)|\neg W] \\ &\leq \mathbb{E}[H(X)|W] + (1 - \Pr(W))\log n = O(\log\log n),\end{aligned}$$

where we have used Theorem 4 to upper bound $1 - \Pr(W)$.

We can thus construct a prefix code (once, for the entire graph) for the observed values of B_i, whose empirical average length is given by $\sum_{x \,:\, \exists j, B_j = x} \ell_x P_X(x) \leq H(X) + 1$, where ℓ_x denotes the length of the code word for x. Now, recalling the definition of $P_X(x)$ in (4), this implies

$$\mathbb{E}\left[\sum_{x \,:\, \exists j, B_j = x} \ell_x |\{j \ : \ B_j = x\}|\right] \leq n\mathbb{E}[H(X)] + n = O(n \log\log n).$$

This completes the proof. ∎

The code for $S(G)$ is uniquely decodable, as shown by the decompression algorithm sketched in Section 6.5. Furthermore, its expected length is at most $(m-1)n\log n + O(n\log\log n)$, which recovers the first term of the structural entropy and bounds the second. Let us analyze the running time. Construction of the Huffman code for the backtracking numbers takes time $O(n \log n)$, and each step of the DFS takes time at most $O(m \log n)$. Thus, the running time is $O(mn \log n)$.

∎

We have thus proven the following:

Theorem 8 (Structural compression)**.** *There exists an algorithm (given above) which, on input a graph G isomorphic to $G' \sim \mathcal{PA}(m; n)$, runs in time $O(mn \log n)$ and outputs a code of expected length $(m-1)n\log n + O(n \log\log n)$ from which we can recover $S(G)$ in time $O(mn \log n)$. If self-loops are removed from G' and G (so that the first vertex is not easy to identify), then the same code length can be achieved in time $O(mn^2 \log n)$.*

Note, from Theorem 6, that our algorithm is optimal at least up to the first term of the lower bound, and we explicitly bound the second term.

We note that it is simple to devise an optimal labeled compression algorithm via arithmetic coding. We omit the details.

3 Proof of Theorem 2

We only sketch the proof of the asymmetry result here. The full proof is in the appendix. Let us define first two properties, $\mathfrak{A}$ and $\mathfrak{B}$ of $\mathcal{PA}(m; n)$ which are crucial for our argument. Here and below we set, for convenience, $k = k(n) = n^{\Delta}$ and $\tilde{k} = \tilde{k}(n) = n^{\Delta'}$ for some small enough $0 < \Delta < \Delta'$ to be chosen.

($\mathfrak{A}$) $\mathcal{PA}(m;n)$ has property $\mathfrak{A}$ if no two vertices t_1, t_2, where $k < t_1 < t_2$, are adjacent to the same m neighbors from the set $[t_1 - 1]$.

($\mathfrak{B}$) $\mathcal{PA}(m;n)$ has property $\mathfrak{B}$ if the degree of every vertex $s \le \tilde{k}$ is unique in $\mathcal{PA}(m;n)$, i.e. for no other vertex s' of $\mathcal{PA}(m;n)$ we have $\deg_n(s) = \deg_n(s')$.

It is easy to see that

$$\Pr(|\mathrm{Aut}(\mathcal{PA}(m;n))| = 1) \ge \Pr(\mathcal{PA}(m;n) \in \mathfrak{A} \cap \mathfrak{B}), \tag{5}$$

and so $\Pr(|\mathrm{Aut}(\mathcal{PA}(m;n))| > 1) \le \Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) + \Pr(\mathcal{PA}(m;n) \notin \mathfrak{B})$. Indeed, let us suppose that $\mathcal{PA}(m;n)$ has both properties $\mathfrak{A}$ and $\mathfrak{B}$, and $\sigma \in \mathrm{Aut}(\mathcal{PA}(m;n))$. Let us assume also that σ is not the identity, and let t_1 be the smallest vertex such that $t_2 = \sigma(t_1) \ne t_1$. Note that $\mathfrak{B}$ implies that for all $s \in [k]$ we have $\sigma(s) = s$, so that we must have $k < t_1 < t_2$. On the other hand from $\mathfrak{A}$ it follows that t_1 and $t_2 = \sigma(t_1)$ have different neigbourhoods in the set $[k]$ which consists of fixed point of σ. This contradiction shows that σ is the identity, i.e. $|\mathrm{Aut}(\mathcal{PA}(m;n))| = 1$ which proves (5).

Thus, in order to prove Theorem 2 it is enough to show that both probabilities $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A})$ and $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{B})$ tend to 0 polynomially fast as $n \to \infty$.

Let us study first the property $\mathfrak{A}$. Our task is to estimate from above the probability that there exist vertices t_1 and t_2 such that $k < t_1 < t_2$, which select the same m neighbours (which, of course, belong to $[t_1 - 1]$). Thus we conclude

$$\begin{aligned}\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) &\le \sum_{k<t_1<t_2} \Pr(t_1, t_2 \text{ choose the same neighbours in } [t_1-1]) \\ &\le \sum_{k<t_1<t_2} \sum_{1 \le r_1 \le r_2 \ldots \le r_m < t_1} \Pr(t_1, t_2 \text{ choose } r_1, \ldots, r_m). \end{aligned}\tag{6}$$

The event in the last expression is an intersection of dependent events but, conditioned on the degrees $\deg_{t_\ell}(r_s)$ of the chosen vertices r_s at times t_1, t_2, the choice events become independent.

Let us define $\mathfrak{D}$ as an event that for some $\ell = 1, 2$, and $s = 1, 2, \ldots, m$, $\deg_{t_\ell}(r_s) \le \sqrt{t_\ell / r_s} (\log t_\ell)^3$. Then from Lemma 15 it follows that $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{D}) \le t_1^{-10m/\Delta}$. Conditioning on $\mathfrak{D}$ and further manipulation shows that for $k < t_1 < t_2$ we get

$$\Pr(t_1, t_2 \text{ choose } r_1, \ldots, r_m) \le (\log t_2)^{6m} \prod_{\ell=1}^{2} \prod_{s=1}^{m} \frac{1}{\sqrt{t_\ell r_s}} + n^{-10m}$$

Thus, (6) becomes, again, after some work, $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) \le k^{2-m} (\log k)^{9m} + n^{-1}$. Hence $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) \le n^{\Delta(2.0001-m)}$, which is polynomially decaying since $m \ge 3$. We remark that this holds for arbitrary $\Delta > 0$.

Next we show that, with probability close to 1, the $\tilde{k} = n^{\Delta'}$ oldest vertices of $\mathcal{PA}(m;n)$ have unique degrees and so these are fixed points of every automorphism. The key ingredient of our argument is Lemma 20.

To estimate the probability that $\mathcal{PA}(m;n) \notin \mathfrak{B}$, we reason as follows: from Lemma 20 we know that with probability at least $1 - O(n^{-c})$, for some positive constant c, the degrees of all vertices smaller than $\tilde{k}'^2 = n^{2\Delta''}$ are pairwise different (provided that we choose Δ'' small enough to satisfy Lemma 20).

Furthermore, using Corollary 1, one can deduce that we can choose $\Delta' > 0$ small enough so that, with probability at least $1 - O(n^{-c})$ (for another positive constant $c > 0$) all vertices $s < \tilde{k}$ have degrees larger than those of all vertices $t > \tilde{k}'^2$ (in particular using the left tail bound to show that vertices $< \tilde{k}$ all have high degree and the right tail bound to show that vertices $> \tilde{k}'^2$ have low degree whp). Consequently, with probability $1 - O(n^{-c})$ degrees of vertices from $[\tilde{k}]$ are unique, i.e. $\mathcal{PA}(m;n) \notin \mathfrak{B}$.

Finally, Theorem 2 follows directly from (11) and our estimates for $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A})$ and $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{B})$, provided that we choose $\Delta < \Delta'$.

4 Proof of Theorem 4

Let us start with the following, surprising at first sight, observation.

Fact 9. *Let $w < v$. Then the degree $\deg_v(w)$ as well as the probability that v is adjacent to w does not depend on the structure of the graph induced by the first w vertices.* ∎

Let $p_m(n,k)$ denote the probability that $\mathrm{DAG}(G_n)$ contains a path of length k. From Fact 9 and Corollary 2, it follows that

$$
\begin{aligned}
p_m(n,k) &\le \sum_{v_0<v_1<\cdots<v_k} \prod_{i=1}^{k} \Pr(v_{i-1}\to v_i) \le \sum_{v_0<v_1<\cdots<v_k} \prod_{i=1}^{k} \frac{5m\log(3v_i/v_{i-1})}{\sqrt{v_{i-1}v_i}} \\
&\le \sum_{v_0=1}^{n-k} \frac{1}{\sqrt{v_0}} \prod_{i=1}^{k} \sum_{v_i=v_{i-1}+1}^{n-k-i} \frac{5m\log(3v_i/v_{i-1})}{v_i} \,. \qquad (7)
\end{aligned}
$$

In order to estimate the above sum we split all the vertices $v_1,\dots,v_k$ of the path P into several classes. Namely we say that a vertex v_i is of type t in P if t is the smallest natural number such that $v_i/v_{i-1} \le (1+a)^t$, where a is a small constant to be chosen later, i.e. $t = \lceil \log(v_i/v_{i-1})/\log(1+a)\rceil$. Then, given v_{i-1}, the contribution of terms related to v_i can be estimated from above by

$$
\sum_{v_i=v_{i-1}(1+a)^{t-1}}^{v_{i-1}(1+a)^t} \frac{5m\log(3v_i/v_{i-1})}{v_i} \le 5m\log[(1+a)]\log[3(1+a)^t] \le \alpha t\,, \qquad (8)
$$

where, to simplify notation, we put $\alpha = 5m\log(1+a)\log(3(1+a))$. Let s_t denote the number of vertices of type t in P. Note that $\prod_{t\ge2}\left[(1+a)^{t-1}\right]^{s_t} \le n$ and so

$$
\sum_{t\ge2} ts_t \le 2\sum_{t\ge2}(t-1)s_t \le \frac{2\log n}{\log(1+a)}\,. \qquad (9)
$$

