Compression of Dynamic Graphs Generated by a Duplication Model

Krzysztof Turowski, Abram Magner, and Wojciech Szpankowski
Center for Science of Information,
Department of Computer Science
Purdue University, West Lafayette, IN, USA
Email: krzysztof.szymon.turowski@gmail.com, abram10@gmail.com, spa@cs.purdue.edu

I. Introduction

Complex systems can often be modeled as a set of dynamic (time-evolving) graphs. In these systems, the spatial structure of the domain, as well as patterns of interactions evolve in time – this evolution critically determines emergent properties, associated function, robustness, and security of the system. Understanding the evolution of such systems, identifying criticalities, and controlling driving processes, pose foundational problems in system design and analysis.

Compression schemes for advanced data structures have become the challenge of today. To meet this challenge, we have recently initiated a systematic study of advanced data structures such as trees [17], [18], unlabeled Erdős-Renyi graphs [3], and preferential attachment graphs [11]. In this paper we continue this quest and consider another dynamic graphs generated by a *vertex-duplication model*.

The full duplication model was almost exclusively analyzed in the context of the typical properties such as degree distribution [4]. It is shown that average degree depends strongly on the initial conditions[10]. It is also proved that the asymptotic degree distribution is nonstationary, yet it exhibits power-law behaviour with exponent dependent on the lowest nonzero degree in the initial graph[15]. Other parameters studied in the context of duplication models are number of small cliques [8] or degree-degree correlation [2].

The rest of the paper is organized as follows: in Section II we define a model together and present its basic properties. In Section III we establish main results of this paper concerning entropy of the unlabelled and labelled graphs generated by a full duplication model. Algorithms which achieve this bounds within a constant term are presented in Section IV with Section V being devoted to the experimental verification of these findings.

II. Full duplication model

A. Definition

Definition of the full duplication model can be states as follows: let us denote by G_0 an original graph on n_0 vertices for some fixed constant n_0. Then, for any $1 \leq i \leq n$ we obtain G_{i-1} from G_i by choosing one of vertices of G_i (denoted by v) uniformly at random, attaching to a 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 neighbours of v, but also between v_i and v.

Every G_n has $n+n_0$ vertices. The distinction between n and n_0, not present in other models, is based on the fact that the properties of the G_n heavily depend on G_0 and its structure. However, since n_0 is fixed, we may treat G_0 also as fixed beforehand.

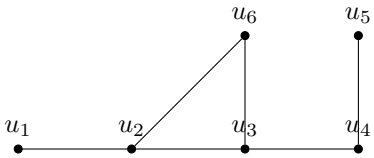
Throughout this paper, we will refer to the vertices of G_0 as $U = \{u_1, \ldots, u_{n_0}\}$ and to all other vertices from G_n as $V = \{v_1, \ldots, v_n\}$. We denote by $N_n(v)$ the neighbourhood of the vertex v, that is, all vertices there are adjacent to v in G_n. Sometimes we drop the subscript, if the size of the graph is clear from the context.

An example of an evolving graph is presented in Figure 1. On the left, we see original G_0 on 6 vertices, on the right there is G_3 with added vertices such that v_1 as a copy of u_2, v_2 – a copy of u_1, and v_3 – a copy of v_1.

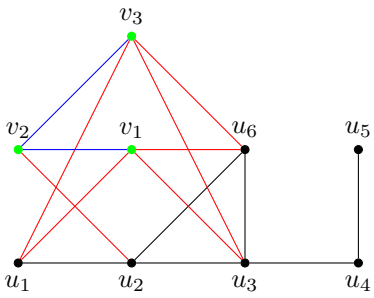
Here, due to the limited space and the brievity of analysis, we restict the analysis to asymmetric G_0, that is, $|\operatorname{Aut}(G_0)| = 1$. This is not much of a concern, since the analysis can be, with certain reservations and complications, extended to arbitrary G_0. Moreover, it is supported by the fact that even medium-sized graphs (up to 100 vertices) are likely to be asymmetric.

B. Basic properties

Let us introduce concepts of a parent and an ancestor of a vertex. We say that w is *a parent* of v (denoted by $w = P(v)$), when v was copied from w at some time $1 \leq i \leq n$. We say that w is *an ancestor* of v (denoted by $w = A(v)$), when $w \in U$ and there exists vertices $v_{i_1}, \ldots, v_{i_k}$ such that $w =$



(a) Initial G_0



(b) G_3

Fig. 1: Example graph growth in pure duplication model

$P(v_{i_1})$, $v_{i_j} = P(v_{i_{j+1}})$ for $1 \leq j \leq k-1$, and $v_{i_k} = v$. For convenience we assume that if $v \in U$, then $P(u) = A(u) = u$. In our example from Figure 1 u_2 is an ancestor of both v_1 and v_3, but only a parent of v_1 and not v_3.

