Algorithmica (2020) 82:2687–2707
https://doi.org/10.1007/s00453-020-00699-2

# Compression of Dynamic Graphs Generated by a Duplication Model

**Krzysztof Turowski[1,3] · Abram Magner[2] · Wojciech Szpankowski[1]**

Received: 9 October 2018 / Accepted: 12 March 2020 / Published online: 3 April 2020


**Abstract**
We continue building up the information theory of non-sequential data structures such as trees, sets, and graphs. In this paper, we consider dynamic graphs generated by a full duplication model in which a new vertex selects an existing vertex and copies all of its neighbors. We ask how many bits are needed to describe the labeled and unlabeled versions of such graphs. We first estimate entropies of both versions and then present asymptotically optimal compression algorithms up to two bits. Interestingly, for the full duplication model the labeled version needs $\Theta(n)$ bits while its unlabeled version (structure) can be described by $\Theta(\log n)$ bits due to significant amount of symmetry (i.e. large average size of the automorphism group of sample graphs).

**Keywords** Random graphs · Structural entropy · Graph compression · Duplication model

## 1 Introduction

Complex systems can often be modeled as dynamic graphs. In these systems, patterns of interactions evolve in time, determining emergent properties, associated function, robustness, and security of the system. There are several broad questions whose answers shed light on the evolution of such dynamic networks: (i) how many bits are required to best describe such a network and its structure (i.e., unlabeled

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grant CCF-1524312, and National Science Center, Poland, under Grant UMO-2016/21/B/ST6/03146.

✉ Krzysztof Turowski
krzysztof.szymon.turowski@gmail.com

[1] Center for Science of Information, Purdue University, West Lafayette, IN, USA

[2] Department of Computer Science, University at Albany, SUNY, Albany, NY, USA

[3] Theoretical Computer Science Department, Jagiellonian University, Krakow, Poland

underlying graph); (ii) how to infer underlying dynamic processes governing network evolution; (iii) how to infer information about previous states of the network; and (iv) how to predict the forward evolution of the network state. In this paper we deal with the first question (i.e., labeled and unlabeled graph compression).

To better understand the evolution of network structural properties, several probabilistic models have been proposed, including, e.g., the preferential attachment, duplication-divergence, Cooper-Frieze, and fit-get richer models [2, 6, 10, 24].

Clearly, some models are more suitable to certain types of data than others. For example, it has been claimed that the preferential attachment mechanism [2] plays a strong role in the formation of citation networks [23]. However, due to the high power law exponent of their degree sequence (greater than 2) and lack of community structure [6], preferential attachment graphs are not likely to describe well biological networks such as protein interaction networks or gene regulatory networks [19]. For such networks another model, known as the vertex-copying model, or simply the *duplication model*, has been claimed as a better fit [25]. In the vertex-copying model, one picks an existing vertex and inserts its clone, possibly with some random modifications, depending on the exact variation of the model [6, 14, 20]. Experimental results show that these variations on the duplication model better capture salient features of protein interaction networks than does the preferential attachment model [22].

In this paper we present comprehensive information-theoretic results for the full duplication model in which every new vertex is a copy of some older vertex. We establish precisely (that is, within a $o(1)$ additive error) the entropy for both unlabeled and labeled graphs generated by this model and design asymptotically optimal compression algorithms that match the entropies up to a constant additive term. Interestingly, we shall see that the entropy of labeled graphs is $H(G_n) = \Theta(n)$, while the structural entropy (the entropy of the isomorphism class of a random graph from the model, denoted by $S(G_n)$) is significantly smaller: $H(S(G_n)) = \Theta(\log n)$. Thus, the vast majority of information of the labeled graphs in this model is present in the labeling itself, not in the underlying graph structure. In contrast, the entropy of the labeled and generated by, e.g., the preferential attachment model is $\Theta(n \log n)$ [17].

Clearly, given its simplicity, this model should be regarded as a stepping stone toward a better understanding of more advanced models of this type. The extensions are typically defined by a fixed-probability mix of the full duplication model and other rules, such as no-duplication or uniform attachment. We shall deal with such models in a forthcoming paper.

Graph compression has enjoyed a surge in popularity in recent years, as the recent survey [3] shows. However, rigorous information-theoretic results are still lacking, with a few notable exceptions. The rigorous information-theoretic analysis of graph compression (particularly in the unlabeled case) was initiated by Choi and Szpankowski [5], who analyzed structural compression of Erdős-Rényi graphs (see also [1]). The authors of [5] presented a compression algorithm that provably achieves asymptotically the first two terms of the structural entropy. In Łuczak et al. [17] the authors precisely analyzed the labeled and structural entropies and gave asymptotically optimal compression algorithms for preferential attachment graphs. There has been recent work on universal compression schemes, including

in a distributed scenario, by Delgosha and Anantharam [8, 9]. Additionally, several works deal with compression of trees [11, 12, 18, 26].

The full duplication model was almost exclusively analyzed in the context of the typical properties such as degree distribution [6]. It was shown that the average degree depends strongly on the initial conditions [16]. It was also proved that the asymptotic degree distribution fails to converge, yet it exhibits power-law behavior with exponent dependent on the lowest nonzero degree in the initial graph [21]. Other parameters studied in the context of duplication models are the number of small cliques [13] or degree-degree correlations [4]. To the best of our knowledge the entropy and compression of duplication models were not discussed previously in any available literature.

The rest of the paper is organized as follows: In Sect. 2 we define the full duplication model and present its basic properties. In Sect. 3 we establish main results concerning the entropy of the unlabeled and labeled graphs with Sect. 4 being devoted to the construction of algorithms that achieve these bounds within a constant additive term.

## 2 Full Duplication Model

In this section we define the full duplication model and present some of its properties.

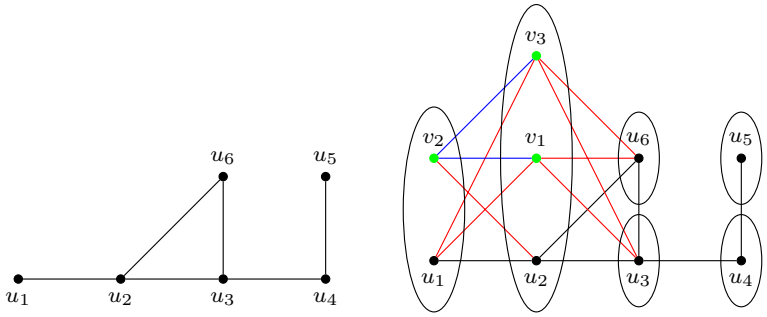
### 2.1 Definitions

The full duplication model is defined as follows: let us denote by $G_0$ a given graph on $n_0$ vertices for some fixed constant $n_0$. Then, for any $1 \leq i \leq n$ we obtain $G_i$ from $G_{i-1}$ by choosing one of the vertices of $G_{i-1}$ (denoted by $v$) uniformly at random, attaching to the graph a new vertex $v_i$ and adding edges between $v_i$ and all vertices adjacent to $v$. Note that $v$ and $v_i$ are not connected – although if one wants to achieve higher clustering, the results in this paper can be straightforwardly applied to the model in which we add not only edges between $v_i$ and the neighbors of $v$, but also between $v_i$ and $v$. Observe that $G_n$ has $n + n_0$ vertices. Also, properties of $G_n$ heavily depend on $G_0$ and its structure, which we assume to be fixed.