Let us set $J = 2\log n/\log(1+a)$. Thus, we arrive at the following estimate for $p_m(n,k)$

$$
\begin{aligned}
p_m(n,k) &\le \sum_{v_0=1}^{n-k} \frac{1}{\sqrt{v_0}} \binom{k}{s_1} \alpha^{s_1} \sum_{\sum_t s_t t\le J} \binom{k-s_1}{s_2, s_3, \dots, s_k} \prod_{t\ge2}^{k} (\alpha t)^{s_t} \\
&\le 3\sqrt{n}\binom{k}{s_1}\alpha^{s_1} \sum_{\sum_t s_t t\le J} \binom{k-s_1}{s_2, s_3, \dots, s_k} \exp\Big(\sum_{t\ge2} s_t\log(\alpha t)\Big) \\
&\le 3\sqrt{n}\binom{k}{s_1}\alpha^{s_1} 2^{2J} \max_{\sum_t s_t t\le J} \exp\Big(\sum_{t\ge2} s_t \log\Big(\frac{e\alpha t(k-s_1)}{s_t}\Big)\Big)\,.
\end{aligned}
$$

In order to estimate the expression $\sigma(J,S) = \max_{\sum_t s_t t\le J}\exp\Big(\sum_{t\ge2} s_t\log\Big(\frac{e\alpha tS}{s_t}\Big)\Big)$ where $S=\sum_{t\ge2}s_t$, we split the set of all t's into two parts. Thus, let $T_1 = \{t : \log(e\alpha tS/s_t) \le t\}$ and $T_2 = \{2,3,\dots,k\}\setminus T_1$. Then, clearly,

$$
\max_{\sum_t s_t t\le J} \exp\Big(\sum_{t\in T_1} s_t\log\Big(\frac{e\alpha tS}{s_t}\Big)\Big) \le \max_{\sum_t s_t t\le J}\exp\Big(\sum_{t\in T_1} s_t t\Big) \le \exp(J)\,.
$$

Observe that for every $t\in T_2$ we have $\log(eS\alpha t/s_t) \ge t$ and so $s_t \le e\alpha te^{-t}S$. It is easy to check that then $s_t\log\Big(\frac{e\alpha tS}{s_t}\Big) \le 6\cdot 2^{-t}S$, so

$$
\max_{\sum_t s_t t\le J}\exp\Big(\sum_{t\in T_2} s_t\log\Big(\frac{e\alpha tS}{s_t}\Big)\Big) \le \max_{\sum_t s_t t\le J}\exp\Big(6S\sum_{t\in T_2}2^{-t}\Big) \le \exp(3S) \le \exp(3J)\,.
$$

Thus, $\sigma(J,S) \le \exp(4J)$, and, since $s_1 = k - S \ge k - J$,

$$p_m(n,k) \le 3\sqrt{n}\binom{k}{s_1}\alpha^{s_1}2^{2J}\sigma(J,k-s_1) \le 3\sqrt{n}2^k\alpha^{k-J}\exp(6J)$$
$$\le 3\exp(\log n + k + (k-J)\log\alpha + 6J)\,.$$

Since for $0 < a < 1$ we have $a/2 < \log(1+a) < a$, if we set $a = 1/(310m)$, then $\alpha < 1/61$ and $\log\alpha < -4$. Now let us recall that $J = 2\log n/\log(1+a)$ and $k = 5000m\log n > 4J$. Thus,

$$p_m(n,k) \le 3\exp(\log n + k + (k-J)\log\alpha + 6J)$$
$$\le 3\exp(\log n + k - 3k + 3k/2) = \exp(\log n - k/2) = o(n^{-1})\,.$$

■

5 Proof of Theorem 6

We only sketch the derivation of the structural entropy here. The full proof is in the appendix.

We start from Lemma 1. Since Theorem 5 precisely gives $H(G)$, and Theorem 2 implies that $\mathbb{E}[\log|\mathrm{Aut}(G)|] = o(n)$, it remains to estimate $H(\sigma|\sigma(G))$ for a uniformly random $\sigma \in \mathbb{S}_n$. To do this, we show that it can be written in terms of a combinatorial parameter of the directed version of G. To describe it, we make a few (somewhat nontrivial) observations: (i) the probability assigned to any graph g by $\mathcal{PA}(m;n)$ only depends on its unlabeled directed graph structure; (ii) for any unlabeled graph generated by $\mathcal{PA}(m;n)$, there is precisely one positive-probability orientation of the edges (i.e., one unlabeled directed graph structure). The latter is a consequence of the fact that our model starts with a vertex having m self-edges; the full proof does not rely on such small details of the model, and in fact replaces this observation with the more general one that there are at most $2^{O(n)}$ unlabeled directed graphs associated with a given unlabeled, undirected graph.

From observation (ii), we find that $H(G|S(G)) = H(\mathrm{DAG}(G)|S(G)) + H(G|\mathrm{DAG}(G)) = H(G|\mathrm{DAG}(G))$, since $\mathrm{DAG}(G)$ is fully determined by $S(G)$. Then observation (i) says that $H(G|\mathrm{DAG}(G)) = \mathbb{E}[\log|\mathrm{Adm}(G)|]$, where we define $\mathrm{Adm}(G)$ to be the set of labeled graphs isomorphic to G which could have arisen by preferential attachment (we call these the *admissible representatives* of $S(G)$). More formally, a labeled graph could have arisen by preferential attachment if, for any $t \le n$, the subgraph induced by the vertices $\{1,...,t\}$ is such that the degree of vertex t is m.

Now, $|\mathrm{Adm}(G)|$ can be written in terms of $|\mathrm{Aut}(G)|$ and another quantity: $|\Gamma(G)|$, which is the set of permutations $\pi \in \mathbb{S}_n$ such that $\pi(G) \in \mathrm{Adm}(G)$; alternatively, viewing $\mathrm{DAG}(G)$ as a partial order, this is the number of *linear extensions* of $\mathrm{DAG}(G)$. Precisely, we have

$$\mathbb{E}[\log|\mathrm{Adm}(G)|] = \mathbb{E}[\log|\Gamma(G)|] - \mathbb{E}[\log|\mathrm{Aut}(G)|],$$

which implies that $H(\sigma|\sigma(G)) = \mathbb{E}[\log|\Gamma(G)|]$. Thus, to estimate $H(S(G))$, it suffices to estimate $\mathbb{E}[\log|\Gamma(G)|]$. A trivial upper bound is $\mathbb{E}[\log|\Gamma(G)|] \le \log n! = n\log n - n + o(n)$. The lower bound follows by noting that any product of permutations that only permute vertices within levels is a member of $\Gamma(G)$. That is, recalling that L_j denotes the jth level of $\mathrm{DAG}(G)$, $|\Gamma(G)| \ge \prod_{j\ge1}|L_j|!$. It follows from Theorem 3 and some work that this lower bound, in turn, is at least $\exp(n\log n - O(n\log\log n))$. Putting all of this together completes the proof.

References

[1] Emmanuel Abbe. Graph compression: the effect of clusters. In *Proc. of the fifty-fourth annual Allerton conference*, 2016.

[2] Emmanuel Abbe and Sergio Verdú. Compressing data on graphs with clusters. In *Proc. of the International Symposium on Information Theory (to appear)*, 2017.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys*, 74:47–97, 2002.

[4] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi, and Prabhakar Raghavan. On compressing social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 219–228, New York, NY, USA, 2009. ACM.

[5] Yongwook Choi and W. Szpankowski. Compression of graphical structures: Fundamental limits, algorithms, and experiments. *Information Theory, IEEE Transactions on*, 58(2):620–638, 2012.

[6] Colin Cooper and Alan Frieze. A general model of web graphs. *Random Structures & Algorithms*, 22:311–335, 2003.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[8] Payam Delgosha and Venkat Anantharam. Universal lossless compression of graphical data. In *Proc. of the International Symposium on Information Theory (to appear)*, 2017.

[9] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2016.

[10] T. Łuczak, A. Magner, and W. Szpankowski. Asymmetry and structural information in preferential attachment graphs. `http://arxiv.org/abs/1607.04102`.

[11] Abram Magner, Ananth Grama, Jithin Sreedharan, and Wojciech Szpankowski. Recovery of vertex orderings in dynamic graphs. *Proceedings of the IEEE International Symposium on Information Theory (to appear)*, 2017.

[12] Abram Magner, Ananth Grama, Jithin K. Sreedharan, and Wojciech Szpankowski. Times: Temporal information maximally extracted from structure. `http://www.cs.purdue.edu/homes/spa/papers/timesnips2017-subm.pdf`.

[13] Abram Magner, Svante Janson, Giorgos Kollias, and Wojciech Szpankowski. On symmetries of uniform and preferential attachment graphs. *Electronic Journal of Combinatorics*, 21, 2014.

[14] Martin Sauerhoff. On the entropy of models for the web graph. *In submission*, 2016.

[15] Remco van der Hofstad. *Random Graphs and Complex Networks: Volume 1*. Cambridge University Press, 2016.

[16] Jie Zhang, En-Hui Yang, and John C. Kieffer. A universal grammar-based code for lossless compression of binary trees. *IEEE Transactions on Information Theory*, 60(3):1373–1386, 2014.

6 Appendix

6.1 Proof of Theorem 2

In this section we shall give a complete proof of Theorem 2. Let us define first two properties, $\mathfrak{A}$ and $\mathfrak{B}$ of $\mathcal{PA}(m;n)$ which are crucial for our argument. Here and below we set, for convenience, $k = k(n) = n^{\Delta}$ and $\tilde{k} = \tilde{k}(n) = n^{\Delta'}$ for some small enough $0 < \Delta < \Delta'$ to be chosen.

($\mathfrak{A}$) $\mathcal{PA}(m;n)$ has property $\mathfrak{A}$ if no two vertices t_1, t_2, where $k < t_1 < t_2$, are adjacent to the same m neighbors from the set $[t_1 - 1]$.