Le us also define the neighbourhood of v $N(v) = \{u\, uv \in G_n\}$. It is closely tied to A as the following lemma shows:

Lemma 1. *If $A(v_i) = A(v_j)$ if and only if $N(v_i) = N(v_j)$.*

Proof. If $u = A(v_i)$, then obviously $N(v_i) = N(u)$ in G_j. All vertices v_k, $k > i$, are either adjacent to both u and v_j or to none of them – therefore $N(v_i) = N(u)$ in any G_n, $n \geq j$.

If G_0 is simple and asymmetric, then $N(u) \neq N(u')$ in G_0 for any u and u', so there exists a vertex u'' which is adjacent to exactly one of u, u'. Clearly, $N(u) - N(u')$ in any G_n also constains u'' so $N(u) \neq N(u')$ in G_n for any N

Since the relations A and N are transitive, we may conclude the proof. □

For $1 \leq i \leq n_0$ $C_{i,n} = |\{w \in G_n \colon A(w) = u_i\}$, that is, a count of vertices from G_n such that they are ultimately copies of u_i (including u_i itself).

The sequence of variables $(C_{i,n})_{i=1}^{n_0}$ can be described as a urn-ball model with n_0 urns. At $n = 0$ 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 $\frac{C_{i,n}}{\sum_{j=1}^{n_0} C_{j,n}}$ – removing this ball and adding two new balls to this urn. Therefore, the joint distribution of $(C_{i,n})_{i=1}^{n_0}$ is the Dirichlet multinomial distribution with $K = n_0$ and $\alpha_i = 1$ for $1 \leq i \leq n_0$:

$$\mathbb{P}\{((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0}\} = \frac{\Gamma(n+1)\Gamma(n_0)}{\Gamma(n+n_0)} \prod_{i=1}^{n_0} \frac{\Gamma(k_i+1)}{\Gamma(k_i+1)} \tag{1}$$

Each variable $C_{i,n}$ is identically distributed – though not independent, as $\sum_{i=1}^{n_0} C_{i,n} = n$ – so we may analyze the properties of $C_n \sim C_{i,n}$ for every $1 \leq i \leq n_0$. It has beta-binomial distribution with parameters $\alpha = 1$, $\beta = n_0 - 1$. That is:

$$\mathbb{P}\{C_n = k+1\} = \binom{n}{k} \frac{B(k+1, n+n_0-k-1)}{B(1, n_0-1)} \tag{2}$$

$$= (n_0-1)\binom{n}{k} B(k+1, n+n_0-k-1) \tag{3}$$

for any $k \geq 0$, where B is a beta function.

Therefore, since $C_n \sim BBin(n, 1, n_0-1)+1$ we know immediately that $\mathbb{E}C_n = \frac{n}{n_0} + 1$. We will also need two other properties, proved in the Appendix A.

Lemma 2. *If $X \sim BBin(n, \alpha, \beta) + 1$, then* $\mathbb{E}[\log X] = \log n + (\psi(\alpha) - \psi(\alpha+\beta))\log e + O(n^{-1})$.

where ψ is a digamma function.

Using the fact that for all integers r, s $\psi(r) - \psi(s) = H_{r-1} - H_{s-1}$ we obtain:

Corollary 1. *For all $n > 0$* $\mathbb{E}[\log C_n] = \log n - H_{n_0-1}\log e + O(n^{-1})$.

Lemma 3. *If $X \sim BBin(n, \alpha, \beta) + 1$, then* $\mathbb{E}[X \log X] = n \log n \frac{\alpha}{\alpha+\beta} + n\frac{\alpha(\psi(\alpha+1)-\psi(\alpha+\beta+1))\log e}{\alpha+\beta} + \log n + \left(\psi(\alpha) - \psi(\alpha+\beta) + \frac{3}{2} - \frac{\alpha}{2(\alpha+\beta)}\right)\log e + O(n^{-1})$.

As previously, in our case we can simplify some terms using the fact that $H_s = H_{s-1} + \frac{1}{s}$ to get:

Corollary 2. *For all $n > 0$* $\mathbb{E}[C_n \log C_n] = n \log n \frac{1}{n_0} + n\frac{(1-H_{n_0})\log e}{n_0} + \log n + (\frac{3}{2} + \frac{1}{2n_0} - H_{n_0})\log e + O(n^{-1})$.