Throughout this paper, we will refer to the vertices of the starting graph $G_0$ as $\{u_1, \ldots, u_{n_0}\}$ and to all other vertices from $G_n$ as $\{v_1, \ldots, v_n\}$. We denote by $V(G)$ and $E(G)$ the set of vertices and the set of edges of a graph $G$, respectively. Moreover, we denote by $N_n(v)$ the neighborhood of the vertex $v$, that is, all vertices that are adjacent to $v$ in $G_n$. Sometimes we drop the subscript, if the size of the graph is clear from the context.

An example of the duplication process is presented in Fig. 1. On the top, we show the original $G_0$ on 6 vertices, and on the bottom we plot $G_3$ with new vertices such that $v_1$ is a copy of $u_2$, $v_2$ is a copy of $u_1$, and $v_3$ is a copy of $v_1$.

Here, due to the limited space, we restrict our analysis to asymmetric $G_0$ (i.e., the underlying automorphism group is of size 1); however, extensions to general




**(a)** Initial $G_0$ **(b)** $G_3$

**Fig. 1** Example graph growth in the full duplication model

$G_0$ are rather straightforward. We observe that typically even moderate-sized graphs are likely to be asymmetric.

## 2.2 Basic Properties

Let us introduce the concept of a parent and an ancestor of a vertex. We say that $w$ is the *parent* of $v$ (denoted by $w = P(v)$), when $v$ was copied from $w$ at some time $1 \le i \le n$. We say that $w \in U$ is the *ancestor* of $v$ (denoted by $w = A(v)$), when there exist vertices $v_{i_1}, \ldots, v_{i_k}$ such that $w = P(v_{i_1})$, $v_{i_j} = P(v_{i_{j+1}})$ for $1 \le j \le k-1$, and $v_{i_k} = v$. For convenience we write that if $u \in U$, then $P(u) = u$ and $A(u) = u$. Note that the ancestor of any given vertex is unique. In our example from Fig. 1 $u_2$ is the ancestor of both $v_1$ and $v_3$, but only a parent of $v_1$ and not $v_3$.

Let now define the set of descendants of $u_i \in U$: $\mathcal{C}_{i,n} := \{w \in G_n : A(w) = u_i\}$ for $1 \le i \le n_0$. The neighborhood of a vertex is closely tied to its ancestor, as the following lemma shows:

**Lemma 1** *Let us fix any $1 \le i \le n_0$. For all $n \ge 0$ and any $v \in \mathcal{C}_{i,n}$ we have*

$$N_n(v) = \bigcup_{u_i u_j \in E(G_0)} \mathcal{C}_{j,n}.$$

***Proof*** We prove this by induction. For $n = 0$ we have $\mathcal{C}_{i,0} = \{u_i\}$ and the claim holds.

Now suppose that the claim holds for some $n \ge 0$ and that $P(v_{n+1}) = w$. If $A(w) = u_k$, then $A(v_{n+1}) = u_k$. Moreover,

$$\mathcal{C}_{k,n+1} = \mathcal{C}_{k,n} \cup \{v_{n+1}\}$$
$$\mathcal{C}_{i,n+1} = \mathcal{C}_{i,n} \quad \text{for } i \ne k.$$

We split the remaining part of the proof into several cases:

Case 1, $i = k$, $v = v_{n+1}$: by induction hypothesis we have

$$N_{n+1}(v_{n+1}) = N_{n+1}(P(v_{n+1})) = \bigcup_{j:u_k u_j \in E(G_0)} \mathcal{C}_{j,n} = \bigcup_{j:u_k u_j \in E(G_0)} \mathcal{C}_{j,n+1}.$$

Case 2, $i = k$, $v \neq v_{n+1}$: similarly,

$$N_{n+1}(v) = N_n(v) = \bigcup_{j:u_k u_j \in E(G_0)} \mathcal{C}_{j,n} = \bigcup_{j:u_k u_j \in E(G_0)} \mathcal{C}_{j,n+1}.$$

Case 3, $i \neq k$, $u_i u_k \in E(G_0)$: for any $v \in \mathcal{C}_{i,n+1} = \mathcal{C}_{i,n}$ we have

$$\begin{aligned} N_{n+1}(v) = N_n(v) \cup \{v_{n+1}\} &= \bigcup_{j:u_i u_j \in E(G_0)} \mathcal{C}_{j,n} \cup \{v_{n+1}\} \\ &= \bigcup_{\substack{j\,:\,u_i u_j \in E(G_0) \\ j \neq k}} \mathcal{C}_{j,n} \cup \mathcal{C}_{k,n} \cup \{v_{n+1}\} \\ &= \bigcup_{\substack{j\,:\,u_i u_j \in E(G_0) \\ j \neq k}} \mathcal{C}_{j,n+1} \cup \mathcal{C}_{k,n+1} = \bigcup_{j:u_i u_j \in E(G_0)} \mathcal{C}_{j,n+1}. \end{aligned}$$

Case 4, $i \neq k$, $u_i u_k \notin E(G_0)$: for any $v \in \mathcal{C}_{i,n+1} = \mathcal{C}_{i,n}$ we have

$$N_{n+1}(v) = N_n(v) = \bigcup_{j:u_i u_j \in E(G_0)} \mathcal{C}_{j,n} = \bigcup_{j:u_i u_j \in E(G_0)} \mathcal{C}_{j,n+1}.$$

Therefore, the proof is completed. □

This means that effectively $G_n$ is composed of clusters such that every vertex of $i$-th cluster is connected to every vertex of $j$-th cluster if and only if $u_i u_j \in E(G_0)$. For example, for a graph in Fig. 1b we may identify (marked with ellipses in the figure) the following classes of vertices with identical neighborhoods: $\mathcal{C}_{1,n} = \{u_1, v_2\}$, $\mathcal{C}_{2,n} = \{u_2, v_1, v_3\}$, $\mathcal{C}_{3,n} = \{u_3\}$, $\mathcal{C}_{4,n} = \{u_4\}$ and $\mathcal{C}_{5,n} = \{u_5\}$.

Let now $C_{i,n} = |\mathcal{C}_{i,n}|$, that is, the number of vertices from $G_n$ that are ultimately copies of $u_i$ (including $u_i$ itself).

It is not hard to see that the sequence of variables $(C_{i,n})_{i=1}^{n_0}$ can be described as a ball and urn model with $n_0$ urns. At time $n = 0$ each urn contains exactly one ball. Each iteration consists of picking an urn at random, proportionally to the number of balls in each bin – that is, with probability $\frac{C_{i,n}}{\sum_{j=1}^{n_0} C_{j,n}}$ – and adding a new ball to this urn. It is known [15] that the joint distribution of $(C_{i,n})_{i=1}^{n_0}$ is directly related to the *Dirichlet multinomial distribution* denoted as $Dir(n, \alpha_1, \ldots, \alpha_K)$, with $K = n_0$ and $\alpha_i = 1$ for $1 \leq i \leq n_0$:

$$\Pr\left((C_{i,n})_{i=1}^{n_0} = (k_i + 1)_{i=1}^{n_0}\right) = \begin{cases} nB(n, n_0) & \text{if } \sum_{i=1}^{n_0} k_i = n, \forall_{1 \leq i \leq n_0} k_i \in \mathbb{N}_+, \\ 0 & \text{otherwise.} \end{cases}$$

where $B(x, y)$ is the Euler beta function.

