Summary of Professional Achievements

Krzysztof Turowski

September 27, 2022

1. Diplomas and scientific degrees

- 2010 – MSc in Telecommunications. Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology. Thesis title: *„Interior design project using acoustic modeling systems”*. Supervisor: prof. Bożena Kostek.
- 2011 – MSc in Computer Science. Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology. Thesis title: *„Backbone colouring problem”*. Supervisor: prof. Marek Kubale.
- 2016 – MA in Philosophy. Faculty of Social Sciences, University of Gdańsk University of Technology. Thesis title: *„Values and free will in the ethics of Nicolai Hartmann”*. Supervisor: prof. Stanisław Judycki.
- 16.06.2015 – PhD in Technical Sciences, discipline: Computer Science. Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology. Thesis title: *„Analysis of algorithmic properties of backbone colouring problem”*. Supervisor: prof. Marek Kubale.

2. Employment

- November 2010 – September 2011 – assistant at Department of Algorithms and Systems Modelling, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland.
- October 2011 – September 2015 – lecturer at Department of Algorithms and Systems Modelling, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland.
- October 2015 – February 2016 – assistant professor at Department of Algorithms and Systems Modelling, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland.
- August 2015 – November 2015 – post-doctoral research scholar at Center for Science of Information, Purdue University, IN, USA.
- March 2016 – April 2018 – software engineer at Google Cloud Platform/Compute Engine, Google Poland sp. z.o.o., Warsaw.
- May 2018 – September 2019 – post-doctoral research scholar at Center for Science of Information, Purdue University, IN, USA.
- since October 2019 – assistant professor at Theoretical Computer Science Department, Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland.

3. Habilitation achievement

This section contains the description of the achievements, set out in art. 219 para 1 point 2 of the Act.

3.1. The title of the achievement

Structural analysis and compression for duplication random graph models

3.2. Articles included in the habilitation achievement

[A1] Krzysztof Turowski, Wojciech Szpankowski, Towards Degree Distribution of a Duplication-Divergence Graph Model, *The Electronic Journal of Combinatorics*, 28(1) (2021), P1.18.

[A2] Alan Frieze, Krzysztof Turowski, Wojciech Szpankowski, *Degree Distribution for Duplication-Divergence Graphs: Large Deviations*, 46th International Workshop on Graph-Theoretic Concepts in Computer Science, WG 2020, Leeds, UK, June 24-26, 2020. Lecture Notes in Computer Science 12301, pages 226-237.

[A3] Alan Frieze, Krzysztof Turowski, Wojciech Szpankowski, *The concentration of the maximum degree in the duplication-divergence models*, Proceedings of 27th International Conference of Computing and Combinatorics, COCOON 2021, Tainan, Taiwan, October 24-26, 2021. Lecture Notes in Computer Science 13025, pages 413-424.

[A4] Philippe Jacquet, Krzysztof Turowski, Wojciech Szpankowski, *Power-Law Degree Distribution in the Connected Component of a Duplication Graph*, 31st International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, AofA 2020, June 15-19, 2020, Klagenfurt, Austria (Virtual Conference). LIPIcs 159, pages 16:1-16:14.

[A5] Krzysztof Turowski, Abram Magner, Wojciech Szpankowski, Compression of Dynamic Graphs Generated by a Duplication Model, *Algorithmica* 82(9) (2020), pages 2687-2707.

- conference version: Krzysztof Turowski, Abram Magner, Wojciech Szpankowski, *Compression of Dynamic Graphs Generated by a Duplication Model*, 56th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2018, Monticello, IL, USA, October 2-5, 2018, pages 1089-1096.

4. Discussion of the papers in the series and the results therein

4.1. Introduction

Graph theory has been applied to describe many complex systems found in the world, such as biological networks or social networks. This approach seems natural when we want to describe a system using some basic elements (represented by graph vertices) and their interactions (represented by graph edges). This allows us to analyze these structures both from a static point of view, i.e. description of their properties in a certain steady state, and from a dynamic point of view, i.e. description of network evolution and changes of its parameters in time.

The issues of structural analysis and compression of random graphs naturally developed from the analyses performed in classical graph theory and information theory. First, they grew out of the research on random graphs and their properties initiated by Paul Erdős and Alfréd Rényi [40] as early as 1959. This line of work, continued over the following decades, resulted in a number of results on popular graph parameters, such as the size of connected components, the maximum degree of a graph, the size of the maximum matching, and the chromatic number (see summaries of results e.g. in [16, 48, 67]).

Second, they also point directly to the search for appropriate random graph models that can be fitted to the structures found in the real world. For example, since the 1970s, there have been hypotheses that the evolutionary dynamics of protein interaction networks can be described by simple duplication and mutation rules [86, 106]. Moreover, the study of the properties of real networks has highlighted a number of their characteristic features:

- short paths between arbitrary vertices – e.g. *Six Degrees of Separation* in social networks, popularized by sociologist Stanley Milgram,
- degree distribution of vertices – the existence of few vertices of high degree (so-called hubs) and many of low degree.

Researchers have been primarily interested in scale-free networks, formally defined as those in which the percentage of vertices of degree k tends to a fixed value, proportional to $k^{-\gamma}$ for some fixed parameter γ. The popularity of this line of research has been enhanced by claims that such a characterization fits very well with many types of biological or social networks, e.g., protein interactions, scientific article citations, or Internet hyperlinks [2, 30]. As part of the theoretical search for ways to generate graphs with such properties, more models were created, the most famous proposed by Barabási and Albert [8] or Watts and Strogattz [102]. And these models, too, were becoming the subjects of careful probabilistic-theoretic analysis [48, 85, 99]. For example, it is proved that the graphs generated by the Barabási-Albert model are indeed characterized by the scale-free property with parameter $\gamma = 3$ [17].

Within information theory, on the other hand, the need to go beyond the treatment proposed by Shannon has been growing for some time. In the keynote text, "Three Great Challenges for Half Century Old Computer Science," Frederick P. Brooks Jr. noted that the challenge for the future is to find the proper definition and measure of information embodied in structures [22]. A natural direction for research is to apply this approach to data in the form of graphs and to try to quantify information created and expressed through the emergence and evolution of such structures [96].

There is a related, but purely practical motivation: in the era of Big Data, one of the types of large-scale data stored and processed is graph data of e.g. social networks. Finding an efficient way to represent them succinctly could bring tangible practical benefits. Even the development of universal, asymptotically optimal compression algorithms, such as the Lempel-Ziv algorithms of [108, 109], did not end the research on compression at all, as there is still a sufficiently large efficiency gap in practice, achievable at the cost of, e.g., some additional assumptions about the nature of the data source.