III. Main theoretical results

From [11] we know that

Lemma 4. $H(G_n) - H(S(G_n)) = \mathbb{E}[\log \Gamma(G_n)] - \mathbb{E}[\log |\operatorname{Aut}(G_n)|]$

where $|\operatorname{Aut}(G_n)|$ is the number of automorphisms of G_n and $\Gamma(G_n)$ is a size of a set of relabelings of G_n such that it produces a graph generated with positive probability from G_0 in this model. Here $H(G_n)$ denote the entropy of a *labelled* graph and $H(S(G_n))$ denote the entropy of an *unlabelled* graph, or a structural entropy of G_n.

The proofs of the following theorems were moved to the respective appendices.

Lemma 5. $\mathbb{E}[\log |\operatorname{Aut}(G_n)|] = n \log n - nH_{n_0} \log e + \frac{3n_0}{2} \log n + O(1)$

Lemma 6. $\mathbb{E}[\log \Gamma(G)] = n \log n - n \log e + \frac{2n_0+1}{2} \log n + O(1)$

Theorem 1. $H(S(G_n)) = (n_0 - 1)\log n + O(1)$

Theorem 2. $H(G_n) = n(H_{n_0} - 1)\log e + \frac{n_0-1}{2}\log n + O(1)$

Proof. The theorem follows straightly from the application of Lemma 4 to Theorems 5, 6 and 1. □

IV. Algorithmic results

A. Retrieval of parameters from G_n

Theorem 3. *For a given labelled G_n or its unlabelled structure $S(G_n)$, then we can retrieve its n, n_0 and G_0 – in case of structure up to automorphisms of G_n.*

Proof. For a given labelled G_n let $(w_1, w_2, \ldots, w_{n+n_0})$ be its order of vertices. It is sufficient to find the smallest k such that $N(w_k) = N(w_i)$ for some $1 \le i < k$. Then $n_0 = k - 1$ and G_0 is induced by a sequence $(w_1, \ldots, w_{k-1})$.

The case for unlabelled graphs is similar: we know (for details see Lemma 7) that the sequence of the first n_0 vertices of the graph (that is, G_0) contains exactly one vertex from each set $\{v\colon A(v) = u_i\}$.

Since from Lemma 1 we know that $A(v) = A(w)$ iff $N(v) = N(w)$, it sufficient to scan all vertices of G_n and split them into sets such that v and w belongs to the same set iff $N(v) = N(w)$. 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. □

B. Unlabelled graphs

A sequence $(C_{i,n})_{i=1}^{n_0}$ together with G_0 describes completely unlabelled graph $S(G_n)$ – yet one may have up to $n_0 \log n_0$ redundant descriptions due to the possible permutations of variables. Nevertheless, it is obvious that $H(S(G_n)) = H(C_{i,n}) + O(1)$.

A trivial algorithm CompressUnlabelledOptimal would write down a sequence $(C_{i,n})_{i=1}^{n_0}$ associated with our G_n as $\log n$-bit numbers. This requires always $n_0 \log n$ bits, so $\mathbb{E}L_{SU}(n) = n_0 \log n + O(1)$. Due to Theorem 2 this is $1 + \frac{1}{n_0 - 1}$-approximate algorithm to our problem.

The pseudo code of the optimal algorithm, called CompressUnlabelledOptimal and based upon arithmetic coding, is as follows:

function CompressUnlabelledOptimal($S(G_n)$, $S(G_0)$)

Fix a permutation π from $[1, 2, \ldots, n_0]$ to the vertices of $S(G_0)$

Fix any ordering $(v_1, \ldots, v_n)$ of the vertices of $S(G_n) \setminus S(G_0)$

$a \leftarrow 0$, $b \leftarrow 1$

for $i = 1, 2, \ldots, n_0$ **do**

$C[i] = 1$

for $j = 1, 2, \ldots, n$ **do**

if $N(v_j) = N(u_i)$ **then**

$C[i] = C[i] + 1$

for $i = n_0, n_0 - 1, \ldots, 1$ **do**

$start \leftarrow \frac{\Gamma(n+1)\Gamma(i)}{\Gamma(n+i)} \sum_{j=1}^{k-1} \frac{\Gamma(n+n_0-j)}{\Gamma(n+n_0-1)}$

$end \leftarrow start + \frac{\Gamma(n+1)\Gamma(k)\Gamma(n+n_0-k)}{\Gamma(n+k)\Gamma(n+n_0-j)}$