($\mathfrak{B}$) $\mathcal{PA}(m;n)$ has property $\mathfrak{B}$ if the degree of every vertex $s \le \tilde{k}$ is unique in $\mathcal{PA}(m;n)$, i.e. for no other vertex s' of $\mathcal{PA}(m;n)$ we have $\deg_n(s) = \deg_n(s')$.

It is easy to see that

$$\Pr(|\mathrm{Aut}(\mathcal{PA}(m;n))| = 1) \ge \Pr(\mathcal{PA}(m;n) \in \mathfrak{A} \cap \mathfrak{B}), \tag{10}$$

and so

$$\Pr(|\mathrm{Aut}(\mathcal{PA}(m;n))| > 1) \le \Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) + \Pr(\mathcal{PA}(m;n) \notin \mathfrak{B}). \tag{11}$$

Indeed, let us suppose that $\mathcal{PA}(m;n)$ has both properties $\mathfrak{A}$ and $\mathfrak{B}$, and $\sigma \in \mathrm{Aut}(\mathcal{PA}(m;n))$. Let us assume also that σ is not the identity, and let t_1 be the smallest vertex such that $t_2 = \sigma(t_1) \ne t_1$. Note that $\mathfrak{B}$ implies that for all $s \in [k]$ we have $\sigma(s) = s$, so that we must have $k < t_1 < t_2$. On the other hand from $\mathfrak{A}$ it follows that t_1 and $t_2 = \sigma(t_1)$ have different neigbourhoods in the set $[k]$ which consists of fixed point of σ. This contradiction shows that σ is the identity, i.e. $|\mathrm{Aut}(\mathcal{PA}(m;n))| = 1$ which proves (10).

Thus, in order to prove Theorem 2 it is enough to show that both probabilities $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A})$ and $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{B})$ tend to 0 polynomially fast as $n \to \infty$.

Let us study first the property $\mathfrak{A}$. Our task is to estimate from above the probability that there exist vertices t_1 and t_2 such that $k < t_1 < t_2$, which select the same m neighbours (which, of course, belong to $[t_1 - 1]$). Thus we conclude

$$\begin{aligned}\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) &\le \sum_{k<t_1<t_2} \Pr(t_1, t_2 \text{ choose the same neighbours in } [t_1 - 1]) \\ &\le \sum_{k<t_1<t_2} \sum_{1 \le r_1 \le r_2 \ldots \le r_m < t_1} \Pr(t_1, t_2 \text{ choose } r_1, \ldots, r_m).\end{aligned} \tag{12}$$

The event in the last expression is an intersection of dependent events but, if we condition on the degrees $\deg_{t_\ell}(r_s)$ of the chosen vertices r_s at times t_1, t_2, then the choice events become independent.

Let us define $\mathfrak{D}$ as an event that for some $\ell = 1, 2,$ and $s = 1, 2, \ldots, m$,

$$\deg_{t_\ell}(r_s) \le \sqrt{t_\ell / r_s}(\log t_\ell)^3.$$

Then from Lemma 15 it follows that

$$\Pr(\mathcal{PA}(m;n) \notin \mathfrak{D}) \le t_1^{-10m/\Delta}.$$

Consequently, for $k < t_1 < t_2$ we get

$$\begin{aligned}\Pr(t_1, t_2 \text{ choose } r_1, \ldots, r_m) &\le \Pr(t_1, t_2 \text{ choose } r_1, \ldots, r_m \big| \mathfrak{D}) + \Pr(\neg\mathfrak{D}) \\ &\le \prod_{\ell=1}^{2} \prod_{s=1}^{m} \frac{\sqrt{t_\ell / r_s} \log^3 t_\ell}{2 t_\ell} + t_1^{-10m/\Delta} \\ &\le (\log t_2)^{6m} \prod_{\ell=1}^{2} \prod_{s=1}^{m} \frac{1}{\sqrt{t_\ell r_s}} + n^{-10m}\end{aligned}$$

Thus, (12) becomes

$$\begin{aligned}\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) &\le \sum_{k<t_1<t_2} (\log t_2)^{6m} \sum_{1\le r_1\le r_2\ldots\le r_m<t_1} \prod_{\ell=1}^{2}\prod_{s=1}^{m} \frac{1}{\sqrt{t_\ell r_s}} + n^{-1} \\ &\le \sum_{k<t_1<t_2} (t_1t_2)^{-m/2}(\log t_2)^{6m} \sum_{1\le r_1\le r_2\ldots\le r_m<t_1} \prod_{s=1}^{m}\frac{1}{r_s} + n^{-1} \\ &\le \sum_{k<t_1} t_1^{-m+1}(\log t_1)^{9m} + n^{-1} \\ &\le k^{2-m}(\log k)^{9m} + n^{-1}\end{aligned}$$

Hence

$$\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A}) \le n^{\Delta(2.0001-m)}\,, \tag{13}$$

which is polynomially decaying since $m \ge 3$. We remark that this holds for *arbitrary* $\Delta > 0$.

In this section we show that, with probability close to 1, the $\tilde{k} = n^{\Delta'}$ oldest vertices of $\mathcal{PA}(m;n)$ have unique degrees and so these are fixed points of every automorphism. The key ingredient of our argument is Lemma 20.

To estimate the probability that $\mathcal{PA}(m;n) \notin \mathfrak{B}$, we reason as follows: from Lemma 20 we know that with probability at least $1 - O(n^{-c})$, for some positive constant c, the degrees of all vertices smaller than $\tilde{k}'^2 = n^{2\Delta''}$ are pairwise different (provided that we choose Δ'' small enough to satisfy Lemma 20).

Furthermore, using Corollary 1, one can deduce that we can choose $\Delta' > 0$ small enough so that, with probability at least $1-O(n^{-c})$ (for another positive constant $c > 0$) all vertices $s < \tilde{k}$ have degrees larger than those of all vertices $t > \tilde{k}'^2$ (in particular using the left tail bound to show that vertices $< \tilde{k}$ all have high degree and the right tail bound to show that vertices $> \tilde{k}'^2$ have low degree whp). Consequently, with probability $1 - O(n^{-c})$ degrees of vertices from $[\tilde{k}]$ are unique, i.e. $\mathcal{PA}(m;n) \notin \mathfrak{B}$.

Finally, Theorem 2 follows directly from (11) and our estimates for $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{A})$ and $\Pr(\mathcal{PA}(m;n) \notin \mathfrak{B})$, provided that we choose $\Delta < \Delta'$. ■

6.2 Proof of Theorem 3

We define $X = X(\epsilon, k)$ to be the number of vertices $w > \epsilon n$ that are at level $\ge k$ in DAG(G). In other words, w is counted in X if there exist vertices $v_1 < v_2 < \cdots < v_k$ for which $w < v_1$ and the path $v_k \cdots v_1 w$ exists in DAG(G). We have the following lemma bounding $\mathbb{E}[X]$:

Lemma 10. *For any $\epsilon = \epsilon(n) > 0$, there exists $k = k(\epsilon)$ for which $\mathbb{E}[X(\epsilon, k)] \le \epsilon n$. In particular, we can take any k satisfying $k \ge 15\frac{m}{\epsilon^2}\log(3/\epsilon)$.*

Proof. Suppose that $w > \epsilon n$. We want to upper bound the probability that there exist vertices $v_1 < \cdots < v_k$, with $w < v_1$, such that there is a path $v_k \cdots v_1 w$ in G. Applying Corollary 2, this probability is upper bounded by

$$\binom{n}{k} \cdot \frac{((5m/\epsilon)\log(3/\epsilon))^k}{n^k} \le \frac{e\,((5m/\epsilon)\log(3/\epsilon))^k}{k^k}$$

Now, it is sufficient to show that we can choose k so that this is $\le \epsilon$. In fact, we can choose $k \ge 3 \cdot \frac{5m}{\epsilon^2}\log(3/\epsilon)$. This completes the proof. ■

To complete the proof, we have to extend the above to remove the assumption that vertices are $> \epsilon n$ (i.e., we need to study $Y = Y(k)$, the number of vertices $w \ge 1$ that are at level $\ge k$ in DAG(G)). This is a simple consequence of the above lemma, the fact that $X \le Y \le X + \epsilon n$ with probability 1, and Markov's inequality. This completes the proof. ■

6.3 Proof of Theorem 5

In this section we prove Theorem 5 on the entropy of labeled preferential attachment graphs.

We start by noting that, using the chain rule for entropy, we can write

$$H(G_n) = \sum_{t=1}^{n} H(v_{t+1}|G_t), \tag{14}$$

where we denote by v_{t+1} the multiset of connection choices of vertex $t+1$ (i.e., a value for v_{t+1} takes the form of a multiset of m vertices $< t+1$). This follows because G_n corresponds precisely to exactly one n-tuple $(v_1, v_2, ..., v_n)$ of vertex choice multisets.

To calculate the remaining conditional entropy for each t, we first note that it would be simpler if v_{t+1} were a sequence of vertex choices, rather than a multiset (i.e., an equivalence class of sequences). First, let us denote by $\tilde{v}_{t+1}$ the sequence of m choices made by vertex $t+1$. I.e., $\tilde{v}_{t+1,1}$ is the first choice that it makes, and so on. Then we have the following observation:

$$H(\tilde{v}_{t+1}|G_t) = H(\tilde{v}_{t+1}, v_{t+1}|G_t) = H(v_{t+1}|G_t) + H(\tilde{v}_{t+1}|v_{t+1}, G_t), \tag{15}$$

where the first equality is because v_{t+1} is a deterministic function of $\tilde{v}_{t+1}$, and the second is by the chain rule for conditional entropy. We thus have

$$H(v_{t+1}|G_t) = H(\tilde{v}_{t+1}|G_t) - H(\tilde{v}_{t+1}|v_{t+1}, G_t). \tag{16}$$

The second term on the right-hand side is at most a constant with respect to n, so its total contribution to $H(G_n)$ is at most $O(n)$. We will estimate it precisely later, but will first compute $H(\tilde{v}_{t+1}|G_t)$.