In the approach, called "Shannon meets Turing," one can primarily study the entropy of graphs (for labeled vertices) and structures (for indistinguishable vertices) according to Shannon's definition over the probabilistic space of graphs generated from a given model. The results obtained in this way provide a lower bound estimate of the compressibility of graphs generated from this model. The other side of this research is the search for algorithms that can achieve the appropriate compression ratio with the best possible accuracy guarantees.

In the course of this research, it became clear that there is a momentous relationship between the structural study of random graphs and their information limitations. First of all, the knowledge of the distribution of the number of automorphisms – or, more precisely: the knowledge that the generated graphs are asymmetric with high probability – turned out to be the cornerstones for determining the entropy values and the structural entropy for both Erdős-Rényi and Barabási-Albert random graph models [28, 80]. For many other graph models, in turn, it was possible to obtain some asymptotic lower entropy estimates based, among other things, on proofs of the probabilistic behavior of the maximum degree of graphs or other simple graph structures [27].

In the development of lossless compression approaches, in addition to universal algorithms and the creation of heuristics that are effective in practice (see the review of existing algorithms in [10]), an important task is to develop algorithms that provide some average and pessimistic guarantees for specific random graph models. In particular, algorithms that have been developed for the Erdős-Renyí and Barabási-Albert models mentioned above achieve optimal compression asymptotically with respect two leading terms of entropy [28, 80].

As it can be seen from the above introduction, much of the research interest has focused on the two most popular models: Erdős-Renyí and Barabási-Albert. However, despite their simple definitions and very interesting structural properties, it seems that they do not accurately describe many types of real-world networks [33, 92]. In particular, note that for biological networks, for example, a scale-free factor of $\gamma < 2$ is often identified, which does not fit any of the above [30] models. Moreover, it seems that the fit of these models to existing networks could not be supported by the rationale behind just such and not other characteristics of the network formation mechanism.

The intuition that biological and social networks may arise through an intrinsic mechanism of duplication and mutation (also called *duplication-divergence*) provides a line of research to address these challenges. On the one hand, the evolutionary motivation for such duplication models is well rooted in biological considerations [106]. Also, in comparisons of different models with existing networks of, e.g., protein interactions or citations, we find that duplication models perform best when it comes to matching the degrees of vertices in a graph [33] or the distribution of small subgraphs [92]. On the other hand, non-rigorous computations for such models showed that for certain parameters they can generate scale-free graphs with appropriate coefficients [61, 85, 89], which makes them equally interesting for theorists who value precisely these features[1].

It is this very issue that became the focus of research in the series of publications that make up the presented achievement. The individual papers considered the probabilistic behavior of the duplication model by Solé and Pastor-Satorras [87, 93], which is a popular model that is also a generalization of another important model called pure duplication [30]. The exact and asymptotic behavior of various random variables such as the average degree of a graph, the degree of a fixed vertex, the maximum degree, and the scale factor γ were studied. The problem of graph entropy and compression for some special case of this model was also addressed. To achieve this goal, a number of combinatorial, probabilistic, and algorithmic approaches have been used, including tools developed in the field of analytic combinatorics to solve recursive relations and to derive theorems about the asymptotic behavior of variables.

4.2. Basic definitions

In this summary we use standard graph notation, e.g. after [37]: $V(G)$ and $E(G)$ denote the sets of vertices and edges of the graph G, $\mathcal{N}(G)$ – the set of neighbours of vertex u in G, $\deg_G(u) = |\mathcal{N}_G(u)|$ – the degree of vertex u in G. All the considered graphs are simple, without loops or multiple edges. Since the notation G_t is often used to denote a random graph on t vertices, for brevity we use the notation $\deg_t(u)$ instead of $\deg_{G_t}(u)$, $cN_t(u)$ instead of $cN_{G_t}(u)$ etc.

The basic parameters are the *average degree* $D(G)$ of the graph G, defined as

$$D(G) = \frac{1}{|V(G)|} \sum_{v \in |V(G)} \deg_G(u),$$

and the *average square of degree* $D_2(G)$, i.e.

$$D_2(G) = \frac{1}{|V(G)|} \sum_{v \in |V(G)} \deg_G^2(u).$$

Any random graph model induces some probability distribution on the set of all graphs with a fixed set of vertices. Since for the vast majority of random graphs the number of vertices belongs to the input of the model, it is natural to define random variables over the set of all graphs with a given number of vertices t, denoted as $\mathcal{G}_t$. For example, the average degree and the average square of degree for random graph models are defined as:

$$\mathbb{E}[D(G_t)] = \sum_{G \sim \mathcal{G}_t} \Pr[G_t = G] D(G),$$
$$\mathbb{E}[D_2(G_t)] = \sum_{G \sim \mathcal{G}_t} \Pr[G_t = G] D_2(G).$$

In the literature, the variables $D(G_t)$ and $D_2(G_t)$ are also encountered under the name of the first and second moments of the degree distribution, respectively. Analogous definitions of random

[1] Note, however, that these computations were not rigorous enough, so that at least some of them were later refuted and corrected e.g. in [57]

variables can also be formulated for any other known graph parameter, e.g. for the degree of a fixed vertex $\deg_t(s)$ or for the maximum degree of a graph $\Delta(G_t)$.

Another important parameter is the number and fraction of vertices of a given degree, defined formally:

$$\begin{aligned}\mathbb{E}[F_k(G_t)] &= \sum_{G\sim\mathcal{G}_t} \Pr[G_t = G]|\{v \in V(G)\colon \deg_G(v) = k\}|,\\ \mathbb{E}[f_k(G_t)] &= \frac{\mathbb{E}[F_k(G_t)]}{t}.\end{aligned}$$

In a similar way, we can also define the *entropy* of a graph according to the classical definition from information theory for any discrete probability distribution:

$$H(G_t) = -\sum_{G\sim\mathcal{G}_t} \Pr[G_t = G] \log_2 \Pr[G_t = G].$$

In this definition, we assume that $\mathcal{G}_t$ is the set of all pairwise non-isomorphic graphs with a set of vertices $\{1, \ldots, t\}$ i.e. each vertex has a unique label assigned to it[2]. Accordingly, one can define the *structural entropy* as

$$H(S_t) = -\sum_{S\sim\mathcal{S}_t} \Pr[S_t = S] \log_2 \Pr[S_t = S],$$

where $\mathcal{S}_t$ is the set of all non-isomorphic unlabeled graphs on t vertices.

The formal definition of the Solé and Pastor-Satorras duplication model $\mathtt{DD}(t, p, r)$ [87, 93], which is the main subject of research in the papers [A1]-[A5], is as follows: for given parameters $0 \le p \le 1$ and $0 \le r \le t_0$ and (often implicitly[3]) of a given initial graph G_{t_0} with $V(G_{t_0}) = \{1, \ldots, t_0\}$ for each $t = t_0, t_0 + 1, \ldots$ we build G_{t+1} from G_t according to the following procedure:

1. we add a new vertex $t + 1$ to the graph,
2. we randomly select a vertex u uniformly from the set $V(G_t) = \{1, \ldots, t\}$ and label u as *parent*$(t + 1)$,
3. for each vertex $i \in V(G_t)$:
   (a) if $i \in \mathcal{N}_t(\mathit{parent}(t + 1))$, then we add an edge between i and $t + 1$ with probability p,
   (b) if $i \notin \mathcal{N}_t(\mathit{parent}(t + 1))$, then we add an edge between i and $t + 1$ with probability $\frac{r}{t}$.