$b \leftarrow a + (b-a) * start$

$a \leftarrow a + (b-a) * end$

$n \leftarrow n - C[i] + 1$

$p \leftarrow b - a$, $x \leftarrow \dfrac{a+b}{2}$

return $S(G_0)\|\pi\|$[first $\lceil -\log p \rceil + 1$ bits of x]

Theorem 4. *Algorithm* CompressUnlabelledOptimal *is optimal up to a constant term for unlabelled graphs compression, when graph is generated by a full duplication model.*

Proof. It is sufficient to observe that

$$\mathbb{P}((C_{i,n})_{i=1}^{n_0} = (k_i + 1)_{i=1}^{n_0}) \tag{4}$$

$$= \mathbb{P}((C_{i,n})_{i=1}^{n_0} = (k_i + 1)_{i=1}^{n_0} | C_{n_0,n} = k_{n_0} + 1)\mathbb{P}(C_{n_0,n} = c_{n_0}) \tag{5}$$

$$= \mathbb{P}((C_{i,n})_{i=1}^{n_0-1} = (k_i + 1)_{i=1}^{n_0-1} | \sum_{i=0}^{n_0-1} C_{i,n} = k_{n_0} + 1)\mathbb{P}(C_{n_0,n} = k_{n_0} + \tag{6}$$

$$= \mathbb{P}((C_{i,n-k_{n_0}})_{i=1}^{n_0-1} = (k_i + 1)_{i=1}^{n_0-1}) f(k_i | \sum_{j \le i} k_j, 1, n_0 - 1) \tag{7}$$

where f is a probability mass funtion of beta-binomial distribution.

That is, the marginal distribution of a Dirichlet multinomial distribution is beta-binomial distribution. And if we fix a value of one coordinate to c for Dirichlet multinomial distribution, it is known that the resulting distribution is also Dirichlet multinomial, but with $n_0 - 1$ coordinates and all values summing up to $n - c$.

We repeat this process – but note that the sequence stops for $n_0 = 2$:

$$\mathbb{P}((C_{i,n})_{i=1}^{2} = (k_i + 1)_{i=1}^{2}) \tag{8}$$

$$= \mathbb{P}(C_{1,n} = k_1 + 1 | C_{1,n} = k_1 + 1)\mathbb{P}(C_{2,n} = k_1 + 1) \tag{9}$$

$$= f(k_1 | k_1 + k_2, 1, n_0 - 1) \tag{10}$$

Putting all together we get

$$\mathbb{P}((C_{i,n})_{i=1}^{n_0} = (k_i+1)_{i=1}^{n_0}) \tag{11}$$

$$= \prod_{i=2}^{n_0} f(c_i|\sum_{j\le i} k_j, 1, n_0 - i) \tag{12}$$

$$= \prod_{i=2}^{n_0} (n_0-1)\frac{\Gamma(\sum_{j\le i} k_j + 1)\Gamma(\sum_{j<i} k_j + n_0 - i)}{\Gamma(\sum_{j\le i} k_j + n_0 - i + 1)\Gamma(\sum_{j<i} k_j + 1)} \tag{13}$$

$$= (n_0-1)!\frac{\Gamma(n+1)}{\Gamma(n+n_0-1)} = nB(n,n0) \tag{14}$$

It matches the entropy, so by the properties of arithmetic encoding (see [5]), we know that $\mathbb{E}L_O(S(G_n)) \le H((C_{i,n})_{i=0}^{n_0}) + 2 + \mathbb{E}L_O(S(G_0)) + \mathbb{E}L_O(\pi) = H(S(G_n)) + O(1)$. □

C. Labelled graphs

Labelled graph G_n is equivalent to a sequence $(A(v_i))_{i=1}^{n}$ for a given (labelled) G_0 (which obviously can be encoded on the constant number of bits).

A trivial algorithm CompressLabelledSimple just writes all $A(v_i)$ as a $\log n_0$-bit number. Clearly, this gives us always a codeword with length $n \log n_0$ and therefore (as we need to compress G_0) $\mathbb{E}L_{SL}(n) = n\log n_0 + O(1)$. Note that from Theorem 2 we know that this algorithm is asymptotically $(1 + \frac{1-\gamma}{\log n_0})$-approximate, where γ is Euler-Mascheroni constant.