By definition of conditional entropy,

$$H(\tilde{v}_{t+1}|G_t) = \sum_{G \text{ on } t \text{ vertices}} \Pr(G_t = G) H(\tilde{v}_{t+1}|G_t = G).$$

Next, note that, conditioned on $G_t = G$, the m choices that vertex $t+1$ makes are independent and identically distributed. So the remaining conditional entropy is just m times the conditional entropy of a single vertex choice made by $t+1$. Using the definition of entropy (as a sum over all possible vertex choices, from 1 to t) and grouping together terms corresponding to vertices of the same degree (which all have the same conditional probability), we get

$$H(\tilde{v}_{t+1}|G_t) = m \sum_{G} \Pr(G_t = G) \sum_{d=m}^{t} N_d(G) p_{t,d} \log(1/p_{t,d}), \tag{17}$$

where $N_d(G)$ denotes the number of vertices of degree d in the fixed graph G, and we define (using the notation of [14])

$$p_{t,d} = \frac{d}{2mt}.$$

Note that the d sum starts from $d = m$, since m is the minimum possible degree in the graph.

Next, we bring the G sum inside the d sum, and we note that

$$\sum_{G} \Pr(G_t = G) N_d(G) = \mathbb{E}[N_d(G)],$$

which we denote by $\bar{N}_{t,d}$.

Thus, we can express $H(\tilde{v}_{t+1}|G_t)$ as

$$H(\tilde{v}_{t+1}|G_t) = m \sum_{d=m}^{t} \bar{N}_{t,d} p_{t,d} \log(1/p_{t,d}), \tag{18}$$

Plugging this into (14), we get

$$H(G_n) + \sum_{t=1}^{n} H(\tilde{v}_{t+1}|v_{t+1}, G_t) = m \sum_{t=1}^{n} \sum_{d=m}^{t} \bar{N}_{t,d} p_{t,d} \log(1/p_{t,d}). \tag{19}$$

Now, we split the inner sum into two parts:

$$\begin{aligned} H(G_n) + \sum_{t=1}^{n} H(\tilde{v}_{t+1}|v_{t+1}, G_t) &= m \sum_{t=1}^{n} \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \bar{N}_{t,d} p_{t,d} \log(1/p_{t,d}) \\ &\quad + m \sum_{t=1}^{n} \sum_{d=\lfloor t^{1/15} \rfloor + 1}^{t} \bar{N}_{t,d} p_{t,d} \log(1/p_{t,d}). \end{aligned} \tag{20}$$

The first part provides the dominant contribution, of order $\Theta(n \log n)$, and we will show that the second part is $o(n)$, due to the smallness of $\bar{N}_{t,d}$.

Estimating the small d terms: To estimate the contribution of the first sum, we apply Lemma 18 to estimate $\bar{N}_{t,d}$ and we use the definition of $p_{t,d}$:

$$\begin{aligned} &\sum_{t=1}^{n} \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \bar{N}_{t,d} p_{t,d} \log(1/p_{t,d}) \\ &= 2m(m+1) \sum_{t=1}^{n} t \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \frac{d}{d(d+1)(d+2) \cdot 2mt} \log\left(\frac{2mt}{d}\right) + \sum_{t=1}^{n} \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \frac{Cd}{2mt} \log(2mt/d) \\ &= 2m(m+1) \sum_{t=1}^{n} t \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \frac{d}{d(d+1)(d+2) \cdot 2mt} \log\left(\frac{2mt}{d}\right) + o(n) \\ &= (m+1) \sum_{t=1}^{n} \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \frac{(\log t + \log 2m - \log d)}{(d+1)(d+2)} + o(n). \end{aligned} \tag{21}$$

Here, the second sum on the right-hand side of the first equality is the error in approximation incurred by invoking Lemma 18. It is easily seen to be $o(n)$. The final equality is simple algebra.

We can further simplify this expression using the following identity: for any $\alpha > m$,

$$\sum_{d=m}^{\alpha} \frac{1}{(d+1)(d+2)} = \frac{1}{m+1} - O_{\alpha \to \infty}(1/\alpha). \tag{22}$$

This can be seen by expressing the dth term of the sum as its partial fraction decomposition, and then noting cancellations in the resulting expression.

Applying this identity to (21) yields

$$\sum_{t=1}^{n} \sum_{d=m}^{\lfloor t^{1/15} \rfloor} \bar{N}_{t,d} p_{t,d} \log(1/p_{t,d}) = \log n! + (\log 2m - A)\, n + o(n),$$

where we define A as in the statement of Theorem 5. Here, the error term from (22) is captured in the $o(n)$ term.

Upper bounding the large d terms: Our goal is now to show that the second sum of (20), which we denote by E, is $o(n)$.

We apply Lemma 19 to upper bound $\bar{N}_{t,d}$, which yields

$$E \le C \sum_{t=1}^{n} \sum_{d=\lfloor t^{1/15} \rfloor + 1}^{t} \frac{t}{d^3} \cdot \frac{d}{2tm} \log(2tm/d) \le C' \sum_{t=1}^{n} \log t \sum_{d=\lfloor t^{1/15} \rfloor + 1}^{t} d^{-2},$$

where we canceled factors in the numerator and denominator of each term, and we upper bounded the expression inside the logarithm using the fact that $d > \lfloor t^{1/15} \rfloor$.

The inner sum is easily seen to be $O(t^{-1/15})$, so that, finally,

$$E \le C' \sum_{t=1}^{n} t^{-1/15} \log t = o(n),$$

as desired.

We thus end up with

$$\sum_{t=1}^{n} H(\tilde{v}_{t+1} | G_t) = m \log n! + m(\log 2m - A)n + o(n). \tag{23}$$

Estimating $H(\tilde{v}_{t+1}|v_{t+1}, G_t)$: The final step is to estimate the contribution of $H(\tilde{v}_{t+1}|v_{t+1}, G_t)$. Let $\mathcal{C}_t$ denote the set of multisets of m elements coming from $[t]$ having no repeated elements. Then we can write

$$\begin{aligned} H(\tilde{v}_{t+1}|v_{t+1}, G_t) &= \sum_{G, v \in \mathcal{C}_t} \Pr(G_t = G, v_{t+1} = v) H(\tilde{v}_{t+1}|v_{t+1} = v, G_t = G) \\ &+ \sum_{G, v \notin \mathcal{C}_t} \Pr(G_t = G, v_{t+1} = v) H(\tilde{v}_{t+1}|v_{t+1} = v, G_t = G). \end{aligned} \tag{24}$$

The first sum can be estimated as follows: we trivially upper bound

$$H(\tilde{v}_{t+1}|v_{t+1} = v, G_t = G) \le \log m!$$

and take it outside the sum. This gives

$$\begin{aligned} \sum_{G, v \in \mathcal{C}_t} \Pr(G_t = G, v_{t+1} = v) H(\tilde{v}_{t+1}|v_{t+1} = v, G_t = G) &\le \log m! \sum_{G, v \in \mathcal{C}_t} \Pr(G_t = G, v_{t+1} = v) \\ &= \log m! \Pr(v_{t+1} \in \mathcal{C}_t). \end{aligned}$$

Now we can upper bound the remaining probability in this expression by noting that with high probability, the maximum degree in G_t is $\tilde{O}(\sqrt{t})$ [9]. Using this fact, we have, for arbitrarily small fixed $\epsilon > 0$,

$$\begin{aligned} \Pr(v_{t+1} \in \mathcal{C}_t) &= \Pr(v_{t+1} \in \mathcal{C}_t, \text{max. degree of } G_t \le C t^{1/2+\epsilon}) \\ &+ \Pr(v_{t+1} \in \mathcal{C}_t, \text{max. degree of } G_t > C t^{1/2+\epsilon}) \end{aligned} \tag{25}$$

The first term is at most

$$\begin{aligned} \Pr(v_{t+1} \in \mathcal{C}_t, \text{max. degree of } G_t \le C t^{1/2+\epsilon}) &\le 1 - \left(1 - \frac{C t^{1/2+\epsilon}}{2mt}\right)^{m-1} \\ &= 1 - \left(1 - \Theta(t^{-1/2+\epsilon}/m)\right)^{m-1} \\ &= \Theta(t^{-1/2+\epsilon}). \end{aligned}$$

Now, the second term of (25) is at most

$$\begin{aligned}\Pr(v_{t+1} \in \mathcal{C}_t, \text{max. degree of } G_t > Ct^{1/2+\epsilon}) &\leq \Pr(\text{max. degree of } G_t > Ct^{1/2+\epsilon}) \\ &= O(e^{-t^{\epsilon}})\end{aligned}$$

and is thus negligible compared to the first term.

Thus, the first sum in (24) is at most

$$\sum_{G, v \in \mathcal{C}_t} \Pr(G_t = G, v_{t+1} = v) H(\tilde{v}_{t+1} | v_{t+1} = v, G_t = G) = O(t^{-1/2+\epsilon}). \tag{26}$$

We will now show that the second sum in (24), over all multisets v of size m with no repeated elements, is $(1 + o(1)) \log m!$. This is trivial, since vertex $t + 1$ is equally likely to have chosen the elements of v in any order. Thus,

$$H(\tilde{v}_{t+1} | v_{t+1} = v, G_t = G) = \log m!. \tag{27}$$

This implies that

$$\begin{aligned}\sum_{G, v \notin \mathcal{C}_t} \Pr(G_t = G, v_{t+1} = v) H(\tilde{v}_{t+1} | v_{t+1} = v, G_t = G) &= \log m! \cdot \Pr(v_{t+1} \notin \mathcal{C}_t) \\ &= \log m! (1 - O(t^{-1/2+\epsilon})).\end{aligned}$$

Thus,

$$H(\tilde{v}_{t+1} | v_{t+1}, G_t) = \log m! (1 + O(t^{-1/2+\epsilon})).$$

Summing over all t yields a total contribution of

$$-\sum_{t=1}^{n} H(\tilde{v}_{t+1} | v_{t+1}, G_t) = -n \log m! + o(n). \tag{28}$$

Putting everything together: From (19), (23), and (28), we get

$$H(G_n) = mn \log n + m(\log 2m - 1 - A - \log m!) n + o(n), \tag{29}$$

where A is as in the statement of Theorem 5. ∎

6.4 Proof of Theorem 6

We now prove the claimed estimate of the structural entropy.