Each variable $C_{i,n}$ is identically distributed – though not independent, as we know that $\sum_{i=1}^{n_0} C_{i,n} = n$ – so we may analyze the properties of $C_n \sim C_{i,n}$ for every $1 \le i \le n_0$. Actually, $C_n - 1$ has the *beta-binomial distribution* $BBin(n, \alpha, \beta)$ with parameters $\alpha = 1, \beta = n_0 - 1$. That is, for any $k \ge 0$:

$$\Pr(C_n = k+1) = \binom{n}{k} \frac{B(k+1, n+n_0-k-1)}{B(1, n_0-1)} \tag{1}$$

$$= (n_0 - 1)\binom{n}{k} B(k+1, n+n_0-k-1). \tag{2}$$

Chung et al. claimed in [6] that the distribution of $C_n$ can be approximated by a density function $f(x) = \exp\left(-\frac{x}{\mathbb{E}C_n}\right)$. Instead, here we have an exact formula.

Moreover, since $C_n \sim BBin(n, 1, n_0 - 1) + 1$ we know immediately that $\mathbb{E}C_n = \frac{n}{n_0} + 1$. For further results we will also need further properties of the beta binomial distribution (with proofs provided in the appendices).

Note that all the logarithms used in subsequent theorems (unless explicitly noted as ln) have base 2.

**Lemma 2** *If $X \sim BBin(n, \alpha, \beta)$, then it is true that $\mathbb{E}[\log(X+1)] = \log n + (\psi(\alpha) - \psi(\alpha+\beta)) \log e + o(1)$ where $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the Euler digamma function.*

Since for all integers $r, s$ we have $\psi(r) - \psi(s) = H_{r-1} - H_{s-1}$ (where $H_j$ denotes the $j$-th harmonic number), it follows that

**Corollary 1** $\mathbb{E}[\log C_n] = \log n - H_{n_0-1} \log e + o(1)$ *for large $n$.*

Similarly, we may prove that:

**Lemma 3** *If $X \sim BBin(n, \alpha, \beta)$, then*

$$\mathbb{E}[(X+1)\log(X+1)] = n \log n \frac{\alpha}{\alpha+\beta} + n \frac{\alpha(\psi(\alpha+1) - \psi(\alpha+\beta+1)) \log e}{\alpha+\beta} + \log n + \left(\psi(\alpha) - \psi(\alpha+\beta) + 1 + \frac{\beta}{2(\alpha+\beta)}\right) \log e + o(1).$$

From the above lemma it is straightforward that:

**Corollary 2** *Asymptotically*

$$\mathbb{E}[C_n \log C_n] = \frac{1}{n_0} n \log n + n \frac{(1 - H_{n_0}) \log e}{n_0} + \log n + \left( \frac{3}{2} - \frac{1}{2n_0} - H_{n_0 - 1} \right) \log e + o(1).$$

## 3 Main Theoretical Results

As discussed in the introduction, our goal is to present results for the duplication graphs 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. In graph structure compression the goal is to remove label information to produce a compact description of a graph structure.

Formally, the labeled graph compression problem can be phrased as follows: one is given a probability distribution $\mathcal{G}_n$ on graphs on $n$ vertices, and the task is to exhibit a pair of mappings (i.e., a source code) $(E, D)$, where $E$ maps graphs to binary strings satisfying the standard prefix code condition, and $D$ maps binary strings back to graphs, such that, for all graphs $G$, $D(E(G)) = G$, and the *expected code length* $\mathbb{E}[|E(G)|]$, with $G \sim \mathcal{G}_n$, is minimized. The standard source coding theorem tells us that the fundamental limit for this quantity is $H(G)$, the Shannon entropy, defined as:

$$H(G) = -\sum_{G \in \mathcal{G}_n} P(G) \log P(G), \tag{3}$$

where $G$ is a functional of the distribution, not a fixed graph.

The *unlabeled* version of this problem relaxes the invertibility constraint on the encoder and decoder. In particular, we only require $D(E(G)) \cong G$; i.e., the decoder only outputs a graph isomorphic to $G$. Again, the optimization objective is to minimize the expected code length. Thus, in effect, the source code efficiently describes the isomorphism type of its input. Denoting by $S(G)$ the isomorphism type of $G$, the fundamental limit for the expected code length is the *structural entropy* of the model, which is given by $H(S(G))$.

There is a relation between the labeled entropy $H(G)$ and structural entropy $H(S(G))$. To express it succinctly for a broad class of graph models we need the automorphism group[1] $\mathrm{Aut}(G)$, and the set $\Gamma(G)$ of *feasible permutations* of $G$; i.e., the set of permutations of $G$ that yield a graph that has positive probability under the random graph model in question. See [5, 17] for more details.

Now, we are ready to present a relation between $H(G)$ and $H(S(G))$. The following lemma was proved in [17]:

[1] An automorphism of a graph is a permutation that preserves edge relations. In other words, it is a permutation which, when applied to the graph, yields the same graph (note that, in mathematical literature, a graph is by default labeled).

**Lemma 4** *We have, for any graph model $G_n$ in which all positive-probability labeled graphs that are isomorphic have the same probability*,

$$H(G_n) - H(S(G_n)) = \mathbb{E}[\log|\Gamma(G_n)|] - \mathbb{E}[\log|\mathrm{Aut}(G_n)|].$$

Now we prove the following results regarding the expected logarithms of the sizes of the automorphism group and feasible permutation set for samples $G_n$ from the full duplication model.

**Lemma 5** *We have*

$$\mathbb{E}[\log|\mathrm{Aut}(G_n)|] = n\log n - nH_{n_0}\log e + \frac{3n_0}{2}\log n + \left(\frac{n_0}{2} - \frac{1}{2} - \frac{3n_0}{2}H_{n_0-1}\right)\log e + \frac{n_0}{2}\log(2\pi) + o(1)$$

*for large n.*

***Proof*** Under the assumption that $|\mathrm{Aut}(G_0)| = 1$ we have $\mathbb{E}[\log|\mathrm{Aut}(G_n)|] = \mathbb{E}\left[\log\prod_{i=1}^{n_0} C_{i,n}!\right]$. To prove it, it is sufficient to notice that all vertices $v$, $w$ such that $A(v) = A(w)$ can be mapped on one another arbitrarily (since by Lemma 1 they have equal neighborhoods)—but if $A(v) \neq A(w)$, there does not exist any automorphism $\sigma$ for which $v$ and $w$ are in the same orbit. Precisely, this is because, if such a $\sigma$ did exist, then one may show that it induces an automorphism of $G_0$.