This sequence is random with $\mathbb{P}(A(v_i) = u_j) = \frac{C_{j,i-1}}{i+n_0-1}$ 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 pseudo code of the optimal algorithm is as follows:

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

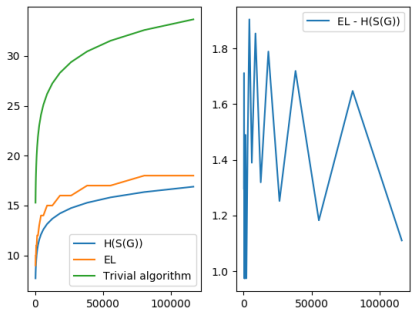
Theorem 5. *Algorithm* CompressLabelledOptimal *is optimal up to a constant term for unlabelled graphs compression, when graph is generated by a full duplication model.*

Proof. By the properties of arithmetic encoding (see [5]), we know that $\mathbb{E}L_O(G_n) \le H(G_n|G_0) + 2 + \mathbb{E}L_O(G_0) = H(G_n) + O(1)$. □

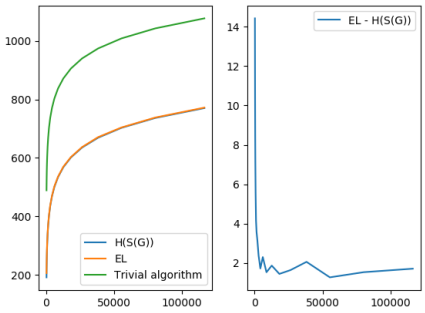
Note that CompressLabelledOptimal and CompressLabelledSimple differ only in that the first updates at each step the probabilities and the second fixes them to a constant $\frac{1}{n_0}$.

V. Experimental results

We implemented our compression algorithms for both labelled and unlabelled cases and run on a synthetic data, generated according to the vertex duplication rules. Figure **??** reports the results of compression for unlabelled graphs, figure 3 for labelled graphs.



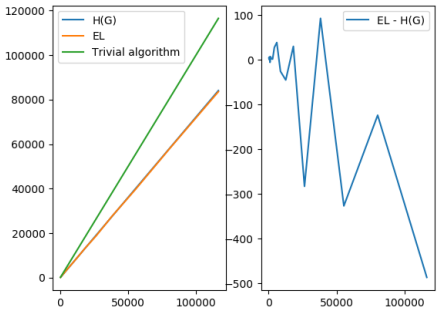
(a) $n_0 = 2$



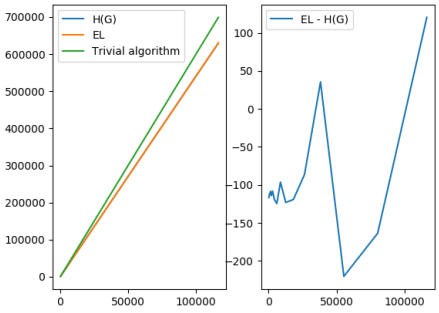
(b) $n_0 = 64$

Fig. 2: Compression for unlabelled graphs

As we see, for unlabelled graphs the error does converge to values less than 2, which is consistent with the properties of arithmetic encoding. For labelled graphs we observe divergence in error – which is apparently very small compared to the size of graphs. Unfortunately, the size of sample space for is very large compared to the size of sampled graphs, therefore the average length of compressed graphs cen be even negative.



(a) $n_0 = 2$



(b) $n_0 = 64$

Fig. 3: Compression for labelled graphs

ACKNOWLEDGMENT

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, by NSF Grant CCF-1524312, and NIH Grant 1U01CA198941-01.

REFERENCES

[1] R. Albert, A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys*, 74:47-97, 2002.

[2] S. Boccaletti, D.-U. Hwang, and V. Latora. Growing hierarchical scale-free networks by means of nonhierarchical processes. *International Journal of Bifurcation and Chaos*, 17(7):2447-2452, 2007.

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

[4] F. Chung, L. Lu, T. G. Dewey, D. J. Galas, Duplication Models for Biological Networks, *Journal of Computational Biology* 10(5):677-687, 2003.

[5] T. Cover, J. Thomas, *Elements of information theory*, John Wiley and Sons, 2006.

[6] P. Flajolet. Singularity analysis and asymptotics of binomial sums. *Theor. Comput. Sci.*, 215(1-2):371-381, 1999.

[7] I. Ispolatov, P. L. Krapivsky, and A. Yuryev. Duplication-divergence model of protein interaction network. Phys. Rev. E, 71:061911, 2005.

[8] I. Ispolatov, P. L. Krapivsky, I. Mazo, and A. Yuryev. Cliques and duplication-divergence network growth. New Journal of Physics, 2005.

[9] P. Jacquet, W. Szpankowski. Entropy computations via analytic depoissonization. *IEEE Transactions on Information Theory*, 45(4):1072-1081, 1999.