We first show that the contribution of $\mathbb{E}[\log |\mathrm{Aut}(G)|]$ is negligible (in particular, $o(n)$). From Theorem 2, we immediately have

$$\mathbb{E}[\log |\mathrm{Aut}(G)|] \leq n \log n \cdot n^{-\delta} = o(n).$$

We now move on to estimate $H(\sigma | \sigma(G))$, which we will show to satisfy

$$n \log n - O(n \log \log n) \leq H(\sigma | \sigma(G)) \leq n \log n - n + O(\log n). \tag{30}$$

To go further, we need to define a few sets which will play a role in our derivation. We define the *admissible set* $\mathrm{Adm}(S)$ of a given unlabeled graph S to be the set of all labeled graphs g with $S(g) = S$ such that g could have been generated according to the preferential attachment model with given parameters. That is, denoting by g_t the subgraph of g induced by the vertices

$1, \ldots, t$ for each $t \in [n]$, we have that the degree of vertex t in g_t is exactly m. We can similarly define $\mathrm{Adm}(g) = \mathrm{Adm}(S(g))$. Then, for a graph g, we define $\Gamma(g)$ to be the set of permutations π such that $\pi(g) \in \mathrm{Adm}(g)$. We will also define, for an arbitrary set of graphs B,

$$\mathrm{Adm}_B(g) = \mathrm{Adm}(g) \cap B, \qquad \Gamma_B(g) = \{\pi \ : \ \pi(g) \in \mathrm{Adm}_B(g)\}.$$

For a given graph g, these sets are related by the following formula (the simple proof of this fact is a tweak of that given in [11]):

$$|\mathrm{Adm}_B(g)| = \frac{|\Gamma_B(g)|}{|\mathrm{Aut}(g)|}. \tag{31}$$

We next need to consider some directed graphs associated with G: we start with $\mathrm{DAG}(G)$, which is defined on the same vertex set as G; there is an edge from u to $v < u$ in $\mathrm{DAG}(G)$ if and only if there is an edge between u and v in G (in other words, $\mathrm{DAG}(G)$ is simply the graph G before we remove edge directions). Note that, if we ignore self-loops, $\mathrm{DAG}(G)$ is a directed, acyclic graph.

We denote the *unlabeled* version of $\mathrm{DAG}(G)$ (i.e., the set of all labeled directed graphs with the same structure as $\mathrm{DAG}(G)$) by $\mathrm{UDAG}(G)$. We will also, at times, abuse notation and write $\mathrm{UDAG}(G)$ as the set of all labeled, undirected graphs with the same structure as $\mathrm{UDAG}(G)$ and with labeling consistent with $\mathrm{UDAG}(G)$ as a partial order.

We have the following observations regarding these directed graphs.

Lemma 11. *For any two graphs g_1, g_2 satisfying $\mathrm{UDAG}(g_1) = \mathrm{UDAG}(g_2)$, we have*

$$\Pr(G = g_1) = \Pr(G = g_2).$$

Proof. This can be seen by deriving a formula for the probability assigned to a given graph g by the model and noting that it only depends on the structure and admissibility (a graph is said to be admissible if it is in $\mathrm{Adm}(S)$ for some unlabeled graph S). If g is not admissible, then there exists some $t \in [n]$ such that the degree of vertex t at time t is not equal to m. This has probability 0, so $\Pr(G = g) = 0$.

Now, if g is an admissible graph, then we can write $\Pr(G = g)$ as a product over possible degrees of vertices at time n: let $\deg_g(v)$ denote the degree of vertex v in g. We consider the immediate ancestors (i.e., the parents, the vertices that chose to connect to v) of v in $\mathrm{DAG}(g)$, denoting the number of edges that they supply to v by $d_1(v), \ldots, d_{k(v)}(v)$, where $k(v)$ is the number of parents of v. We also denote by $K_g(v)$ the number of orders in which the parents of v could have arrived in the graph (which is only a function of $\mathrm{UDAG}(g)$. Then we can write $\Pr(G = g)$ as follows:

$$\Pr(G = g) = \frac{\prod_{d \geq m} \prod_{v \ : \ \deg_g(v) = d} K_g(v) \prod_{j=1}^{k_g(v)} \binom{m}{j} (m + d_1(v) + \cdots d_{j-1}(v))^{d_j(v)}}{\prod_{i=1}^{n-1} (2mi)^m}. \tag{32}$$

Here, each factor of the v product corresponds to the sequence of $d - m$ choices to connect to vertex v, which can be ordered in a number of ways determined by the structure of $\mathrm{DAG}(g)$. The innermost product gives the contribution of each such choice. Since this formula is only in terms of the degree sequence of the graph and $\mathrm{UDAG}(g)$, two graphs that are admissible and have the same unlabeled DAG must have the same probability, which completes the proof. ∎

Lemma 12. *Fix an unlabeled graph S on n nodes with $\Pr(S(G) = S) > 0$ with some fixed $m \geq 1$. Then the number of distinct unlabeled directed graphs with undirected structure S is at most $e^{\Theta(n)}$.*

Proof. Observe that the number of edges in S is $\Theta(n)$, as it arises with positive probability from $\mathcal{PA}(m;n)$ and m is fixed.

Then note that each of the $\Theta(n)$ edges may be given one of two orientations, resulting in at most $2^{\Theta(n)}$ distinct directed graphs, which completes the proof. ∎

The next lemma shows that $H(\sigma|\sigma(G))$ may be expressed in terms of the quantities just defined.

Lemma 13. *Fix $m \geq 1$ and consider $G \sim \mathcal{PA}(m;n)$. Let $\sigma \in \mathbb{S}_n$ be a uniformly random permutation. Then*

$$H(\sigma|\sigma(G)) = \mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|] + O(n). \tag{33}$$

Proof. First, we give an alternative representation of $H(\sigma|\sigma(G))$. Recall that $H(G|S(G)) = H(\sigma|\sigma(G)) - \mathbb{E}[\log |\mathrm{Aut}(G)|]$. The plan is to derive an alternative expression for $H(G|S(G))$ as follows: by the chain rule for entropy, we have

$$\begin{aligned} H(G|S(G)) &= H(G, \mathrm{UDAG}(G)|S(G)) \\ &= H(\mathrm{UDAG}(G)|S(G)) + H(G|\mathrm{UDAG}(G)) \\ &= O(n) + H(G|\mathrm{UDAG}(G)). \end{aligned}$$

Here, the last equality is a result of Lemma 12. Now, by Lemma 11, we have

$$H(G|\mathrm{UDAG}(G)) = \mathbb{E}[\log |\mathrm{Adm}_{\mathrm{UDAG}(G)}(G)|] = \mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}|] - \mathbb{E}[\log |\mathrm{Aut}(G)|] + O(n),$$

where the second equality is an application of (31). This completes the proof. ∎

Remark 1. *Note that Lemma 13 is robust to small variations in the model.*

Now, to calculate $H(\sigma|\sigma(G))$, it thus remains to estimate $\mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|]$.

We will lower bound $|\Gamma_{\mathrm{UDAG}(G)}(G)|$ in terms of the sizes of the *levels* of $\mathrm{DAG}(G)$, defined as follows: L_1 consists of the vertices with in-degree 0 (i.e., with total degree m). Inductively, L_j is the set of vertices incident on edges coming from vertices in L_{j-1}. Equivalently, a vertex w is an element of some level $\geq j$ if and only if there exist vertices $v_1 < \cdots < v_j$ such with $v_1 > w$ and the path $v_j v_{j-1} \cdots v_1 w$ exists in G.

Then it is not too hard to see that any product of permutations that only permute vertices within levels is a member of $\Gamma_{\mathrm{UDAG}(G)}(G)$. Thus, we have, with probability 1,

$$|\Gamma_{\mathrm{UDAG}(G)}(G)| \geq \prod_{j\geq 1} |L_j|!.$$

We now use Theorem 3 to finish our lower bound on $\mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|]$. Fix $\epsilon = \frac{1}{\log^2 n}$, so that $\delta = \sqrt{2\epsilon} = \Theta(1/\log n)$, and choose $\ell = \frac{15m}{2\delta^4}\log(3/(2\delta^2))$. Then, defining A to be the event that the number of vertices in layers $> \ell$ is at most $\delta n = \Theta(n/\log n)$, we have

$$\mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|] \geq \mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)| \;\big|\; A](1-\delta).$$

Among the ℓ layers, there are at most $\ell - 1$ that satisfy, say, $|L_i| < \log\log n$, since $\sum_{i=1}^{\ell} |L_i| \geq (1-\delta)n$. So we have the following:

$$\sum_{i=1}^{\ell} \log(|L_i|!) = O(\ell \log\log n \log\log\log n) + \sum_{i\in B}(|L_i|\log|L_i| + O(|L_i|)),$$

where $B = \{i \le \ell \ : \ |L_i| \ge \log\log n\}$, and we used Stirling's formula to estimate the terms $i \in B$.