Thus,

$$\mathbb{E}[\log|\mathrm{Aut}(G_n)|] = \mathbb{E}\left[\log\prod_{i=1}^{n_0} C_{i,n}!\right] = \sum_{i=1}^{n_0}\mathbb{E}[\log C_{i,n}!] = n_0\mathbb{E}[\log C_n!].$$

We use Stirling's approximation together with Corollarys 1 and 2 to obtain

$$\begin{aligned}\mathbb{E}[\log C_n!] &= \mathbb{E}[C_n\log C_n] - \mathbb{E}C_n\log e + \frac{1}{2}\mathbb{E}[\log C_n] + \frac{1}{2}\log(2\pi) + o(1)\\ &= \mathbb{E}[C_n\log C_n] - n\frac{\log e}{n_0} - \log e + \frac{1}{2}\mathbb{E}[\log C_n] + \frac{1}{2}\log(2\pi) + o(1)\\ &= n\log n\frac{1}{n_0} - n\frac{H_{n_0}\log e}{n_0} + \frac{3}{2}\log n\\ &\quad + \left(\frac{1}{2} - \frac{1}{2n_0} - \frac{3}{2}H_{n_0-1}\right)\log e + \frac{1}{2}\log(2\pi) + o(1).\end{aligned}$$

Finally,

$$\mathbb{E}[\log|\mathrm{Aut}(G_n)|] = n\log n - nH_{n_0}\log e + \frac{3n_0}{2}\log n + \left(\frac{n_0}{2} - \frac{1}{2} - \frac{3n_0}{2}H_{n_0-1}\right)\log e + \frac{n_0}{2}\log(2\pi) + o(1).$$

The proof is completed. □

Observe that $G_n$ has $n + n_0$ vertices; therefore, the trivial upper bound on $|\Gamma(G_n)|$ is $(n + n_0)!$. We can do the exact computation of $\Gamma(G_n)$ using the following lemma:

**Lemma 6** *For a permutation $\pi$ of all vertices in $G_n$, the following two claims are equivalent*:

1. *$\pi$ is a relabeling of $G_n$ which produces a positive-probability graph under the full duplication model*,
2. *$\pi$ is a permutation such that for every $1 \le i \le n_0$ there exists $v \in \mathcal{C}_{i,n}$ such that $\pi(v) = u_i$.*

***Proof*** In the whole proof we denote by $u'_1, \ldots, u'_{n_0}$ the vertices that are mapped by $\pi$ to the starting graph vertices $u_1, \ldots, u_{n_0}$. That is, $u'_i = \pi^{-1}(u_i)$ for each $i \in \{1, 2, \ldots, n_0\}$.

($\Rightarrow$) Let $\pi$ produce a graph under the considered model with positive probability.

Suppose now that there exists $1 \le k \le n_0$ such that $u'_k \notin \mathcal{C}_{k,n}$, but $u'_k \in \mathcal{C}_{l,n}$ for some $l \ne k$. Then, by Lemma 1 we know that $N_n(u_k) = \bigcup_{u_k u_j \in E(G_0)} \mathcal{C}_{j,n}$ and $N_n(u'_k) = N_n(u_l) = \bigcup_{u_l u_j \in E(G_0)} \mathcal{C}_{j,n}$.

Since $|\mathrm{Aut}(G_0)| = 1$ by assumption, $N_0(u_k) \ne N_0(u_l)$ and therefore

$$\begin{aligned} N_n(u'_k) \setminus N_n(u_k) &= \bigcup_{u_l u_j \in E(G_0)} \mathcal{C}_{j,n} \setminus \bigcup_{u_k u_j \in E(G_0)} \mathcal{C}_{j,n} \\ &\supseteq \bigcup_{u_l u_j \in E(G_0)} \mathcal{C}_{j,0} \setminus \bigcup_{u_k u_j \in E(G_0)} \mathcal{C}_{j,0} \\ &= N_0(u_l) \setminus N_0(u_k) \ne \emptyset \end{aligned}$$

which proves that $N_n(u'_k) \ne N_n(u_k)$ and therefore $G'_0$ cannot be identical to $G_0$.

($\Leftarrow$) Denote by $v'_1$, ..., $v'_n$ the vertices $\pi^{-1}(v_1)$, ..., $\pi^{-1}(v_n)$; i.e., these vertices are mapped by $\pi$ to vertices outside the seed graph.

By assumption, for every $v'_i$, $1 \le i \le n$, there exists some $u'_j = \pi^{-1}(u_j)$, $1 \le j \le n_0$, such that $v'_i, u'_j \in \mathcal{C}_{j,n}$. Now, in $i$-th step we may just copy $v'_i$ from its respective $u'_j$. It is easy to check that for the neighborhoods $N'(v'_i)$ in the graph created in this way for every $1 \le k \le n_0$ and every $v'_i \in \mathcal{C}_{k,n}$ we have

$$N'_n(v'_i) = N'_n(u'_k) = \bigcup_{j : u_k u_j \in E(G_0)} \mathcal{C}_{j,n} = N_n(u_k) = N_n(v'_i),$$

which concludes the proof. □

**Lemma 7** *Asymptotically*

$$\mathbb{E}[\log|\Gamma(G_n)|] = n\log n - n\log e + \left(n_0 + \frac{1}{2}\right)\log n - H_{n_0-1}\log e + \frac{1}{2}\log(2\pi) + o(1).$$

***Proof*** From Lemma 6, we may construct all admissible permutations by choosing for each $\mathcal{C}_{i,n}$ exactly one vertex which would be mapped to $u_i$ and then arranging remaining $n$ vertices in any order. Therefore:

$$|\Gamma(G_n)| = n!\prod_{i=1}^{n_0}\binom{C_{i,n}}{1} = n!\prod_{i=1}^{n_0} C_{i,n}.$$

Then

$$\begin{aligned}\mathbb{E}[\log|\Gamma(G_n)|] &= \log n! + \sum_{i=1}^{n_0}\mathbb{E}[\log C_{i,n}] = \log n! + n_0\mathbb{E}[\log C_n]\\ &= \log n! + n_0\log n - H_{n_0-1}\log e + o(1),\end{aligned}$$

and the final result follows from the Stirling approximation. □

We now proceed to estimate the structural entropy.

**Theorem 1** *For large $n$ we have*

$$H(S(G_n)\mid G_0) = (n_0-1)\log n - \log(n_0-1)! + o(1).$$

***Proof*** Recalling that we assume throughout that the initial graph $G_0$ is asymmetric, it may be seen that the isomorphism type of $G_n$ is entirely specified by the vector $(C_{i,n})_{i=1}^{n_0}$. We know that $(C_{i,n})_{i=1}^{n_0}$ has the Dirichlet multinomial distribution with $\alpha_i = 1$ for $1 \le i \le n_0$.

Therefore

$$\begin{aligned}H(S(G_n)\mid G_0) &= H\big((C_{i,n})_{i=1}^{n_0}\big)\\ &= -\sum_{(k_i)}\Pr((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0})\log\Pr((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0})\\ &= -\log(nB(n,n_0))\sum_{(k_i)}\Pr((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0})\\ &= -\log n - \log B(n,n_0) = (n_0-1)\log n - \log(n_0-1)! + o(1).\end{aligned}$$

The last two lines follow respectively from the Stirling approximation and the Taylor expansion of $\log B(n,n_0)$, which completes the proof. □

To compute the graph entropy $H(G)$ we can use Lemmas 4, 5 and 7 together with Theorem 1, therefore obtaining the following result.

**Theorem 2** *For large n*

$$H(G_n \mid G_0) = n(H_{n_0} - 1)\log e + \log n \frac{n_0 - 1}{2} - \log(n_0 - 1)! + \left(\frac{1 - n_0}{2} + \frac{3n_0 - 2}{2} H_{n_0 - 1}\right)\log e + \frac{n_0}{2}\log(2\pi) + o(1).$$

Clearly, to compress the whole $G_n$ we would have to encode $G_0$ as well, but since $n_0$ is fixed, this does only affect the constant term. Moreover, by the conditional entropy property, any optimal $G_0$ compression algorithm yields an asymptotically optimal compression for $G_n$.

## 4 Algorithmic Results

In this section we present asymptotically optimal algorithms for compression of labeled and generated according to the full duplication model.

### 4.1 Retrieval of Parameters from $G_n$

In order to present efficient compression algorithms for the duplication model, we must first reconstruct $G_0$ from $G_n$ and find values of $n_0$ and $n$. This is relatively easy to accomplish, as the proof of the next theorem shows.