[10] J. Kim, P.L. Krapivsky, B. Kahng, S. Redner. Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev. E*, 66:055101, 2002.

[11] T. Łuczak, A. Magner, W. Szpankowski. Structural Information in Graphs: Symmetries and Admissible Relabelings.

[12] Mark Newman, Networks: An Introduction, Oxford University Press 2010.

[13] R. V. Solé, R. Pastor-Satorras, E. Smith. Evolving protein interaction networks through gene duplication, *Journal of Theoretical Biology*, 222:199-210, 2003.

[14] D. J. de S. Price. A general theory of bibliometric and other cumulative advantage processes *J. Amer. Soc.Inform. Sci.*, 27(5):292-306, 1976.

[15] A. Raval. Some asymptotic properties of duplication graphs. *Phys. Rev. E*, 68:066119, 2003.

[16] M. Shao, Y. Yang, J. Guan, S. Zhou, Choosing appropriate models for protein-protein interaction networks: a comparison study. *Briefings in Bioinformatics*, 15(5):823-838, 2014.

[17] A. Magner, W. Szpankowski, K. Turowski. Lossless Compression of Binary Trees with Correlated Vertex Names, *ISIT 2016*: 1217-1221.

[18] Z. Gołębiewski, A. Magner, W. Szpankowski. Entropy of Some General Plane Trees (with Z. Golebiewski and A. Magner), *ISIT 2017*: 1563-1567.

APPENDIX

A. Proof of the Lemma 2

From the definition of beta-binomial distribution as a compound of beta and binomial distributions:

$$\mathbb{E}[\log X] = \sum_{k=0}^{n} \log(k+1) \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} b(p,\alpha,\beta) dp \tag{15}$$

$$= \int_0^1 b(p,\alpha,\beta) dp \sum_{k=0}^{n} \log(k+1) \binom{n}{k} p^k (1-p)^{n-k} \tag{16}$$

(17)

where $b(p,\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$ – a probability density function for beta distribution.

We know from [6], [9] that

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \log(k+1) = \log(np+1) + O(n^{-1}) \tag{18}$$

and

$$\log(np+1) = \log n + \log p + O(n^{-1}) \tag{19}$$

hence

$$\mathbb{E}[\log X] = \int_0^1 b(p,\alpha,\beta)[\log n + \log p + O(n^{-1})] dp \tag{20}$$

$$= \log n \int_0^1 b(p,\alpha,\beta) dp + \int_0^1 b(p,\alpha,\beta) \log p dp + O(n^{-1}) \tag{21}$$

$$= \log n + (\psi(\alpha) - \psi(\alpha+\beta)) \log e + O(n^{-1}) \tag{22}$$

B. Proof of the Lemma 3

Similarly as above:

$$\mathbb{E}[X\log X]=\sum_{k=0}^{n}(k+1)\log(k+1)\int_0^1\binom{n}{k}p^k(1-p)^{n-k}b(p,\alpha,\beta)dp \tag{23}$$

$$=\int_0^1 b(p,\alpha,\beta)dp\sum_{k=0}^{n}(k+1)\log(k+1)\binom{n}{k}p^k(1-p)^{n-k} \tag{24}$$

$$\tag{25}$$

In this case from [6], [9] we get:

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}(k+1)\log(k+1) \tag{26}$$

$$=(np+1)\log(np+1)+\frac{np(1-p)\log e}{2(np+1)}+O(n^{-1}) \tag{27}$$

$$=np\log n+np\log p+\log n+\log p+\frac{(3-p)\log e}{2}+O(n^{-1}) \tag{28}$$

Then,

$$\mathbb{E}[X\log X]=\int_0^1 b(p,\alpha,\beta)[np\log n+np\log p \tag{29}$$

$$+\log n+\log p+\frac{(3-p)\log e}{2}+O(n^{-1})]dp \tag{30}$$

$$=n\log n\int_0^1 b(p,\alpha,\beta)pdp+n\int_0^1 b(p,\alpha,\beta)p\log pdp \tag{31}$$

$$+\log n\int_0^1 b(p,\alpha,\beta)dp+\int_0^1 b(p,\alpha,\beta)\log pdp \tag{32}$$

$$+\frac{3\log e}{2}\int_0^1 b(p,\alpha,\beta)dp-\frac{1}{2}\int_0^1 b(p,\alpha,\beta)pdp+O(n^{-1}) \tag{33}$$