The sum $\sum_{i\in B} O(|L_i|) = O((1-\delta)n) = O(n)$, so it remains to estimate

$$\sum_{i\in B} |L_i| \log |L_i|.$$

Let $N = \sum_{i\in B} |L_i|$. Then, multiplying and dividing each instance of $|L_i|$ by N in the above expression, it becomes

$$\sum_{i\in B} |L_i| \log |L_i| = N \sum_{i\in B} \frac{|L_i|}{N} \log \frac{|L_i|}{N} + N \sum_{i\in B} \frac{|L_i|}{N} \log N.$$

The first sum is simply $-NH(X)$, where X is a random variable distributed according to the empirical distribution of the vertices on the levels $i \in B$. Since $|B| \le \ell$, we have that $|-NH(X)| \le N \log \ell$. Thus, the first term in the above expression is $O(N \log \ell) = O(n \log\log n)$. Meanwhile, the second term is $N \log N \sum_{i\in B} \frac{|L_i|}{N} = N \log N = n \log n - O(n \log\log n)$. Thus, in total, we have shown

$$\mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|] \ge n \log n - O(n \log\log n). \tag{34}$$

Compare this with the trivial upper bound on $\mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|]$:

$$\mathbb{E}[\log |\Gamma_{\mathrm{UDAG}(G)}(G)|] \le \log n! = n \log n - n + O(\log n). \tag{35}$$

This implies that we have recovered the first term, but there is a gap in our lower and upper bounds on the second term.

6.5 More Details of the Structural Compression Algorithm

Decompression: We next sketch the decompression algorithm. Given a string $S = s_1...s_N$ from the compression algorithm, we produce a labeled graph as follows:

1. Read the prefix of S to recover k, n, and the prefix code for the backtracking numbers. Create a node called 1, with m self-edges. Initialize a stack $U \leftarrow 1$.
2. For $j = 2$ to n,
   (a) Set $x \leftarrow$ the top number on U. Push j onto the stack. Read the next $\lceil \log k \rceil$ bits to recover the out-degree d_j of vertex j. Read the next $(d_j - 1)\lceil \log n \rceil$ bits to recover a list of $d_j - 1$ choices made by vertex j, and append x to this list to produce a list ℓ of d_j vertices. Output $j \to \ell$.
   (b) Read the codeword for the next backtracking number B, and pop U B times.

The worst-case running time is $O(mn \log n)$: the first step takes time at most $O(n \log n)$ (as the backtracking numbers have length at most $\log n$), the total number of pops of the stack from the backtracking steps is n, and it takes time $\Theta(m \log n)$ to reconstruct the m choices made by each vertex in the loop. Thus, the total time taken in the loop is $O(mn \log n)$, as claimed. Since this algorithm produces precisely the adjacency list encoded by the compression algorithm, the output is a graph isomorphic to the original, with the isomorphism being given by the mapping from each vertex to its DFS number.

Case where there are no self-loops: In the model where the self-loops on the first vertex are removed from $G' \sim \mathcal{PA}(m;n)$, the algorithm can be adapted to yield the same optimality guarantee, at the expense of some additional running time. Essentially, the idea is to, for each vertex v in the input graph, treat v as the first vertex by adding m self-edges to v, and run the above algorithm, noting the resulting code length. In the end, we take the code with minimum length. By the above analysis, when the input arises from preferential attachment, with high enough probability there exists some v for which the resulting code has length $(m-1)n\log n + O(n\log\log n)$.

6.6 Results on the Degree Sequence

In this section, we present results on the degree sequence of preferential attachment graphs which we will use in the proofs of our main results in subsequent sections.

First, recall that $\deg_t(s)$ is the degree of a vertex $s < t$ after time t (i.e., after vertex t has made its choices). We also define $\mathrm{dg}_t(s) = \deg_t(s) - m$.

Our first lemma gives a bound on the in-degree of each vertex at any given time. This will give a corollary (Corollary 2) that bounds the probability that two given vertices are adjacent at a given time.

Lemma 14. *For any v, w,*

$$\Pr(\mathrm{dg}_v(w) = d) \le \binom{m+d-1}{m-1}\left(1 - \sqrt{\frac{w}{v}} + O\left(\frac{d}{\sqrt{vw}}\right)\right)^d$$

In particular,

$$\Pr(\deg_v(w) = d) \le (2m+d)^m \exp\left(-\sqrt{\frac{w}{v}}d + O\left(\frac{d^2}{\sqrt{vw}}\right)\right).$$

Proof. We estimate this probability as follows. Below we set $t_{d+1} = mv + 1$.

$$\begin{aligned}
\Pr(\mathrm{dg}_v(w) = d) &\le \sum_{mw < t_1 < t_2 < \cdots < t_d \le mv} \prod_{i=1}^{d} \frac{m+i-1}{2t_i} \prod_{j=t_i+1}^{t_{i+1}-1}\Big(1 - \frac{m+i}{2j}\Big) \\
&\le \sum_{mw < t_1 < t_2 < \cdots < t_d \le mv} \frac{(m+d-1)!}{(m-1)!} \prod_{i=1}^{d} \frac{1+O(d/t_i)}{2t_i} \exp\Big(-\sum_{j=t_i}^{t_{i+1}-1} \frac{i}{2j}\Big) \\
&= \sum_{mw < t_1 < t_2 < \cdots < t_d \le mv} \frac{(m+d-1)!}{(m-1)!} \prod_{i=1}^{d} \frac{1+O(d/t_i)}{2t_i} \exp\Big(-\sum_{j=t_i}^{mv} \frac{1}{2j}\Big) \\
&\le \binom{d+m-1}{m-1}\Big(\sum_{i=mw+1}^{mv} \frac{1+O(d/t_i)}{2t_i} \exp\Big(-\sum_{j=t}^{mv}\frac{1}{2j}\Big)\Big)^d.
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{i=mw+1}^{mv} \frac{1+O(d/t_i)}{2t_i} \exp\Big(-\sum_{j=t_i}^{mv}\frac{1}{2j}\Big) \\
&\qquad \le \sum_{i=mw+1}^{mv} \frac{1+O(d/t_i)}{2t_i} \exp\Big(-\frac{1}{2}\log\frac{mv}{t_i} + O\Big(\frac{1}{t_i}\Big)\Big) \\
&\qquad \le \sum_{i=mw+1}^{mv} \frac{1+O(d/t_i)}{2\sqrt{mvt_i}} \\
&\qquad \le 1 - \sqrt{w/v} + O(d/\sqrt{vw}).
\end{aligned}$$

Thus, the assertion follows. ∎

Recall that for $t > s$, the expectation of $\deg_t(s)$ is $O(\sqrt{t/s})$. We first state a simple tail bound to the right of this expectation, which may be found in [9] (it also is a corollary of Lemma 14):

Lemma 15 (Right tail bound for a vertex degree at a specific time)**.** *Let $r < t$. Then*

$$P[\deg_t(r) \geq Ae^m (t/r)^{1/2} (\log t)^2] = O(t^{-A})$$

for any constant $A > 0$ and any t.

We can prove a similar left tail bound for the random variable $\deg_t(s)$ whenever $s \ll t$, as captured in the following lemma.

Lemma 16 (Degree left tail bound)**.** *Let $v = O(T^{1-\epsilon})$ as $T \to \infty$, for some fixed $\epsilon \in (0, 1/2)$. Then there exist some $C, D > 0$ such that*

$$\Pr\left[\deg_T(v) < C\left(\frac{T}{v}\right)^{(1-\epsilon)^2/(2\pm 0.0001)}\right] \leq e^{-D\epsilon^3 \log(T)} = T^{-D\epsilon^3}. \tag{36}$$

To prove this, we need the following coarser lemma.

Lemma 17. *Let $v < T^{1-\epsilon}$, for some fixed $\epsilon > 0$. Then there exist constants $C, D > 0$ independent of ϵ such that*

$$\Pr(\deg_{vT^\epsilon}(v) < C\epsilon \log T) \leq T^{-D\epsilon} \tag{37}$$

for T sufficiently large.

Proof. We observe the graph at exponentially increasing time steps: for some $\beta > 0$, let $t_0 = v$, $t_j = (1+\beta)^j t_0$, $t_k = (1+\beta)^k t_0 = vT^\epsilon$ (so $k = \frac{\epsilon \log T}{\log(1+\beta)}$). Note that $\deg_{t_0}(v) = \deg_v(v) = m$.

Let us upper bound the probability p_{j+1} that no connection to vertex v is made by any vertex in the subinterval $(t_j, t_{j+1}]$:

$$p_{j+1} \leq \left(1 - \frac{m}{2mt_{j+1}}\right)^{m(t_{j+1}-t_j)} = \left(1 - \frac{1}{2t_{j+1}}\right)^{m\beta t_j}, \tag{38}$$

which is at most some positive constant $\rho = \rho(m\beta)$, uniform in j, satisfying $\rho < 1$. This follows from the inequality $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$. Thus, the total number of connections to vertex v in all subintervals can be stochastically lower bounded by a binomial random variable with parameters $k = \Theta(\epsilon \log T)$ and success probability $\rho(m\beta)$: for any $d \geq 0$,

$$\Pr(\deg_{t_k}(v) - m \geq d) \geq \Pr(\text{Binomial}(k, \rho) \geq d). \tag{39}$$

In particular, as $T \to \infty$, this implies (using the Chernoff bound) that with probability $1 - T^{-D\epsilon}$, the number of subintervals which contribute at least one new edge to v is at least $C\epsilon \log T$, for some C, so that $\deg_{vT^\epsilon}(v) \geq C\epsilon \log T$, which completes the proof. ∎

With the previous lemma in hand, we are now ready to prove our left tail bound.

Proof of Lemma 16. Similar to the proof of Lemma 17, we observe the graph at exponentially increasing times: fix a small $\alpha > 0$, and let $t_0 = vT^\epsilon$, $t_j = (1+\alpha)t_0$, $t_k = (1+\alpha)^k t_0 = T$, so that $k = \frac{\log(T/t_0)}{\log(1+\alpha)}$. Denote by $d_j = \deg_{t_j}(v)$ and $\Delta_{j+1} = d_{j+1} - d_j$, for each j.