All draws, i.e. the parent vertex from the set of existing vertices and all decisions whether to add edges between the new vertex and existing vertices, are independent.

A special case of this model, considered in the literature [11, 29, 30], is the so-called *pure duplication* model. It contains only edges resulting from the duplication of vertices, but it has edges connecting new vertices to existing vertices outside the neighbourhood of their parents i.e. it is exactly the $\mathtt{DD}(t, p, 0)$ model.

4.3. Analysis of the behaviour of the degree distribution for duplicate random graph models

The starting point for the work on duplicate random graph models, in particular the model by Solé and Pastor-Satorras, was an observation made for many other random graph models in at least two independent papers [27, 80], according to which the behaviour of the random variables $\deg_t(s)$ and $\Delta(G_t)$ formed the basis for developing entropy estimates for individual models. However, since the Solé and Pastor-Satorras model itself was previously mainly an object of study based on computer simulations and approximation calculations, it was necessary to develop a rigorous theory of the

[2] Note that for parameters such as D or Δ it does not matter whether we are operating on a set of graphs with or without labels.

[3] Typically, G_{t_0} is assumed to be a complete graph, a choice that is as convenient from a theoretical point of view as it is variously evaluated from an application point of view. See also [59].

behaviour of the individual variables in the graphs, in the spirit of the approach of e.g. Bollobása [17]. The paper [A1] contains results describing the behaviour of the mean and variance of the basic variables (the mean degree in the graph and the degree of a fixed vertex), while the papers [A2] and [A3] extend the results to theorems about the concentration of the distribution of these variables and the maximum degree in the graph around the mean value.

The paper [A1] contains results on the distribution properties of two random variables defined for the model $\mathtt{DD}(t,p,r)$: the degree $\deg_t(s)$ of a given vertex s in G_t and the average degree of the graph $D(G_t)$. The first step was to look for the first moments of the relevant random variables:

Problem [A1]: what is the precise asymptotic growth rate of the functions $\mathbb{E}[\deg_t(s)]$, $\mathbb{E}[D(G_t)]$, $\mathrm{Var}[\deg_t(s)]$, and $\mathrm{Var}[D(G_t)]$?

The starting point of the proof was the observation that recursive equations can be constructed for the relevant variables. For example, for the expected values from the model definition, the following relations can be derived:

$$\mathbb{E}[\deg_{t+1}(s)] = \mathbb{E}[\deg_t(s)]\left(1+\frac{p}{t}-\frac{r}{t^2}\right)+\frac{r}{t}, \tag{1}$$

$$\mathbb{E}[D(G_{t+1})] = \mathbb{E}[D(G_t)]\left(1+\frac{2p-1}{t+1}-\frac{2r}{t(t+1)}\right)+\frac{2r}{t+1}. \tag{2}$$

It is then shown that the two variables in this model are closely related through the dependence of the initial value for the first recurrence $\mathbb{E}[\deg_s(s)]$ on the average degree of the graph in the previous iteration $\mathbb{E}[D(G_{s-1})]$. Again directly from the model definition, the degree of the last vertex is defined as

$$\deg_{t+1}(t+1) \sim Bin\left(\deg_t(parent(t+1)), p\right) + Bin\left(t-\deg_t(parent(t+1)), \frac{r}{t}\right),$$

i.e. $\deg_{t+1}(t+1)$ is a random variable that is a sum of two binomial variables with appropriate parameters. Thus, we can derive the following relationship between the expected values of the two variables:

$$\mathbb{E}[\deg_{t+1}(t+1)] = \left(p-\frac{r}{t}\right)\mathbb{E}[D(G_t)]+r. \tag{3}$$

It follows from the relation (3) that the determination of the growth rate $\mathbb{E}[D(G_t)]$ leads directly to the growth rate for the initial condition $\mathbb{E}[\deg_s(s)]$ for the equation (1) in the case where $s\to\infty$.

The techniques developed in the paper for solving the above equations are actually used to determine the asymptotic behaviour from general recursive relations of the form

$$\mathbb{E}[f(G_{n+1}) \mid G_n] = f(G_n)g_1(n)+g_2(n), \tag{4}$$

for the function $g_1(n)=\frac{W_1(n)}{W_2(n)}$ expressible as the quotient of polynomials W_1 and W_2 of the same degree. Clearly, the equations (1) and (2) are special cases of the form (4).

First, it is shown that for the solution of the above recurrence equation written in the form

$$\mathbb{E}[f(G_n)] = \prod_{k=n_0}^{n-1} g_1(k)\left(f(G_{n_0})+\sum_{j=n_0}^{n-1} g_2(j)\prod_{k=n_0}^{j}\frac{1}{g_1(k)}\right). \tag{5}$$

It can be shown that

$$\prod_{k=n_0}^{n-1} g_1(k) = \prod_{k=n_0}^{n-1}\frac{W_1(k)}{W_2(k)} = \prod_{i=1}^{d}\frac{\Gamma(n-a_i)}{\Gamma(n-b_i)}\frac{\Gamma(n_0-b_i)}{\Gamma(n_0-a_i)}$$

where Γ is the Euler gamma function, d is the degree of the polynomials W_1 and W_2, while a_i and b_i (for $i=1,\ldots,d$) are, respectively, the (complex) roots of W_1 and W_2.

Combined with the lemma below, determining the growth rate of the quotient of the Γ functions, this allowed us to determine the asymptotic growth rate[4] for the first factor in the solution of the general recurrence.

[4] More precisely, the full version of the following lemma allows us to determine not only the main factor of the asymptotic growth rate, but also the expansion with any desired precision.