**Theorem 3** *For a given labeled $G_n$ or its unlabeled version $S(G_n)$, we can retrieve its $n$, $n_0$ and $G_0$ (in the case of structure up to isomorphisms of $G_0$) in polynomial times in terms of $n$.*

***Proof*** For a labeled $G_n$ let $(w_1, w_2, \ldots, w_{n+n_0})$ be its vertices in the order of appearance. Since $(w_1, \ldots, w_{n_0}) = (u_1, \ldots, u_{n_0})$ and $(w_{n_0+1}, \ldots, w_{n_0+n}) = (v_1, \ldots, v_n)$, it is sufficient to find the smallest $k$ such that $N_n(w_k) = N_n(w_i)$ for some $1 \le i < k$. Then $n_0 = k - 1$ and $G_0$ is induced by the sequence $(w_1, \ldots, w_{k-1})$.

The case for is similar: we know (for details see Lemma 6) that the sequence of the first $n_0$ vertices of the graph (that is, $G_0$) contains exactly one vertex from each set $\mathcal{C}_{i,n}$.

From Lemma 1 it follows that $A(v) = A(w)$ iff $N_n(v) = N_n(w)$ for every $v, w \in V(G_n)$, so it is sufficient to scan all vertices of $G_n$ and split them into sets such that $v$ and $w$ belongs to the same set iff $N_n(v) = N_n(w)$. Then, we pick one vertex from each set to from $G_0$. Obviously, $n_0$ and $n$ may be extracted from the sizes of $G_0$ and $G_n$. □

Recall for example that in Fig. 1b we identified the clusters $\{u_1, v_2\}$, $\{u_2, v_1, v_3\}$, $\{u_3\}$, $\{u_4\}$ and $\{u_5\}$. Therefore, we know that $n_0 = 6$, $n = 3$ and the $G_0$ is isomorphic to a graph induced, for example, by the set $\{v_2, v_3, u_3, u_4, u_5\}$.

### 4.2 Unlabeled Graphs

A trivial algorithm CompressUnlabeledSimple for unlabeled compression writes down a sequence $(C_{i,n})_{i=1}^{n_0}$ associated with our $G_n$ as $\log n$-bit numbers. This always requires $n_0 \log n$ bits, so $\mathbb{E}L_{SU}(n) = n_0 \log n$, where $L_{SU}$ denotes the code length of our proposed scheme. By Theorem 1 this achieves the fundamental limit to within a multiplicative factor of $1 + \frac{1}{n_0 - 1}$.

However, it is easy to design an optimal algorithm up to a constant additive error, provided we have already compressed $G_0$ or $S(G_0)$ (anyway, a graph of fixed size). The pseudocode of an optimal algorithm, called CompressUnlabeledOpt, based on arithmetic coding, is as follows:

**function** CompressUnlabeledOpt($S(G_n)$, $G_0$)
  Fix any ordering $(v_1, \ldots, v_n)$ of the vertices of $V(G_n) \setminus V(G_0)$
  $a \leftarrow 0$, $b \leftarrow 1$
  **for** $i = 1, 2, \ldots, n_0$ **do**
    $C[i] \leftarrow 0$
    **for** $j = 1, 2, \ldots, n$ **do**
      **if** $N(v_j) = N(u_i)$ **then**
        $C[i] \leftarrow C[i] + 1$
  **for** $i = n_0, n_0 - 1, \ldots, 2$ **do**
    $start \leftarrow (i-1) \sum_{j=0}^{C[i]-1} \binom{n}{j} B(j+1, n+i-j-1)$
    $end \leftarrow start + (i-1)\binom{n}{C[i]} B(C[i]+1, n+i-C[i]-1)$
    $interval \leftarrow b - a$
    $b \leftarrow a + interval \cdot end$, $a \leftarrow a + interval \cdot start$
    $n \leftarrow n - C[i]$
  $p \leftarrow b - a$, $x \leftarrow \dfrac{a+b}{2}$
  **return** first $\lceil -\log p \rceil + 1$ bits of $x$

The next finding proves that CompressUnlabeledOpt is nearly optimal.

**Theorem 4** *Algorithm CompressUnlabeledOpt is optimal up to a two bits for unlabeled graphs compression, when the graph is generated by the full duplication model.*

***Proof*** It is sufficient to observe that

$$
\begin{aligned}
&\Pr((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0}) \\
&= \Pr((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0} \mid C_{n_0,n} = k_{n_0}+1)\Pr(C_{n_0,n} = k_{n_0}+1) \\
&= \Pr((C_{i,n})_{i=1}^{n_0-1} = (k_i+1)_{i=1}^{n_0-1} \mid C_{n_0,n} = k_{n_0}+1)\Pr(C_{n_0,n} = k_{n_0}+1) \\
&= \Pr\left((C_{i,n})_{i=1}^{n_0-1} = (k_i+1)_{i=1}^{n_0-1} \Big| \sum_{i=0}^{n_0-1} C_{i,n} = n+n_0-k_{n_0}-1\right) \\
&\quad \Pr(C_{n_0,n} = k_{n_0}+1) \\
&= \Pr((C_{i,n-k_{n_0}})_{i=1}^{n_0-1} = (k_i+1)_{i=1}^{n_0-1}) \\
&\quad \binom{n}{k_{n_0}}(n_0-1)B(k_{n_0}+1, n+n_0-k_{n_0}-1),
\end{aligned}
$$

The last equality follows from the fact that the marginal distribution of the Dirichlet multinomial distribution is the beta-binomial distribution, given by Eq. 1. Moreover, if we fix value of the last coordinate of $(C_{i,n})_{i=1}^{n_0}$ to $k+1$, then the resulting distribution is also (shifted) Dirichlet multinomial, but with $n_0-1$ coordinates and all values summing up to $n+n_0-k-1$.

We repeat this process until we have 2-dimensional distribution:

$$
\begin{aligned}
&\Pr((C_{i,n})_{i=1}^{2} = (k_i+1)_{i=1}^{2}) \\
&\quad = \Pr(C_{1,n} = k_1+1 \mid C_{1,n} = k_1+1)\Pr(C_{2,n} = k_1+1) \\
&\quad = \binom{k_1+k_2}{k_2} B(k_2+1, k_1+1).
\end{aligned}
$$

By the properties of arithmetic coding (see e.g. [7]), $\mathbb{E}L_O(S(G_n) \mid G_0) \le H((C_{i,n})_{i=0}^{n_0}) + 2 = H(S(G_n) \mid G_0) + 2$, where $L_O$ denotes the code length. This completes the proof. □

### 4.3 Labeled Graphs

We note that the labeled graph $G_n$ is equivalent to a sequence $(A(v_i))_{i=1}^{n}$ for a given (labeled) $G_0$, which obviously can be encoded separately using a constant number of bits.

A trivial algorithm CompressLabeledSimple just writes all $A(v_i)$ as $\log n_0$-bit numbers. Clearly, this always gives us a codeword with length exactly $\mathbb{E}L_{SL}(n) = n \log n_0$. From Theorem 2 it is known that this algorithm is asymptotically $(1 + \frac{1-\gamma}{\log n_0})$-approximately optimal, where $\gamma$ is Euler-Mascheroni constant.