In the interval $(t_j, t_{j+1}]$, conditioned on the graph up to time t_j, Δ_{j+1} is stochastically lower bounded by a binomially distributed random variable with parameters $t_{j+1} - t_j = \alpha t_j$ and

$p_{j+1} = \frac{md_j}{2mt_{j+1}} = \frac{d_j}{2t_{j+1}}$. The former parameter is simply the interval length (in terms of number of vertices). The latter parameter comes from the fact that the degree of v at any point in the interval is at least d_j, and the total degree of the graph is at most t_{j+1}. I.e.,

$$\Delta_{j+1} \big| G_{t_j} \succeq_{st} \text{Binomial}\left(\alpha t_j, \frac{d_j}{2t_{j+1}}\right), \tag{40}$$

where $\succeq_{st}$ denotes stochastic domination.

This suggests that we define the *bad* event $B_j = [\Delta_j < \alpha t_{j-1} p_j (1-\epsilon)]$, for arbitrary $\epsilon > 0$, and for $j \in [1, k]$. We further define $B_0 = [d_0 < C\epsilon \log T]$, for some constant $C > 0$.

Conditioning on all of the B_j (for $j \in \{0, ..., k\}$) failing to hold, we have

$$\Pr\left[\bigcap_{j<k}\left[d_{j+1} \geq d_j\left(1 + \frac{(1-\epsilon)\alpha}{2(1+\alpha)}\right)\right] \mid \bigcap_{j=0}^{k} \neg B_j\right] = 1, \tag{41}$$

recalling that $d_{j+1} = d_j + \Delta_j$ by definition. This in particular implies that (still under the same conditioning)

$$d_k \geq d_0 \cdot \left(1 + \frac{(1-\epsilon)\alpha}{2(1+\alpha)}\right)^k = d_0 \exp\left(\log(T/t)\frac{\log(1 + \frac{(1-\epsilon)\alpha}{2(1+\alpha)})}{\log(1+\alpha)}\right). \tag{42}$$

Taking α close enough to 0, this becomes

$$d_k \geq d_0 \exp\left(\frac{1-\epsilon}{2}\log(T/t)\right) = d_0 (T/t)^{\frac{1-\epsilon}{2 \pm 0.0001}}, \tag{43}$$

as in the statement of the lemma.

Now, it remains to lower bound the probability $\Pr(\bigcap_{j=0}^{k} \neg B_j)$. We may write it as

$$\Pr(\bigcap_{j=0}^{k} \neg B_j) = \Pr(\neg B_0) \prod_{j=1}^{k} \Pr(\neg B_j | \neg B_0, ..., \neg B_{j-1}) \geq (1 - T^{-D\epsilon}) \prod_{j=1}^{k} \Pr(\neg B_j | \neg B_0, ..., \neg B_{j-1}),$$

where the inequality is by Lemma 17.

Now, by the stochastic domination (40), the conditioning, and the Chernoff bound, the jth factor of the product is lower bounded as follows:

$$\Pr(\neg B_j | \neg B_0, ..., \neg B_{j-1}) \geq \Pr(\text{Binomial}(\alpha t_{j-1}, p_j) \geq \alpha t_{j-1} p_j (1-\epsilon) | \neg B_0, ..., \neg B_{j-1}) \tag{44}$$

$$\geq 1 - \exp\left(-\frac{\epsilon^2 \alpha d_j}{2(1+\alpha)}\right). \tag{45}$$

Under the conditioning, d_j is further lower bounded by $\left(1 + \frac{(1-\epsilon)\alpha}{2(1+\alpha)}\right)^j C\epsilon \log T \geq \left(1 + \frac{\alpha}{4(1+\alpha)}\right)^j C\epsilon \log T$ (using the fact that $\epsilon < 1/2$), resulting in

$$\Pr(\neg B_j | \neg B_0, ..., \neg B_{j-1}) \geq 1 - \exp\left(-C\frac{\epsilon^3 \alpha}{2(1+\alpha)} \cdot \left(1 + \frac{\alpha}{4(1+\alpha)}\right)^j \log(T)\right). \tag{46}$$

This implies

$$\Pr(\bigcap_{j=0}^{k} \neg B_j) \geq \Pr(\neg B_0) \cdot \prod_{j=1}^{k}(1 - \exp\left(-C\frac{\epsilon^3 \alpha}{2(1+\alpha)} \cdot \left(1 + \frac{\alpha}{4(1+\alpha)}\right)^j \log(T)\right)). \tag{47}$$

For convenience, set $C' = C\frac{\alpha}{2(1+\alpha)}$ and $D' = 1 + \frac{\alpha}{4(1+\alpha)}$. Note that $D' > 1$. So the product in (47) can be written (after some simple asymptotic analysis) as

$$\prod_{j=1}^{k} \left(1 - \exp\left(-\epsilon^3 C' \cdot D'^j \log(T)\right)\right) = 1 - \Theta(T^{-\epsilon^3 C' D'}).$$

This implies, after combination with the lower bound on $\Pr(\neg B_0)$, that we can write

$$\Pr(\bigcap_{j=0}^{k} \neg B_j) \geq (1 - T^{-D\epsilon})(1 - \Theta(-T^{-\epsilon^3 C' D'})) \geq 1 - T^{-D''\epsilon^3}, \tag{48}$$

for some $D'' > 0$, as claimed. ∎

Using Lemma 16, we can prove a corollary roughly lower bounding the typical minimum degree of the collection of vertices before a given time.

Corollary 1. *Let $\Delta > 0$ be fixed. There exists some small enough $\delta > 0$ and positive constant D such that*

$$\Pr\left[\bigcup_{w < T^\delta} \deg_T(w) < C\left(T^{1-\Delta}\right)^{1/2}\right] \leq T^{-D} \tag{49}$$

as $T \to \infty$.

Proof. This follows immediately from the fact that the probability bound in Lemma 16 is monotone in ϵ and constant with respect to v. We omit the simple details. ∎

Next, we give a lemma on the expected number of vertices of degree d at time t. We denote this quantity by $\bar{N}_{t,d}$ and the random variable itself by $N_{t,d}$. We start by recalling an approximation result on this quantity [15].

Lemma 18 (Expected value of $N_{t,d}$). *We have, for $t \geq 1$ and $1 \leq d \leq t$ and for any fixed $m \geq 1$,*

$$\left|\bar{N}_{t,d} - \frac{2m(m+1)t}{d(d+1)(d+2)}\right| \leq C,$$

for some fixed $C = C(m) > 0$.

This approximation is useful whenever $d = o(t^{1/3})$. For larger d, the error term C dominates. For our proofs, we need to extend this result for larger d as $t \to \infty$. We have the following result along these lines.

Lemma 19 (Upper bound on $\bar{N}_{t,d}$). *We have, for $t \to \infty$, $d \geq t^{1/15}$, and fixed $m \geq 1$,*

$$\bar{N}_{t,d} = O\left(\frac{t}{d(d+1)(d+2)}\right) = O\left(\frac{t}{d^3}\right). \tag{50}$$

Proof. We will prove the claimed upper bound by induction on the number of edge connection choices made so far in the graph (e.g., after vertex t has made all of its choices, this number is mt).

Let us define $\bar{M}_{\tau,d}$ to be the expected number of vertices with degree d in the graph after τ vertex choices have been made in the graph. Note that $\bar{M}_{\tau,d} = \bar{N}_{\tau/m,d}$ whenever τ is divisible by m. Thus, to prove our desired result, it is sufficient to prove that

$$\bar{M}_{\tau,d} = O\left(\frac{\{\tau\}_m}{d(d+1)(d+2)}\right) \tag{51}$$

for $\tau \to \infty$ and $d \geq (\tau/m)^{1/15}$ (for convenience, we denote by $\{\tau\}_m$ the largest integer $\leq \tau$ that is divisible by m). The base case is provided by Lemma 18.

Next, note that $\bar{M}_{\tau,d}$ satisfies the following recurrence:

$$\begin{aligned}\bar{M}_{\tau,d} &\leq \bar{M}_{\tau-1,d}\left(1 - \frac{d-m}{2\{\tau\}_m}\right) + \bar{M}_{\tau-1,d-1}\frac{d-1}{2\{\tau\}_m} - \bar{M}_{\tau-1,d}\frac{d-m}{2\{\tau\}_m} \\ &= \bar{M}_{\tau-1,d}\left(1 - \frac{d-m}{\{\tau\}_m}\right) + \bar{M}_{\tau-1,d-1}\frac{d-1}{2\{\tau\}_m}. \end{aligned} \tag{52}$$

This is because an m-tuple that has degree d after choice τ either had degree d after choice $\tau-1$ and wasn't chosen by the τth choice, or had degree $d-1$ and was chosen by choice τ. Moreover, any m-tuple with degree d at time $\tau-1$ that *was* chosen by choice τ no longer has degree d. The upper bound is a result of the specific details of our model but may be generalized.

Next, we apply the inductive hypothesis, resulting in

$$\bar{M}_{\tau,d} \leq \frac{C\{\tau-1\}_m}{d(d+1)(d+2)}\left(1 - \frac{d-m}{\{\tau\}_m}\right) + \frac{C\{\tau\}_m}{(d-1)d(d+1)}\frac{d-1}{2\{\tau\}_m} \tag{53}$$

$$\leq \frac{C\{\tau-1\}_m}{d(d+1)(d+2)}\left(1 - \frac{d-m}{\{\tau\}_m}\right) + \frac{C}{2d(d+1)}, \tag{54}$$

for some positive constant $C(m) = C$. This can be rearranged to yield

$$\bar{M}_{\tau,d} \leq \frac{C\{\tau-1\}_m}{d(d+1)(d+2)} + \frac{C}{2d(d+1)} - \frac{C\{\tau-1\}_m(d-m)}{d(d+1)(d+2)\{\tau\}_m}. \tag{55}$$

To continue, we split into two cases: either $\{\tau-1\}_m = \{\tau\}_m$ or $\{\tau-1\}_m = \tau - m = \{\tau\}_m - m$. In the first case, (55) becomes

$$\bar{M}_{\tau,d} \leq \frac{C\{\tau\}_m}{d(d+1)(d+2)} + \frac{C}{2d(d+1)} - \frac{C}{(d+1)(d+2)} + \frac{Cm}{d(d+1)(d+2)}.$$

Now, provided that τ is large enough, and since d is $\Omega(\tau^{1/15})$, the sum of the last three factors is negative, so that

$$\bar{M}_{\tau,d} \leq \frac{C\{\tau\}_m}{d(d+1)(d+2)},$$

as desired.