$$=n\log n\frac{\alpha}{\alpha+\beta}+n\frac{\alpha\log e(\psi(\alpha+1)-\psi(\alpha+\beta+1))}{\alpha+\beta} \tag{34}$$

$$+\log n+(\psi(\alpha)-\psi(\alpha+\beta)+\frac{3}{2}-\frac{\alpha}{2(\alpha+\beta)})\log e \tag{35}$$

$$+O(n^{-1}) \tag{36}$$

Under the assumption that $|\operatorname{Aut}(G_0)|=1$ we have $\mathbb{E}[\log\operatorname{Aut}(G_n)]=\mathbb{E}\left[\log\prod_{i=1}^{n_0}C_{i,n}!\right]$. 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 freely – but if $A(v)\neq A(w)$, there does not exist any automorphism σ such that $\sigma(v)=w$ and $\sigma(w)=v$.

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

We use Stirling approximation together with Corollary 1 and 2 to obtain:

$$\begin{aligned}\mathbb{E}[\log C_n!]&=\mathbb{E}[C_n\log C_n]-\log e\mathbb{E}C_n+\frac{1}{2}\mathbb{E}[\log C_n]+O(1)\\&=\mathbb{E}[C_n\log C_n]-n\frac{\log e}{n_0}+\frac{1}{2}\mathbb{E}[\log C_n]+O(1)\\&=n\log n\frac{1}{n_0}+n\frac{(1-H_{n_0})\log e}{n_0}+\log n-n\frac{\log e}{n_0}+\log n\frac{1}{2}+\end{aligned}$$

and finally:

$$\mathbb{E}[\log\operatorname{Aut}(G_n)]=n\log n-nH_{n_0}\log e+\frac{3n_0}{2}\log n+O(1)$$

G_n has $n+n_0$ vertices, therefore the trivial bound is $\Gamma(G_n)\leq(n+n_0)!$.

We can do the exact computation of $\Gamma(G_n)$ using the following lemma:

Lemma 7. *Let π be a relabeling of G_n such that it produces a positive-probability graph under full duplication model. If $1\leq\pi(i)\leq\pi(j)\leq n_0$, then $A(w_i)\neq A(w_j)$ for any $i\neq j$.*

Proof. First, let us note that n_0 vertices in any given sequence has to be isomporphic to a given G_0.

Let $(w_1,w_2,\dots w_{n+n_0})$ be any permutation of vertices of G_n. Let us define $F(i)=\{j\colon A(w_j)=u_i\}$.

If $|F(i)\cap\wedge\{1\leq j\leq n_0\}|\geq 2$ for any i, then there exists in G_0 two identical vertices – which contradicts the assumption that $|\operatorname{Aut}(G)|=1$.

If $|F(i)\cap\wedge\{1\leq j\leq n_0\}|=0$ for any i, then let $k=\min F(i)$. Clearly w_k has to be a copy of some w_j for $1\leq j\leq k$, but all other vertices have different ancestors and (by Lemma 1) nonequal set of neighbours – a contradiction.

Therefore $|F(i)|=1$ for all $1\leq i\leq n_0$. □

Clearly, any permutation of first n_0 vertices is admissible, since every order of vertices in G_0 is admissible. Moreover, any permutation of last n vertices is admissible, because we may assume that all chosen vertices for replication are only these already existing in G_0. This, together with a fact that $C_{i,n}=|F(i)|$ lead us to the formula:

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

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

and the result $\mathbb{E}[\log\Gamma(G_n)]=n\log n-n\log e+(n_0+\frac{1}{2})\log n+O(1)$ follows from the Stirling approximation.

As before mentioned, $(C_{i,n})_{i=1}^{n_0}$ has Dirichlet multinomial distribution for $K = n_0$ and $\alpha_i = 1$ for $1 \leq i \leq n_0$. Therefore we know that

$$\mathbb{P}((C_{i,n})_{i=1}^{n_0} = (k_i{+}1)_{i=1}^{n_0}) = nB(n, n_0) \prod_{i=1}^{n_0} k_i^{-1} B(k_i, 1)^{-1} = nB(n, n_0)$$

leading to the desired result:

$$\begin{aligned}
H(S(G_n)) &= H(C_{i,n}) + O(1) \\
&= -\sum_{(c_i)} \mathbb{P}((C_{i,n})_{i=1}^{n_0}) \log \mathbb{P}((C_{i,n})_{i=1}^{n_0}) \\
&= -\sum_{(c_i)} nB(n, n_0) \log(nB(n, n_0)) = -\log n - \log B(n, n_0) \\
&= -\log n - (n + \frac{1}{2}) \log n - (n + n_0 + \frac{1}{2}) \log(n + n_0) + O(1) \\
&= -\log n - n_0 \log n + O(1) = (n_0 - 1) \log n + O(1)
\end{aligned}$$

The last two lines follow respectively from Stirling approximation and Taylor expansion of $\log(n + n_0)$.