It is easy to design an asymptotically optimal algorithm up to a constant error. Indeed, the sequence of $A(v_i)$ is random with $\Pr(A(v_i) = u_j) = \frac{C_{j,i-1}}{n_0+i-1}$ for $1 \le i \le n$, $1 \le j \le n_0$. Therefore, given $G_{i-1}$ we know the conditional probabilities of $G_i$ and we may construct another algorithm based on arithmetic coding.

The pseudocode of the optimal algorithm is as follows:

**function** CompressLabeledOpt($G_n$, $G_0$)
  $a \leftarrow 0$, $b \leftarrow 1$
  **for** $i = 1, 2, \ldots, n_0$ **do**
    $C[i] \leftarrow 1$
  **for** $i = 1, 2, \ldots, n$ **do**
    **for** $j = 1, 2, \ldots, n_0$ **do**
      **if** $N(v_i) = N(u_j)$ **then**
        $start \leftarrow \sum_{k=1}^{j-1} \frac{C[k]}{n_0+i-1}$
        $end \leftarrow start + \frac{C[j]}{n_0+i-1}$
        $interval \leftarrow b - a$
        $b \leftarrow a + interval \cdot end$, $a \leftarrow a + interval \cdot start$
        $C[j] \leftarrow C[j] + 1$
  $p \leftarrow b - a$, $x \leftarrow \dfrac{a+b}{2}$
  **return** first $\lceil -\log p \rceil + 1$ bits of $x$

The next theorem proves that CompressLabeledOpt is almost optimal up to a known additive constant.

**Theorem 5** *Algorithm CompressLabeledOpt is optimal up to a two bits for labeled graph compression, when the graph is generated by the full duplication model.*

***Proof*** By the well-known properties of arithmetic encoding (see [7]), we know that $\mathbb{E}L_O(G_n \mid G_0) \leq H(G_n \mid G_0) + 2$, where $L_O$ denotes the code length. □

Note that these two algorithms for the labeled graphs differ only in that the optimal one updates the probabilities at each step and the second fixes them to a constant value of $\frac{1}{n_0}$.



## Appendix

## Proof of Lemma 2

We can write $\mathbb{E}[\ln(X+1)]$ as follows:

$$\mathbb{E}[\ln(X+1)] = \int_0^1 \pi(p,\alpha,\beta)\mathbb{E}[\ln(X+1) \mid p]\mathrm{d}p \tag{4}$$

as $X \sim BBin(n,\alpha,\beta)$ can be defined as a compound distribution $X \sim Bin(n,p)$ for $p \sim Beta(n,\alpha,\beta)$. Here $\pi(p,\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$ is the beta probability distribution function.

We proceed by defining an event $A = [|X - np| \le \epsilon np]$ for some fixed $\epsilon > 0$ and then splitting the remaining part into two regions: $M_1 = [0, n^{-2/3}]$ and $M_2 = [n^{-2/3}, 1]$.

First, we use Taylor expansion around $\mathbb{E}[X \mid p] = np$ and get:

$$\begin{aligned}\mathbb{E}[\ln(X+1) \mid p] &= \ln(np+1) - \mathbb{E}\left[\frac{(X-np)^2}{2(c+1)^2} \;\middle|\; p\right] \\ &= \ln n + \ln p + \ln\left(1+\frac{1}{np}\right) - \mathbb{E}\left[\frac{(X-np)^2}{2(c+1)^2} \;\middle|\; p\right],\end{aligned} \tag{5}$$

where $c$ is a random variable with values within the range of $X$.

We know that

$$\begin{aligned}&\int_0^1 \pi(p,\alpha,\beta)\ln n\mathrm{d}p = \ln n \\ &\int_0^1 \pi(p,\alpha,\beta)\ln p\mathrm{d}p = \psi(\alpha) - \psi(\alpha+\beta) \\ &\int_0^1 \pi(p,\alpha,\beta)\ln\left(1+\frac{1}{np}\right)\mathrm{d}p \le \frac{1}{B(\alpha,\beta)}\int_0^1 \ln\left(1+\frac{1}{np}\right)\mathrm{d}p \\ &\qquad = \frac{1}{B(\alpha,\beta)}\left[p\ln\left(1+\frac{1}{np}\right) + \frac{1}{n}\ln(np+1)\right]_0^1 = O\Big(\frac{\ln n}{n}\Big).\end{aligned}$$

For $M_1$ it holds that

$$\begin{aligned}\int_{M_1} \pi(p,\alpha,\beta)\mathbb{E}\left[\frac{(X-np)^2}{2(c+1)^2} \;\middle|\; p\right]\mathrm{d}p &\le \int_{M_1} \pi(p,\alpha,\beta)\frac{np(1-p)}{2}\mathrm{d}p \\ &\le \int_{M_1} \pi(p,\alpha,\beta)\frac{np}{2}\mathrm{d}p \le \frac{n}{2B(\alpha,\beta)}\int_{M_1} p\mathrm{d}p = O\big(n^{-1/3}\big).\end{aligned}$$

Conditioned on $A$, it is true that $np(1-\epsilon) \le c \le np(1+\epsilon)$. Moreover, $\Pr(A \mid p)\mathbb{E}[(X-np)^2 \mid p, A] \le \mathbb{E}[(X-np)^2 \mid p] = np(1-p)$, therefore:

$$\begin{aligned}
&\int_{M_2} \pi(p,\alpha,\beta)\Pr(A \mid p)\mathbb{E}\left[\frac{(X-np)^2}{2(c+1)^2} \;\middle|\; p, A\right]\mathrm{d}p \\
&\quad\le \int_{M_2} \pi(p,\alpha,\beta)\frac{np(1-p)}{2(np(1-\epsilon)+1)^2}\mathrm{d}p \\
&\quad\le \int_{M_2} \pi(p,\alpha,\beta)\frac{np}{2n^2p^2(1-\epsilon+\frac{1}{np})^2}\mathrm{d}p \\
&\quad\le n^{-1}\int_{M_2} \pi(p,\alpha,\beta)\frac{1}{2p(1-\epsilon+\frac{1}{np})^2}\mathrm{d}p \\
&\quad\le n^{-1}\frac{1}{2B(\alpha,\beta)(1-\epsilon)^2}\int_{M_2}\frac{1}{p}\mathrm{d}p = O\Big(\frac{\ln n}{n}\Big).
\end{aligned}$$

Furthermore, for $M_2$ conditioned on $\neg A$, we use the Chernoff bound:

$$\Pr(\neg A \mid p) = \Pr(|X-np| > \epsilon np \mid p) \le 2\exp\left(-\frac{\epsilon^2 np}{3}\right)$$

for a fixed constant $\epsilon > 0$ together with the obvious fact that $(X-np)^2 \le n^2$ to bound the remaining error

$$\begin{aligned}
&\int_{M_2} \pi(p,\alpha,\beta)\Pr(\neg A \mid p)\mathbb{E}\left[\frac{(X-np)^2}{2(c+1)^2} \;\middle|\; p, \neg A\right]\mathrm{d}p \\
&\quad\le \int_{M_2} \pi(p,\alpha,\beta)\exp\left(-\frac{\epsilon^2 np}{3}\right)\frac{n^2}{2}\mathrm{d}p \\
&\quad\le \frac{n^2}{2B(\alpha,\beta)}\int_{M_2}\exp\left(-\frac{\epsilon^2 np}{3}\right)\mathrm{d}p \\
&\quad\le \frac{3n}{2B(\alpha,\beta)\epsilon^2}\exp\left(-\frac{\epsilon^2 n^{-1/3}}{3}\right) = o(1).
\end{aligned}$$