Lemma 4.1 (Abramowitz, Stegun [1])**.** *For any $a, b \in \infty$ and $n \to \infty$ asymptotically it holds that*

$$\frac{\Gamma(n+a)}{\Gamma(n+b)} = n^{a-b}\left(1 + \frac{(a-b)(a+b-1)}{2n} + O\left(\frac{1}{n^2}\right)\right).$$

The next step was to use the observation that if the function $g_2(n)$ can be represented as a quotient of the Euler gamma function, then for finding the asymptotic growth rate of formulae in the form (5) it suffices to find the asymptotic expansion of expressions of the form

$$\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)}.$$

Using more advanced but similar tools to those described above, including also the theory of hypergeometric functions (also present in [1]), the following relations can be obtained:

Lemma 4.2. *Let $a_i, b_i \in \mathbb{R}$ for $i = 1, 2, \ldots, k$ ($k \in \mathbb{N}$) and $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$. Asymptotically, when $n \to \infty$ it holds that*

$$\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} = \begin{cases} \frac{1}{a-b+1} n^{a-b+1} + O\left(n^{\max\{a-b,0\}}\right) & \textit{dla } a+1 > b, \\ \ln n + O\left(1\right) & \textit{dla } a+1 = b. \end{cases}$$

Lemma 4.3. *Let $a_i, b_i \in \mathbb{R}$ for $i = 1, 2, \ldots, k$ ($k \in \mathbb{N}$) and $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$. For any $n \in \mathbb{N}_+$ it holds that*

$$\sum_{j=n}^{\infty} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} = \frac{\prod_{i=1}^{k} \Gamma(n+a_i)}{\prod_{i=1}^{k} \Gamma(n+b_i)} \; {}_{k+1}F_k\left[\begin{smallmatrix} n+a_1,\ldots,n+a_k,1 \\ n+b_1,\ldots,n+b_k \end{smallmatrix}; 1\right]$$

where $c_3 = p + \sqrt{p^2 + 2r}$, $c_4 = p - \sqrt{p^2 + 2r}$, while ${}_pF_q[\begin{smallmatrix} \mathbf{a} \\ \mathbf{b} \end{smallmatrix}; z]$ is a generalised hypergeometric function (see [1]). Moreover, asymptotically when $n \to \infty$ it is true that

$$\sum_{j=n}^{\infty} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} = \frac{1}{b-a\ -1} n^{a-b+1} + O(n^{a-b+2}).$$

With the above lemmas, we are able to find asymptotic growth rates for functions described by recursive equations of the form (5), provided that the functions $g_1(n)$ and $g_2(n)$ are of certain particular form. Since both these conditions are satisfied by the equation (2) describing $D(G_t)$, it can be proved that

Theorem 4.1. *Asymptotically, for $t \to \infty$ it holds that*

$$\mathbb{E}[D(G_t)] = \begin{cases} t^{2p-1} \frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} D(G_{t_0})(1+o(1)) & \textit{if } p \le \frac{1}{2},\ r = 0, \\ \frac{2r}{1-2p}(1+o(1)) & \textit{if } p < \frac{1}{2},\ r > 0, \\ 2r \ln t\,(1+o(1)) & \textit{if } p = \frac{1}{2},\ r > 0, \\ t^{2p-1} \frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}(1+o(1)) & \textit{if } p > \frac{1}{2}, \\ \quad \left(D(G_{t_0}) + \frac{2rt_0}{t_0^2+2pt_0-2r} \, {}_3F_2\left[\begin{smallmatrix} t_0+1,t_0+1,1 \\ t_0+c_3+1,t_0+c_4+1 \end{smallmatrix}; 1\right]\right) & \end{cases}$$

where $D(G_{t_0})$ is the average degree of the initial graph G_{t_0} and

$${}_3F_2\left[\begin{smallmatrix} a_1,a_2,a_3 \\ b_1,b_2 \end{smallmatrix}; z\right] = \sum_{l=0}^{\infty} \frac{(a_1)_l(a_2)_l(a_3)_l}{(b_1)_l(b_2)_l} \frac{z^l}{l!}$$

for the Pochhammer function $(a)_l = a\ (a+1)\ldots(a+l-1)$, $(a)_0 = 1$.

Applying the respective formulae for $\deg_t(s)$ we obtain that

Theorem 4.2. *Asymptotically, for $t \to \infty$, it holds that:*
(i) *for $s = O(1)$*

$$\mathbb{E}[\deg_t(s)] = \Theta(t^p),$$

(ii) *for $s = \omega(1)$ and $s = o(t)$*

$$\mathbb{E}[\deg_t(s)] = \begin{cases} \Theta\left(\left(\frac{t}{s}\right)^p s^{2p-1}\right) & \textit{if } p \le \frac{1}{2} \textit{ and } r = 0 \textit{ or if } p > \frac{1}{2}, \\ \Theta\left(\log\left(\frac{t}{s}\right)\right) & \textit{if } p = 0 \textit{ and } r > 0, \\ \Theta\left(\left(\frac{t}{s}\right)^p\right) & \textit{if } 0 < p < \frac{1}{2} \textit{ and } r > 0, \\ \Theta\left(\sqrt{\frac{t}{s}} \log s\right) & \textit{if } p = \frac{1}{2} \textit{ and } r > 0. \end{cases}$$

(iii) *for $s = \Theta(t)$*

$$\mathbb{E}[\deg_t(s)] = \begin{cases} \Theta\left(t^{2p-1}\right) & \textit{if } p \le \frac{1}{2},\ r = 0 \textit{ or } p > \frac{1}{2}, \\ \Theta(1) & \textit{if } 0 \le p < \frac{1}{2} \textit{ and } r > 0, \\ \Theta(\log t) & \textit{if } p = \frac{1}{2} \textit{ and } r > 0. \end{cases}$$

More precisely, not only asymptotic estimates but also exact formulae with complex leading coefficients (depending on s, p, r) for each case were presented in the paper [A1].

Analogous proofs were also carried out for the variances of the variables discussed above and it was obtained that:

Theorem 4.3. *Asymptotically for $t \to \infty$ it holds that*

$$\mathrm{Var}[D(G_t)] = \begin{cases} \Theta(1) & \textit{if } p < \frac{1}{2}, \\ \Theta(\log^2 t) & \textit{if } p = \frac{1}{2}, \\ \Theta(t^{4p-2}) & \textit{if } p > \frac{1}{2}. \end{cases}$$

(i) *for $s = O(1)$*

$$\mathrm{Var}[\deg_t(s)] = \begin{cases} \Theta(\log t) & \textit{if } p = 0, \\ \Theta(t^{2p}) & \textit{if } p > 0. \end{cases}$$

(ii) *for $s = \omega(1)$*

$$\mathrm{Var}[\deg_t(s)] = \begin{cases} \Theta\big(\log\left(\frac{t}{s}\right)\big) & \textit{if } p = 0, \\ \Theta\big(\left(\frac{t}{s}\right)^{2p}\big) & \textit{if } 0 < p < \sqrt{2} - 1, \\ \Theta\big(\left(\frac{t}{s}\right)^{2p} \log s\big) & \textit{if } p = \sqrt{2} - 1, \\ \Theta\big(\left(\frac{t}{s}\right)^{2p} s^{p^2+2p-1}\big) & \textit{if } p > \sqrt{2} - 1. \end{cases}$$

The paper [A2] extends the research presented in the paper [A1] with results on the asymptotic growth rate of random variables for the tails of the distribution of the variables $D(G_t)$ and $\deg_t(s)$ (for $s = O(1)$):

Problem [A2]: for which functions $f_l(t)$, $f_h(t)$ (respectively, $g_l(t)$, $g_h(t)$) occur $\Pr[D(G_t) \notin [f_l(t), f_h(t)]] = O(t^{-A})$ (respectively, $\Pr[\deg_t(s) \notin [g_l(t), g_h(t)]] = O(t^{-A})$) for any $A > 0$?