First, let us note (for the fixed, asymmetric G_0) the one-to-one relationship between G_n and a sequence $(A(v_1), \ldots, A(v_n))$. Therefore, the entropy of G_n i equal to the entropy of a sequence $(A(v_1), \ldots, A(v_n))$ generated according to the rules.

This is in fact a joint distribution of n sequential steps of ball-and-urn process with n_0 urns, each containing one ball at the beginning. Every chosen ball is removed, its color is reported as i-th value and two new balls of the same color are added instead. The number of balls in urns at a given time is also known as a Dirichlet-multinomial distribution with $K = n_0$ and α_i = 1 for $1 \leq i \leq n_0$.

We can compute the total entropy of this distribution the following way. First, start with chain rule:

$$H(G_n|G_0) = \sum_{i=1}^{n} H(G_i|G_{i-1})$$

Then, for every $1 \leq i \leq n$:

$$\begin{aligned}
H(G_i|G_{i-1}) &= \sum_G p_{i-1}(G) H(G_i|G_{i-1} = G) \\
&= \sum_G p_{i-1}(G) \sum_{G'} p_i(G'|G) \log \frac{1}{p_i(G'|G)}
\end{aligned}$$

where $p_i(G) = \mathbb{P}(G_i = G)$, $p_i(G'|G) = \mathbb{P}(G_i = G'|G_{i-1} = G)$.

Now observe that $p_i(G'|G) = \mathbb{P}(A(v_i) = u_j) = \frac{C_{j,i-1}(G)}{i+n_0-1}$ if G' is equal to G with attached copy of u_j for some $1 \leq j \leq n_0$. If G' and G do not meet this condition, then clearly the conditional probaility is equal to 0. This lead us to:

$$\begin{aligned}
H(G_i|G_{i-1}) &= \sum_G p_{i-1}(G) \sum_{j=1}^{n_0} \frac{C_{j,i-1}(G)}{i + n_0 - 1} \log \frac{i + n_0 - 1}{C_{j,i-1}(G)} \\
&= \sum_{j=1}^{n_0} \mathbb{E}\left[\frac{C_{j,i-1}}{i + n_0 - 1} \log \frac{i + n_0 - 1}{C_{j,i-1}}\right] \\
&= n_0 \mathbb{E}\left[\frac{C_{i-1}}{i + n_0 - 1} \log \frac{i + n_0 - 1}{C_{i-1}}\right] \\
&= \frac{n_0}{i + n_0 - 1} \left(\log(i + n_0 - 1) \mathbb{E} C_{i-1} - \mathbb{E}[C_{i-1} \log C_{i-1}]\right) \\
&= \log(i + n_0 - 1) - \frac{n_0}{i + n_0 - 1} \mathbb{E}[C_{i-1} \log C_{i-1}]
\end{aligned}$$

We recall Corollary 2 and slightly rearrange the terms:

$$\begin{aligned}
\mathbb{E}[C_{i-1} \log C_{i-1}] = \log(i-1) \frac{i + n_0 - 1}{n_0} + (1 - H_{n_0}) \log e \frac{i + n_0 - 1}{n_0} \\
+ (\frac{1}{2} + \frac{1}{2n_0}) \log e + O(i^{-1})
\end{aligned}$$

Note that this formula holds only for $i > 1$, but from the definition it is straightforward that $\mathbb{E}[C_0 \log C_0] = 0$.

Putting this all together we get

$$\begin{aligned}
H(G_n) &= \sum_{i=1}^{n_0} (\log(n + i - 1) - \log i) - (n - 1)(1 - H_{n_0}) \log e \\
&\quad - \frac{n_0 + 1}{2} (H_{n+n_0-1} - H_{n_0-1}) \log e + \sum_{i=1}^{n} O(i^{-1}) + O(1) \\
&= n_0 \log n - (1 - H_{n_0})(n - 1) \log e \\
&\quad - \frac{n_0 + 1}{2} \ln n \log e + \sum_{i=1}^{n} O(i^{-1}) + O(1) \\
&= (H_{n_0} - 1) n + \frac{n_0 - 1}{2} \log n + O(\log n)
\end{aligned}$$

but, as we see from the theorem, the error term is actually smaller ($O(1)$ instead of $O(\log n)$).