The proof follows from using all the bounds presented above and combining them with Eqs. 4 and 5. □

## Proof of Lemma 3

We proceed as before by writing $\mathbb{E}[(X+1)\ln(X+1)]$ as follows:

$$\mathbb{E}[(X+1)\ln(X+1)] = \int_0^1 \pi(p,\alpha,\beta)\mathbb{E}[(X+1)\ln(X+1) \mid p]\mathrm{d}p. \tag{6}$$

Once again we define an event $A = [|X-np| \le \epsilon np]$ for some fixed $\epsilon > 0$ and using Taylor expansion around $\mathbb{E}[X \mid p] = np$:

$$
\begin{aligned}
&\mathbb{E}[(X+1)\ln(X+1) \mid p] \\
&= (np+1)\ln(np+1) + \frac{np(1-p)}{2(np+1)} - \mathbb{E}\left[\frac{(X-np)^3}{6(c+1)^2} \;\middle|\; p\right] \\
&= np\ln n + np\ln p + np\ln\left(1+\frac{1}{np}\right) + \ln n + \ln p + \ln\left(1+\frac{1}{np}\right) \\
&+ \frac{np(1-p)}{2(np+1)} - \mathbb{E}\left[\frac{(X-np)^3}{2(c+1)^2} \;\middle|\; p\right],
\end{aligned} \tag{7}
$$

where $c$ is a random variable with values within the range of $X$.

Moreover,

$$
\begin{aligned}
\int_0^1 \pi(p,\alpha,\beta) np\ln n \mathrm{d}p &= \frac{\alpha}{\alpha+\beta} n\ln n \\
\int_0^1 \pi(p,\alpha,\beta) np\ln p \mathrm{d}p &= \frac{\alpha(\psi(\alpha+1)-\psi(\alpha+\beta+1))}{\alpha+\beta} n \\
\int_0^1 \pi(p,\alpha,\beta)\frac{1}{np+1}\mathrm{d}p &\le \frac{1}{B(\alpha,\beta)}\int_0^1 \frac{1}{np+1}\mathrm{d}p = o(1) \\
\int_0^1 \pi(p,\alpha,\beta)\frac{p}{np+1}\mathrm{d}p &= \frac{1}{n}\int_0^1 \pi(p,\alpha,\beta)\left(1-\frac{1}{np+1}\right)\mathrm{d}p = o(1) \\
\int_0^1 \pi(p,\alpha,\beta)\frac{np(1-p)}{2(np+1)}\mathrm{d}p &= \int_0^1 \pi(p,\alpha,\beta)\left(\frac{1-p}{2}+\frac{1-p}{2(np+1)}\right)\mathrm{d}p \\
&= \frac{\beta}{2(\alpha+\beta)} + o(1).
\end{aligned}
$$

The term $np\ln\left(1+\frac{1}{np}\right)$ can be computed as following:

$$
\begin{aligned}
&\int_0^1 \pi(p,\alpha,\beta) np\ln\left(1+\frac{1}{np}\right)\mathrm{d}p \\
&\quad = 1 + \int_0^{2/n} \pi(p,\alpha,\beta)\left(np\ln\left(1+\frac{1}{np}\right)-1\right)\mathrm{d}p \\
&\qquad + \int_{2/n}^1 \pi(p,\alpha,\beta)\left(np\ln\left(1+\frac{1}{np}\right)-1\right)\mathrm{d}p
\end{aligned}
$$

with

$$
\begin{aligned}
\int_{2/n}^1 \pi(p,\alpha,\beta) np\left(\ln\left(1+\frac{1}{np}\right)-1\right)\mathrm{d}p &\le \int_{2/n}^1 \pi(p,\alpha,\beta)\frac{-1}{np}\mathrm{d}p \\
&\le \frac{1}{nB(\alpha,\beta)}\int_{2/n}^1 -\frac{1}{p}\mathrm{d}p = o(1)
\end{aligned}
$$

and

$$\begin{aligned}
&\int_0^{2/n} \pi(p,\alpha,\beta) np\left(\ln\left(1+\frac{1}{np}\right)-1\right)\mathrm{d}p \\
&\quad\le \frac{1}{B(\alpha,\beta)}\int_0^{2/n} np\left(\ln\left(1+\frac{1}{np}\right)-1\right)\mathrm{d}p \\
&\quad\le \frac{1}{nB(\alpha,\beta)}\int_0^{2} x\left(\ln\left(1+\frac{1}{x}\right)-1\right)\mathrm{d}x \\
&\quad\le \frac{1}{2nB(\alpha,\beta)}\left[x^2\ln\left(1+\frac{1}{x}\right)+x-\ln(x+1)-x^2\right]_0^2 = o(1).
\end{aligned}$$

Finally, we estimate the remainder term for two regions: $M_1 = [0, n^{-2/3}]$ and $M_2 = [n^{-2/3}, 1]$.

For $M_1$ it is true that

$$\begin{aligned}
&\int_{M_1} \pi(p,\alpha,\beta)\mathbb{E}\left[\frac{(X-np)^3}{6(c+1)^2}\;\middle|\; p\right]\mathrm{d}p \\
&\quad\le \int_{M_1} \pi(p,\alpha,\beta)\frac{np(1-p)(1-2p)}{6}\mathrm{d}p \\
&\quad\le \int_{M_1} \pi(p,\alpha,\beta)\frac{np}{6}\mathrm{d}p \\
&\quad\le \frac{n}{6B(\alpha,\beta)}\int_{M_1} p\mathrm{d}p = O\left(n^{-1/3}\right).
\end{aligned}$$

Furthermore, for $A$ defined as above we have

$$\begin{aligned}
&\Pr(A \mid p)\mathbb{E}[(X-np)^3 \mid p, A] + \Pr(\neg A \mid p)\mathbb{E}[(X-np)^3 \mid p, \neg A] \\
&\quad = \mathbb{E}[(X-np)^3 \mid p] \\
&\mathbb{E}[(X-np)^3 \mid p, \neg A] \ge -n^3p^3\Pr(\neg A \mid p) = \Pr(|X-np| \ge \epsilon np \mid p) \\
&\quad \le 2\exp\left(-\frac{\epsilon^2 np}{3}\right).
\end{aligned}$$

and therefore

$$\begin{aligned}
&\Pr(A \mid p)\mathbb{E}[(X-np)^3 \mid p, A] \\
&= \mathbb{E}[(X-np)^3 \mid p] - \Pr(\neg A \mid p)\mathbb{E}[(X-np)^3 \mid p, \neg A] \\
&\le \mathbb{E}[(X-np)^3 \mid p] + 2n^3p^3\exp\left(-\frac{\epsilon^2 np}{3}\right) \\
&\le np(1-p)(1-2p) + o(1).
\end{aligned}$$