Can it be shown that the above relation holds for the functions $f_l(t), f_h(t) \in \widetilde{\Theta}(\mathbb{E}[D(G_t)])$ (respectively, $g_l(t), g_h(t) \in \widetilde{\Theta}(\mathbb{E}[\deg_t(s)])$)?

The first main result of [A2] is the proof of the lower bound for the random variable $D(G_t)$:

Theorem 4.4. *Asymptotically when $t \to \infty$ for $G_t \sim \mathtt{DD}(t, p, r)$ it holds that*

$$\begin{aligned} &\Pr[D(G_t) \ge A\,C\,\log^2(t)] = O(t^{-A}) && \textit{if } p < \frac{1}{2}, \\ &\Pr[D(G_t) \ge A\,C\,\log^3(t)] = O(t^{-A}) && \textit{if } p = \frac{1}{2}, \\ &\Pr[D(G_t) \ge A\,C\,t^{2p-1}\log^2(t)] = O(t^{-A}) && \textit{if } p > \frac{1}{2}. \end{aligned}$$

for some constant $C > 0$ and for any $A > 0$.

The proof of this theorem, like the rest of the results in the paper [A2], is based on a precise estimation of the behaviour of the moment-generating function (*moment-generating function*). Generalizing the model analysis leading to the equation (2), it can be shown that, for the random variable $D(G_t)$, we know that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda_{t+1}D(G_{t+1})\right) \mid G_t\right] &= \mathbb{E}\left[\exp\left(\lambda_{t+1}\left(\frac{t}{t+1}D(G_t) + \frac{2}{t+1}\deg_{t+1}(t+1)\right)\right) \mid G_t\right] \\ &= \exp\left(\frac{\lambda_{t+1}t}{t+1}D(G_t)\right)\mathbb{E}\left[\exp\left(\frac{2\lambda_{t+1}}{t+1}\deg_{t+1}(t+1)\right) \mid G_t\right].\end{aligned}$$

It follows from this that for any sequence of parameters $\lambda_t \to 0$ we have

$$\mathbb{E}\left[\exp\left(\lambda_{t+1}D(G_{t+1})\right) \mid G_t\right] \leq \exp\left(\lambda_{t+1}D(G_t)\left(1-\frac{2p-1}{t+1}\right)(1+O(\lambda_{t+1})) + \frac{2r\lambda_{t+1}}{t+1}(1+o(t^{-1})\right).$$

By choosing the appropriate values of λ_k for $k = t_0, \ldots, t-1, t$ and taking $\varepsilon_t \geq \lambda_k$ for all $k \leq t$ we obtain

$$\mathbb{E}\left[\exp\left(\lambda_{t+1}D(G_{t+1})\right)\right] \leq \exp\left(\lambda_{t_0}D(G_{t_0})\right)\left(\frac{t}{t_0}\right)^{2r\varepsilon_{t+1}+C_1}.$$

Finally, using Chernoff bound it follows that

$$\begin{aligned}\Pr[D(G_t) \geq \alpha\mathbb{E}[D(G_t)]] &= \Pr[\exp(D(G_t) - \alpha\mathbb{E}[D(G_t)]) \geq 1] \\ &\leq \exp\left(-\alpha\lambda_t\mathbb{E}[D(G_t)]\right)\mathbb{E}[\exp\left(\lambda_t D(G_t)\right)] \\ &\leq \exp\left(-\alpha\lambda_t\mathbb{E}[D(G_t)]\right)\exp\left(\lambda_{t_0}D(G_{t_0})\right)\left(\frac{t}{t_0}\right)^{2r\varepsilon_{t+1}+C_1}\end{aligned}$$

which, for appropriately chosen values of ε_t, λ_t and α, completes the proof of Theorem 4.4.

By an analogous approach, it is possible to show an analogous relation also for the random variable $\deg_t(s)$:

Theorem 4.5. *Asymptotically when $t \to \infty$ for $G_t \sim \mathtt{DD}(t,p,r)$ and $s = O(1)$ it holds that*

$$\Pr[\deg_t(s) \geq A\,C\,t^p\log^2(t)] = O(t^{-A})$$

for some constant $C > 0$ and for any $A > 0$.

The estimation of the left-hand side of the distributions of the variables $D(G_t)$ and $\deg_t(s)$ is performed by similar, although slightly more complex techniques. Ultimately, we obtain that:

Theorem 4.6. *Asymptotically when $t \to \infty$ for $G_t \sim \mathtt{DD}(t,p,r)$ it holds that*

$$\Pr\left[D(G_t) \leq \frac{C}{A}t^{2p-1}\log^{-3-\varepsilon}(t)\right] = O(t^{-A}).$$

for some constant $C > 0$ and for any $\varepsilon, A > 0$.

Analogous results were developed for the variable $begin_t(s)$ with $s = O(1)$:

Theorem 4.7. *Asymptotically when $t \to \infty$ for $G_t \sim \mathtt{DD}(t,p,r)$ and $s = O(1)$ it holds that*

$$\Pr\left[\deg_t(s) \leq \frac{C}{A}t^p\log^{-3-\varepsilon}(t)\right] = O(t^{-A})$$

for some constant $C > 0$ and for any $\varepsilon, A > 0$.

After obtaining a proof of the concentration of the random variables $D(G_t)$ and $\deg_t(s)$ around the respective mean values, one can ask the question of the analogous behaviour of the maximal degree of the graph $\Delta(G_t)$ – and the paper [A3] is devoted to answering exactly this question for $\frac{1}{2} < p \leq 1$. Of course, Theorem 4.7 directly implies an identical lower bound for the variable $\Delta(G_t)$, but a similar result from the upper bound does not carry over.

Problem [A3]: for which functions $f_l(t)$ and $f_h(t)$ $\Pr[\Delta(G_t) \notin [f_l(t), f_h(t)]] = O(t^{-A})$ holds for any $A > 0$?

The main result of [A3] is a theorem showing the concentration of maximum degree in this model:

Theorem 4.8. *For $\frac{1}{2} < p < 1$ asymptotically when $t \to \infty$ for $G_t \sim \mathtt{DD}(t, p, r)$ it is true that*

$$\Pr[\Delta(G_t) \notin [(1-\varepsilon)t^p, (1+\varepsilon)t^p \log^{5-4p}(t)]] = O(t^{-A})$$

for any constants $\varepsilon, A > 0$.