Now we handle the second case (where $\tau - 1_m = \{\tau\}_m - m$):

$$\begin{aligned}\bar{M}_{\tau,d} &\leq \frac{C\{\tau\}_m - Cm}{d(d+1)(d+2)} + \frac{C}{2d(d+1)} - \frac{C(\{\tau\}_m - m)(d-m)}{d(d+1)(d+2)\{\tau\}_m} \\ &= \frac{C\{\tau\}_m}{d(d+1)(d+2)} + \frac{C}{2d(d+1)} - \frac{Cd}{d(d+1)(d+2)} + \frac{Cm(d-m)}{d(d+1)(d+2)\{\tau\}_m} \\ &\leq \frac{C\{\tau\}_m}{d(d+1)(d+2)} + \frac{C}{2d(d+1)} - \frac{Cd}{d(d+1)(d+2)} + \frac{Cmd}{d(d+1)(d+2)\{\tau\}_m}.\end{aligned}$$

We then proceed exactly as in the previous case, which completes the proof. ∎

The next result, a corollary of Lemma 14, gives an upper bound on the probability that two given vertices are adjacent.

Corollary 2. *Let $w < v$. Then the probability that v is adjacent to w is bounded above by $5m\sqrt{1/(vw)}\log(3v/w)$. In particular, each two vertices $v, w \geq \epsilon n$ are adjacent with probability smaller than $(5m/\epsilon)\log(3/\epsilon)/n$.*

Proof. The probability that v and w are adjacent is bounded from above by

$$\sum_{d\ge 0}\frac{md}{2mv}\Pr(\mathrm{dg}_v(w)=d-m)\,.$$

When $d \le d_0 = 8m\sqrt{v/w}\log(3v/w)$ the above sum is clearly smaller $d_0/2 = 4m\sqrt{1/vw}\log(3v/w)$. If $d \ge d_0$ one can use Lemma 14 to estimate this sum by $m\sqrt{1/vw}\log(3v/w)$. ∎

The next result gives a bound on the probability that two early vertices have the same degree.

Lemma 20. *There exist positive constants $\Delta < 1$ and c such that the probability that for some $s < s' < k^2 = n^{2\Delta}$ we have $\deg_n(s) = \deg_n(s')$ is $O(n^{-c})$.*

Proof. Let $s < s' < k^2 = n^{2\Delta}$, for some $\Delta > 0$ to be chosen. We first estimate the probability that $\deg_n(s) = \deg_n(s')$. In order to do so we set $n' = n^{0.6}$ and define

$$\underline{\deg}(s) = \deg_{n-n'}(s) \quad\text{ and }\quad \underline{\underline{\deg}}(s) = \deg_n(s) - \underline{\deg}(s)\,.$$

Note that

$$\begin{aligned}
\Pr(\deg_n(s)=\deg_n(s')) &= \sum_{\underline{d},\underline{d}',\underline{\underline{d}}'}\Pr(\deg_n(s)=\deg_n(s')|\underline{\deg}(s)=\underline{d},\underline{\deg}(s')=\underline{d}',\underline{\underline{\deg}}(s')=\underline{\underline{d}}')\\
&\qquad\times\Pr(\underline{\deg}(s)=\underline{d},\underline{\deg}(s')=\underline{d}',\underline{\underline{\deg}}(s')=\underline{\underline{d}}')\\
&= \sum_{\underline{d},\underline{d}',\underline{\underline{d}}'}\Pr(\underline{\underline{\deg}}(s)=\underline{d}'+\underline{\underline{d}}'-\underline{d}|\underline{\deg}(s)=\underline{d},\underline{\deg}(s')=\underline{d}',\underline{\underline{\deg}}(s')=\underline{\underline{d}}')\\
&\qquad\times\Pr(\underline{\deg}(s)=\underline{d},\underline{\deg}(s')=\underline{d}',\underline{\underline{\deg}}(s')=\underline{\underline{d}}')\,. \qquad (56)
\end{aligned}$$

Observe that due to Lemma 16 (alternatively, Corollary 1) and Lemma 15, with probability $1-O(n^{-c})$, for some appropriate $c > 0$ and small enough $k = n^{\Delta}$, a vertex $s \in [k^2]$ has degree between $n^{0.488}$ and $n^{0.51}$ at any time in the interval $[n-n', n]$. Importantly, note that if this holds with probability $1-O(n^{-c})$ for a given choice of Δ, then the same holds for all smaller choices of Δ, with the same value for c (this is a consequence of the fact that the probability bound in Lemma 16 is a function of ϵ and not of v).

Furthermore, one can estimate the random variable $\underline{\underline{\deg}}(s)$ conditioned on $\underline{\deg}(s) = \underline{d}$ from above and below by binomial distributed random variables and use Chernoff bound to show that with probability at least $1-O(n^{-c})$ we have

$$\Big|\frac{\underline{d}n'}{2mn} - \underline{\underline{\deg}}(s)\Big| = \Big|0.5m\underline{d}n^{-0.4} - \underline{\underline{\deg}}(s)\Big| \le \Big(\frac{\underline{d}n'}{2mn}\Big)^{0.6} \le n^{0.08}\,. \qquad (57)$$

Thus, in order to estimate $\Pr(\deg_n(s) = \deg_n(s'))$, it is enough to bound

$$\rho(\underline{d}',\underline{\underline{d}}',\underline{d}) = \Pr(\underline{\underline{\deg}}(s)=\underline{d}'+\underline{\underline{d}}'-\underline{d}|\underline{\deg}(s)=\underline{d},\underline{\deg}(s')=\underline{d}',\underline{\underline{\deg}}(s')=\underline{\underline{d}}')$$

for $n^{0.488} \le \underline{d}, \underline{d}' \le n^{0.51}$ and

$$|0.5\underline{d}n^{-0.4}/m - (\underline{d}'+\underline{\underline{d}}'-\underline{d})| \le n^{0.08}\,.$$

In order to simplify the notation set $\ell = \underline{d}' + \underline{\underline{d}}' - \underline{d}$. Let us estimate the probability that $\underline{\underline{\deg}}(s) = \ell$ conditioned on $\underline{\deg}(s) = \underline{d}$ and $\underline{\deg}(s') = \underline{d}'$. The probability that some vertex $v > n-n'$ is connected to s by more than one edge is bounded from above by

$$Cn'\Big(\frac{m\deg_n(s)}{n-n'}\Big)^2 \le n^{0.6}\,O(n^{-0.98}) = O(n^{-0.38})$$

so we can omit this case in further analysis. The probability that we connect a given vertex $v > n - n'$ with s is given by

$$\frac{m \deg_{v-1}(s)}{2m(v-1)} = \frac{\underline{d} + O(\underline{d}n^{-0.4})}{2(n - O(n'))} = \frac{\underline{d}}{2n}\Big(1 + O(n^{-0.4})\Big). \tag{58}$$

Consequently, the probability that $\underline{\underline{\deg}}(s) = \ell$ conditioned on $\underline{\deg}(s) = \underline{d}$ and $\underline{\deg}(s') = \underline{d}'$ is given by

$$\binom{n'}{\ell}\rho^{\ell}(1-\rho)^{n'-\ell}\Big(1 + O(n^{-0.4})\Big)^{\ell}\Big(1 + O(n^{-0.4}\underline{d}/n)\Big)^{n'-\ell},$$

where $\rho = \underline{d}/2n$.

If we additionally condition on the fact that $\underline{\underline{\deg}}(s') = \underline{\underline{d}}'$ (so that we now have conditioned on $\underline{\deg}(s) = \underline{d}, \underline{\deg}(s') = \underline{d}'$, and $\underline{\underline{\deg}}(s') = \underline{\underline{d}}'$), it will result in an extra factor of the order $\Big(1 + O(\underline{d}/2n)\Big)^{\underline{\underline{d}}'}$ since it means that some $\underline{\underline{d}}'$ vertices already made their choice (and selected s' as their neighbour). Note however that, since $\ell, \underline{\underline{d}}' = O(\underline{d}n'/n) = O(n^{0.11})$ we have

$$\begin{aligned}
\Big(1 + O(n^{-0.4})\Big)^{\ell} &= 1 + O(n^{-0.29}) \\
\Big(1 + O(n^{-0.4}\underline{d}/n)\Big)^{n'-\ell} &= 1 + O(n^{-0.29}) \\
\Big(1 + O(\underline{d}/2n)\Big)^{\underline{\underline{d}}'} &= 1 + O(n^{-0.48})\,.
\end{aligned}$$

Hence, the probability that $\underline{\underline{\deg}}(s) = \ell$ conditioned on $\underline{\deg}(s) = \underline{d}$, $\underline{\deg}(s') = \underline{d}'$, and $\underline{\underline{\deg}}(s') = \underline{\underline{d}}'$ is given by

$$\binom{n'}{\ell}\rho^{\ell}(1-\rho)^{n'-\ell}\Big(1 + O(n^{-0.29})\Big)\,,$$

and so it is well approximated by the binomial distribution. On the other hand, the probability that the random variable with binomial distribution with parameters n' and ρ takes a particular value is bounded from above by $O(1/\sqrt{n'\rho})$. Thus, for a given pair of vertices $s < s' < k^2 = n^{2\Delta}$ we have

$$\Pr(\deg_n(s) = \deg_n(s')) = O(\sqrt{n/n'\underline{d}}) + O(n^{-c}) = O(n^{-c})\,.$$

Hence, the probability that such a pair of vertices, $s < s' < k^2 = n^{2\Delta}$ exists is bounded from above by $O(k^4 n^{-c})$, and, as remarked at the beginning of the proof, $k = n^{\Delta}$ may be chosen small enough so that this yields a bound of the form $O(n^{-c'})$, for $c' > 0$. ∎