Now we proceed similarly as in the previous proof, using the fact that conditioning on $A$ guarantees that $np(1-\epsilon) \le c \le np(1+\epsilon)$. As we may safely assume that $n \ge 3$, we need to consider two subregions separately:

$$\begin{aligned}
&\int_{1/n^{2/3}}^{1/2} \pi(p,\alpha,\beta)\Pr(A \mid p)\frac{np(1-p)(1-2p)}{6(c+1)^2}\mathrm{d}p \\
&\quad\le \int_{1/n^{2/3}}^{1/2} \pi(p,\alpha,\beta)\frac{np}{6(np(1-\epsilon)+1)^2}\mathrm{d}p \\
&\quad\le \frac{1}{B(\alpha,\beta)}\int_{1/n^{2/3}}^{1/2} \frac{np}{6n^2p^2\left(1-\epsilon+\frac{1}{np}\right)^2}\mathrm{d}p \\
&\quad\le \frac{1}{B(\alpha,\beta)}\int_{1/n^{2/3}}^{1/2} \frac{1}{6np(1-\epsilon)^2}\mathrm{d}p = o(1)
\end{aligned}$$

and

$$\begin{aligned}
&\int_{1/2}^{1} \pi(p,\alpha,\beta)\Pr(A \mid p)\frac{np(1-p)(2p-1)}{6(c+1)^2}\mathrm{d}p \\
&\quad\le \int_{1/2}^{1} \pi(p,\alpha,\beta)\frac{2np^2}{6(np(1-\epsilon)+1)^2}\mathrm{d}p \\
&\quad\le \frac{1}{B(\alpha,\beta)}\int_{1/2}^{1} \frac{2np^2}{6n^2p^2\left(1-\epsilon+\frac{1}{np}\right)^2}\mathrm{d}p \\
&\quad\le \frac{1}{B(\alpha,\beta)}\int_{1/2}^{1} \frac{1}{3n(1-\epsilon)^2}\mathrm{d}p = o(1).
\end{aligned}$$

Therefore for $M_2$ conditioned on $A$ we have

$$\begin{aligned}
&\int_{M_2} \pi(p,\alpha,\beta)\Pr(A \mid p)\mathbb{E}\left[\frac{(X-np)^3}{6(c+1)^2}\;\middle|\; p, A\right]\mathrm{d}p \\
&\quad\le \int_{1/n^{2/3}}^{1} \pi(p,\alpha,\beta)\frac{np(1-p)(1-2p)+o(1)}{6(c+1)^2}\mathrm{d}p = o(1).
\end{aligned}$$

Finally, for $M_2$ conditioned on $\neg A$ we have

$$\begin{aligned}
&\int_{M_2} \pi(p,\alpha,\beta)\Pr(\neg A \mid p)\mathbb{E}\left[\frac{(X-np)^3}{6(c+1)^2}\;\middle|\; p, \neg A\right]\mathrm{d}p \\
&\quad\le \int_{M_2} \pi(p,\alpha,\beta)\exp\left(-\frac{\epsilon^2 np}{3}\right)\frac{n^3}{6}\mathrm{d}p \\
&\quad\le \frac{n^3}{6B(\alpha,\beta)}\int_{M_2} \exp\left(-\frac{\epsilon^2 np}{3}\right)\mathrm{d}p \\
&\quad\le \frac{n^2}{2B(\alpha,\beta)\epsilon^2}\exp\left(-\frac{\epsilon^2 n^{-1/3}}{3}\right) = o(1).
\end{aligned}$$

To finish the proof it is sufficient to apply all the bounds presented above to the Eqs. 6 and 7. □

## References


1. Abbe, E.: Graph compression: the effect of clusters. In: Proceedings of the Fifty-fourth Annual Allerton Conference (2016)
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys **74**, 47–97 (2002)
3. Besta, M., Hoefler, T.: Survey and taxonomy of lossless graph compression and space-efficient graph representations. Preprint (2018). https://arxiv.org/pdf/1806.01799
4. Boccaletti, S., Hwang, D.U., Latora, V.: Growing hierarchical scale-free networks by means of non-hierarchical processes. Int. J. Bifurc. Chaos **17**, 2447–2452 (2007)
5. Choi, Y., Szpankowski, W.: Compression of graphical structures: fundamental limits, algorithms, and experiments. IEEE Trans. Inf. Theor. **58**(2), 620–638 (2012). https://doi.org/10.1109/TIT.2011.2173710
6. Chung, F., Lu, L., Dewey, T.G., Galas, D.J.: Duplication models for biological networks. J. Comput. Biol. **10**(5), 677–687 (2003). https://doi.org/10.1089/106652703322539024
7. Cover, T., Thomas, J.: Elements of Information Theory, 2nd edn. Wiley, London (2006)
8. Delgosha, P., Anantharam, V.: Distributed compression of graphical data. In: 2018 IEEE International Symposium on Information Theory (ISIT), pp. 2216–2220 (2018)
9. Delgosha, P., Anantharam, V.: Universal lossless compression of graphical data. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp. 1578–1582 (2017)
10. Frieze, A., Karoński, M.: Introduction to Random Graphs. Cambridge University Press, Cambridge (2016)
11. Gołebiewski, Z., Magner, A., Szpankowski, W.: Entropy of some general plane trees. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp. 301–305 (2017). https://doi.org/10.1109/ISIT.2017.8006538
12. Hucke, D., Lohrey, M.: Universal tree source coding using grammar-based compression. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp. 1753–1757 (2017). https://doi.org/10.1109/ISIT.2017.8006830
13. Ispolatov, I., Krapivsky, P., Mazo, I., Yuryev, A.: Cliques and duplication-divergence network growth. New J. Phys. **7**, 145 (2005)
14. Ispolatov, I., Krapivsky, P.L., Yuryev, A.: Duplication-divergence model of protein interaction network. Phys. Rev. E **71**, 061911 (2005). https://doi.org/10.1103/PhysRevE.71.061911
15. Johnson, N., Kemp, A., Kotz, S.: Univariate Discrete Distributions. Wiley, London (2005)
16. Kim, J., Krapivsky, P.L., Kahng, B., Redner, S.: Infinite-order percolation and giant fluctuations in a protein interaction network. Phys. Rev. E **66**, 055101 (2002). https://doi.org/10.1103/PhysRevE.66.055101
17. Łuczak, T., Magner, A., Szpankowski, W.: Asymmetry and structural information in preferential attachment graphs. Random Struct. Algorithms (2019)
18. Magner, A., Turowski, K., Szpankowski, W.: Lossless compression of binary trees with correlated vertex names. Trans. Inf. Theory **64** (2018)
19. Newman, M.: Networks: An Introduction. Oxford University Press, Oxford (2010)
20. Pastor-Satorras, R., Smith, E., Solé, R.: Evolving protein interaction networks through gene duplication. J. Theor. Biol. **222**(2), 199–210 (2003). https://doi.org/10.1016/S0022-5193(03)00028-6
21. Raval, A.: Some asymptotic properties of duplication graphs. Phys. Rev. E **68**, 066119 (2003)
22. Shao, M., Yang, Y., Guan, J., Zhou, S.: Choosing appropriate models for protein-protein interaction networks: a comparison study. Brief. Bioinform. **15**(5), 823–838 (2014). https://doi.org/10.1093/bib/bbt014
23. Solla, P.D.D.: A general theory of bibliometric and other cumulative advantage processes. J. Am. Soc. Inf. Sci. **27**(5), 292–306 (1976). https://doi.org/10.1002/asi.4630270505
24. van der Hofstad, R.: Random Graphs and Complex Networks, vol. 1. Cambridge University Press, Cambridge (2016)

25. Vázquez, A., Flammini, A., Maritan, A., Vespignani, A.: Modeling of protein interaction networks. Complexus **1**(1), 38–44 (2003)
26. Zhang, J., Yang, E.H., Kieffer, J.C.: A universal grammar-based code for lossless compression of binary trees. IEEE Trans. Inf. Theory **60**(3), 1373–1386 (2014). https://doi.org/10.1109/TIT.2013.2295392