The proof of Theorem 4.8 is divided into three parts: separately we prove (a) the lower bound, (b) the upper bound for early vertices and (c) the upper bound for later vertices.

The main idea of the proof of part (b) is as follows: we are looking for such values of $(t_i)_{i=0}^k$ and such a sequence $(X_{t_i})_{i=0}^k$ that

1. $\deg_{t_0}(s) \leq X_{t_0}$ holds with high probability for $1 \leq s \leq t_0$,
2. $\deg_{t_{i+1}}(s) - \deg_{t_i}(s) \leq X_{t_{i+1}} - X_{t_i}$ holds with high probability for any $i = 0, \ldots, k-1$,
3. $t_k \approx t$ and $X_{t_k} = \widetilde{O}(t^p)$.

As it can easily be seen, finding suitable sequences $(t_i)_{i=0}^k$ and $(X_{t_i})_{i=0}^k$ satisfying the above conditions guarantees us that with high probability there exists $\deg_t(s) = \widetilde{O}(t^p)$. The starting point of the proof is the definition of the functional forms of the strings for the parameters p, α, β_i $(i = 0, 1, \ldots, k)$ and ϕ:

$$\begin{aligned} t_0 &= \phi, & t_{i+1} &= t_i + \frac{\alpha\, t_i \log t_i}{X_{t_i}}, & t_{k-1} &< t \leq t_k, \\ X_{t_0} &= t_0, & X_{t_{i+1}} &= X_{t_i} + \beta_i \log t_i. \end{aligned}$$

It turns out that for a sequence defined in this way there is a simple lower bound close to the one we sought:

Lemma 4.4. *Assume that $\phi \geq \log^2 t$, $\alpha \leq \sqrt{\phi}$ and $\beta_i \geq \alpha(p-\delta)$ for some $\delta \in [0, p)$. Asymptotically as $t \to \infty$ for any $i = 0, 1, \ldots, k$ we have $X_{t_i} \geq t_i^{p-\delta}$.*

The above formula can be proved inductively by using the inequality from Taylor series expansion and the definition of the parameter α:

$$\begin{aligned} t_{i+1}^{p-\delta} - t_i^{p-\delta} &= t_i^{p-\delta}\left(\left(1 + \frac{t_{i+1} - t_i}{t_i}\right)^{p-\delta} - 1\right) \leq t_i^{p-\delta}\frac{(p-\delta)(t_{i+1} - t_i)}{t_i} \\ &\leq X_{t_i}\frac{(p-\delta)\,(t_{i+1} - t_i)}{t_i} = \alpha(p-\delta)\log t_i \leq \beta_i \log t_i = X_{t_{i+1}} - X_{t_i}. \end{aligned}$$

This lower bound from Lemma 4.4 is then used in the proof of the stronger upper bound:

Lemma 4.5. *Assume that $\phi \geq \log^3 t$, $\alpha(p-\delta) \leq \beta_i \leq \alpha p + \frac{\alpha}{2\log t_i}$ for some $\delta \in [0, p)$. It holds asymptotically as $t \to \infty$ that $X_{t_i} \leq \phi^{1-p} t_i^p \log t_i$ for all $i = 0, 1, \ldots, k$.*

In particular, Lemma 4.5 guarantees that the sequences $(t_i)_{i=0}^k$ and $(X_{t_i})_{i=0}^k$ behave asymptotically according to our requirements i.e. in particular $X_{t_i} = O(t_i^p \mathrm{polylog}(t))$, as long as we ensure that $\phi = O(\mathrm{polylog}(t))$.

The second condition, i.e., the existence of the inequality $\deg_{t_{i+1}}(s) - \deg_{t_i}(s) \leq X_{t_{i+1}} - X_{t_i}$ with high probability, is achieved from a Chernoff bound for vertex degree over time respective to the value of X_{t_i}:

Lemma 4.6. *Let $1 \leq s \leq \tau \leq t$. Let $X_\tau \geq 0$, $\varepsilon \in (0,1)$ be the values such that for any $A > 0$, there exists $\deg_\tau(s) \leq X_\tau$ oraz $3A\tau \log t \leq \varepsilon^3 X_\tau(pX_\tau + r)$. Then, for any $h \in \left[\frac{3A\tau \log t}{\varepsilon^2(pX_\tau+r)}, \varepsilon X_\tau\right]$ it holds that*

$$\Pr\left(\deg_{\tau+h}(s) > \deg_\tau(s) + (1+3\varepsilon)\,\tfrac{h(pX_\tau+r)}{\tau}\right) = O(t^{-A}).$$

By choosing the proper values of $\alpha = \Theta(\log^2 t)$, $\beta_i = \alpha p + \frac{\alpha}{2\log t_i}$ and $\phi = \Theta(\log^4 t)$ it can be shown that the assumptions of the last two lemmas are satisfied. The first condition, about the incidence of $\deg_{t_0}(s) \leq X_{t_0}$, is clearly satisfied from the assumption $X_{t_0} = t_0$, and thus for any vertex $1 \leq s \leq t_0 = \phi$ we get that with high probability their degrees at any time t_i do not exceed X_{t_i}.

To prove the part (c) i.e. the upper bound of Theorem 4.8 for vertices $s \in [t_i, t_{i+1}]$ for successive $i = 0, 1, \ldots, k-1$ we use the observation that when $p < 1$, then with high probability the degree of none of the previous vertices exceeds $X_{t_{i+1}}$, and therefore $\deg_s(s)$ cannot exceed $(1+\varepsilon)(pX_{t_{i+1}})$ for any $\varepsilon > 0$. Moreover, it can be proved that the degree of the vertex $\deg_{t_{i+1}}(s)$ does not exceed $X_{t_{i+1}}$ with high probability – because $\deg_s(s)$ is sufficiently small and the degree increment between the moment s and t_{i+1} cannot make up for the difference. Eventually, it turns out that all vertices in a given interval satisfy the relation $\deg_s(s) \leq X_{t_{i+1}}$ with high probability, and so one can apply Lemma 4.6 to them to obtain an upper bound on their degree at time t.

In the proof of the lower bound (i.e. the (a) part) for Theorem 4.8, we also use the sequences $(t_i)_{i=0}^k$ and $(X_{t_i})_{i=0}^k$ as defined above. Again, an appropriate Chernoff-type bound can be proved:

Lemma 4.7. *Let $1 \leq s \leq \tau \leq t$. Let $X_\tau \geq 0$, $\varepsilon \in \left(0, \frac{1}{3}\right)$ be values such that for any $A > 0$ there is $X_\tau(s) \leq \tau$ and $3A \log t\varepsilon^3 pX_\tau$. Then for any $h \in \left[\frac{3A\log t}{\varepsilon^2 pX_\tau}, \varepsilon\right]$ it holds that*

$$\Pr\left(\deg_{\tau+h}(s) \leq \deg_\tau(s) + (1-2\varepsilon)\,\frac{hpX_\tau}{\tau}\right) = O(t^{-A}).$$

By taking the respective values of α, β_i and ϕ, it can be proved that the constraint $\deg_{t_i}(s) \geq X_{t_i} - \phi + 1$ occurs with high probability for all successive $i = 0, 1, \ldots, k$.

Open problems The investigations in papers [A1]-[A3] made it possible to show that both the degree distributions of individual vertices and the maximum degree in the graph are concentrated around their mean values. A question can be raised whether the existing estimates are the best ones, i.e. e.g. whether the functions $f_l(t)$ and $f_h(t)$ are the best possible. For example, one can try to prove results analogous to those for the Barabási-Albert model, where it was shown in [46] that with high probability there is a $\Delta(G_t) \in [\sqrt{t}/f(t), \sqrt{t}f(t)]$ for some slowly increasing function $f(t)$ – but at the same time that this cannot be true for any constant $f(t) = c$.

Certainly, an important further step in the structural analysis of this random graph model, especially from the point of view of compression of such graphs, could be to identify the asymptotic full degree distribution of the graph, i.e. the distribution of the variables $F_k(G_t)$ for all $k = 0, 1, \ldots$.

Another problem that might be of interest to the community, as shown for other models e.g. in [16, 99], may be the search for the asymptotic behaviour of such graph parameters as the clique number, the size of the largest independent set, the size of the largest matching in the graph, or the chromatic number of the graph.

4.4. Research on the scale parameter for duplicate random graph models

The belief that the scale-free property is present for graphs generated from duplicate models has been one of the important motivations for research on these models. In particular, it has been argued, e.g. in [30], that various biological and social networks are characterised by a scale factor within ranges $(1,2)$ and $(2,3)$, respectively, which would be expected to be consistent with the estimated scale factor for graphs generated by the duplicate model $\texttt{DD}(t,p,0)$ – unlike, e.g. from the Erdős-Renyí model (for which the scale-free property does not hold) or the Barabási-Albert model (for which it holds that $\gamma = 3$, see [17]).

A more detailed study in [57] showed that in the graphs generated from the model $\mathtt{DD}(t, p, 0)$ the following occurs.

$$\lim_{t\to\infty} f_0(G_t) = c(p) \in (0, 1),$$
$$\lim_{t\to\infty} f_k(G_t) = 0 \text{ for } k = 1, 2, \ldots,$$

for some convoluted constant $c(p) \in (0, 1]$ (in particular $c(p) = 1$ for $p < 0.5671\ldots$). This means that, asymptotically, almost all vertices in the graphs generated by a given model are isolated, and it cannot be the case $\lim_{k\to\infty} \lim_{t\to\infty} f_k(G_t) \sim k^\gamma$.

Such a predominant presence of isolated vertices in these graphs, rarely found in real-world networks, suggested a modification of the problem to the study of the (unique) connected component of such a graph, i.e. to omit isolated vertices:

$$a_k(G_t) = \frac{f_k(G_t)}{1 - f_0(G_t)} \text{ for } k = 1, 2, \ldots.$$

A formal consideration of this problem was undertaken in [68]. In particular, an equation for the generating function $A(z) = \sum_{k=0}^{\infty} a_k z^k$ was derived:

$$A(pz + 1 - p) = 2pA(z) + pz(1 - z)A'(z) + A(1 - p), \tag{6}$$

and then the following theorem was proved

Theorem 4.9 ([68, Theorem 2.1(3)])**.** *For $0 < p < \exp(-1)$ let $\beta(p) > 2$ be the solution of the equation $p^{\beta-2} + \beta - 3 = 0$. Then, for the asymptotic degree distribution of the connected component $(a_k)_{k=0}^{\infty}$ in the model $\mathtt{DD}(t, p, 0)$ it holds for $k \to \infty$ that*

$$\lim_{k\to\infty} \frac{a_k}{k^q} = 0 \quad \text{for } q < \beta(p),$$
$$\lim_{k\to\infty} \frac{a_k}{k^q} = \infty \quad \text{for } q > \beta(p).$$

The work [A4] is devoted to improving a result from Theorem 4.9, in particular showing the behaviour of the distribution for the missing case when $q = \beta(p)$:

> **Problem [A4]:** is the $\beta(p)$ factor a scale factor in the graphs generated as connected components of the $\mathtt{DD}(t, p, 0)$ model? If so, to what real number does the quotient $\frac{a_k}{k^{\beta(p)}}$ converge when $k \to \infty$?

Theorem 4.10. *For $0 < p < \exp(-1)$ let $\beta(p) > 2$ be the solution of the equation $p^{\beta-2} + \beta - 3 = 0$. Then, for the asymptotic degree distribution of the connected component $(a_k)_{k=0}^{\infty}$ in the model $\mathtt{DD}(t, p, 0)$ it holds for $k \to \infty$ that*

$$\frac{a_k}{k^{\beta(p)}} = \frac{1}{E(1) - E(\infty)} \cdot \frac{p^{-\frac{1}{2}(\beta(p)-\frac{3}{2})^2}\Gamma(\beta(p) - 2)}{D(\beta(p) - 2)(p^{-\beta(p)+2} + \ln(p))\Gamma(-\beta(p) + 1)} \left(1 + O\left(\frac{1}{k}\right)\right)$$

for the Euler gamma function $\Gamma(s)$ and

$$D(s) = \prod_{i=0}^{\infty} \left(1 + p^{1+i-s}(s - i - 2)\right),$$
$$E(1) - E(\infty) = \frac{1}{2\pi i} \int_{\Re(s)=c} p^{-\frac{1}{2}(s-\frac{1}{2})^2} \frac{\Gamma(s)}{D(s)} \mathrm{d}s \text{ for any } c \in (0, 1).$$

The proof of Theorem 4.10 consisted of a transformation by appropriate substitutions in equation 6 to obtain

$$C\left(\frac{w}{p}\right) = 2pC(w) + p(w - 1)C'(w) + A(1 - p)$$

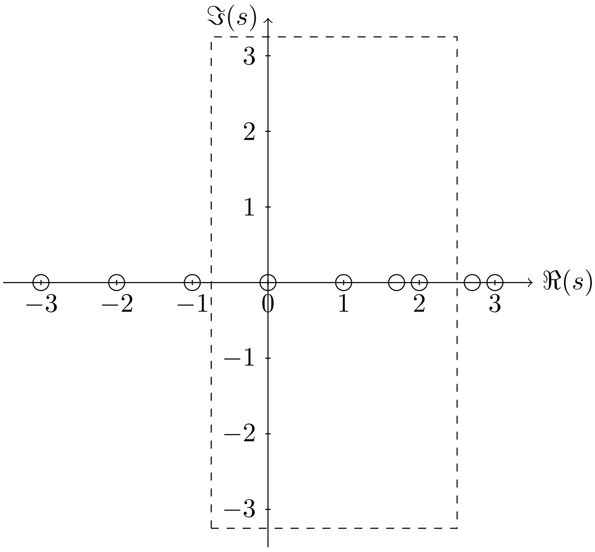

Figure 1: Example integration area for $E^*(s)$ and $E(w)$ with $s^* = 0.7$ i $M = 2.5$.

with boundary conditions $C(1) = A(0) = 0$ and $\lim_{w\to\infty} C(w) = A(1) = 1$.

Since the behaviour of $C(w)$ for $w \to 0$ is unknown, it is not possible to directly determine the *fundamental strip* of the function C and apply the Mellin transformation, a well-known analytical tool for calculating asymptotic expansions [45, 95]. Instead, the reverse procedure was used: first, we defined a function

$$E^*(s) = p^{-\frac{1}{2}(s-\frac{1}{2})^2} \frac{\Gamma(s)}{D(s)}$$

such that for a function satisfying the equation $\mathbb{E}(\frac{w}{p}) = 2pE(w) + p(w-1)E'(w) + K$ for some constant K the following lemma holds:

Lemma 4.8. *The function $\mathbb{E}^*(s)$ is the Mellin transform of the function $E(w)$ with the fundamental strip $\{s \colon \Re(s) \in (-1, 0)\}$.*

To compute the asymptotic expansion of the function $E(w)$ in series with respect to the variable w, the standard approach (see [95]) was used, according to which the integral along a certain vertical line in the fundamental chord $\Re(s) \in (-1, 0)$ is converted into the integral along the corresponding rectangle (see Figure 1) using the fact that, given certain properties of the function E, the upper and lower sides of the rectangle contribute negligible small magnitudes (which follows from the so-called *smallness property* of the Mellin transform) and the right side of the rectangle provides only some asymptotic factor proportional to w^{-M}.

According to Cauchy residue theorem, one can convert this integral into the sum of the values of the corresponding residues. Upon closer examination of the function $E^*(s)$ for $0 < p < \exp(-1)$ one can see that it has only 3 types of simple isolated poles:

- for $s = 0, -1, -2, \ldots$, introduced by the function $\Gamma(s)$,
- for $s = 1, 2, 3, \ldots$, introduced by the function $\frac{1}{D(s)}$,
- for $s = s^* + 1, s^* + 2, s^* + 3, \ldots$, introduced by the function $\frac{1}{D(s)}$,

where s^* is the non-trivial (i.e. different from zero) solution of the equation $p^s + s - 1 = 0$. Calculating the residues for the poles inside the rectangle therefore allows this to compute the expansion of $E(w)$ in series with respect to w with accuracy $O(w^{-M})$ for any $M > 0$. The transformation of the asymptotic expansion of $E(w)$ into $C(w)$ remains a matter of simple linear scaling.

Finally, it suffices only to note that the series expansion of $C(w)$ for $w^{-\alpha}$ is equivalent to the series expansion of $A(z)$ for $(1-z)^{\alpha}$ – and the growth rate of the latter is determined by the word with the smallest $\alpha \in \mathbb{R}_{+} \setminus \mathbb{N}$, since we know from the classical theorems of Flajolet and Odlyzko [44] on the transfer of growth rates that

$$\begin{aligned}
[z^k](1-z)^{\alpha} &= \frac{k^{-\alpha-1}}{\Gamma(-\alpha)}\left(1+O\left(\frac{1}{k}\right)\right), \\
[z^k]o(1-z)^{\alpha} &= o(k^{-\alpha-1}).
\end{aligned}$$

Therefore, since s^*+1 is the smallest non-integer pole of the function $E^*(s)$, then $a_k = [z^k]A(z)$ will be asymptotically proportional to k^{-s^*-2} with an appropriate, explicitly computable constant.

Since for $p \in (0, \exp(-1))$ it is known that $s^* \in (0,1)$, it can also be said that the main result of the work of [A4] is to show that for certain values of the parameter p the model $\mathtt{DD}(t,p,0)$ generates scale-free graphs with a coefficient in the range $(2,3)$, thus matching estimates for real networks of certain types (see [30]).

Open problems A natural extenstion of this line of research would be to find whether the graphs generated according to this model exhibit the scale-free property also for the case of $p \geq \exp(-1)$. The possibility of the existence of a phase transition has already been suggested in [68], where it was shown that a certain infinite Markov chain related to the original graph generation process does not have a stationary distribution (i.e. its Markov chain is *transient*), which may suggest that the same occurs also for the Markov chain describing $\mathtt{DD}(t,p,0)$ model itself.

A similar line of research would be to seek an answer to the same question for the more general model $\mathtt{DD}(t,p,r)$ and to resolve whether the introduction of additional edges preserves the scale factor or at least the scale-free property for $0 < p < \exp(-1)$.

4.5. Compression for duplicate random graph models

The paper [A5] addresses the issue of compression for graphs generated from the $DD(t,1,0)$ model, i.e. according to the so-called full duplication model (*full duplication*) [11, 30].

The starting point of the work on the above problem was the observation that, in this special case, each newly added vertex will be an exact copy of one of the existing vertices – and therefore also a copy of one of the vertices of the starting graph. If we assume that we start with a graph G_{t_0} on t_0 vertices we can therefore represent the graph in terms of $t-t_0$ numbers of length $\log_2 t_0$ describing the labels of the corresponding vertices of the initial graph. Also, the structure of the graph can be described as a sequence of t_0 numbers of length $\log_2 t$ describing the number of vertices that are (also indirect) copies of consecutive vertices from the initial graph. In this way, we also obtain simple entropy and structural entropy estimates for this model:

$$\begin{aligned}
H(G_t) &\leq (t-t_0)\log_2 t_0, \\
H(S_t) &\leq t_0 \log_2 t.
\end{aligned}$$

Naturally, one can ask whether these estimates can be improved, and therefore what are the appropriate lower bounds on compression of such graphs:

Problem [A5]: what are the exact asymptotic growth rates of entropy ($H(G_t)$) and structural entropy ($H(S_t)$) for graphs generated from the model $\mathtt{DD}(t,1,0)$?

The main result of the paper is to determine with accuracy $o(1)$ the asymptotic growth rate of both entropies:

Theorem 4.11. *Asymptotically as $t \to \infty$, it holds that*

$$\begin{aligned}
H(S_t) &= (t_0-1)\log_2 t - \log_2 (t_0-1)! + o(1), \\
H(G_t) &= t(H_{t_0}-1)\log_2 e + \frac{n_0-1}{2}\log_2 n - \log_2(n_0-1)! \\
&\quad + \left(\frac{1-t_0}{2} + \frac{3t_0-2}{2}H_{t_0-1}\right)\log_2 e + \frac{t_0}{2}\log_2(2\pi) + o(1),
\end{aligned}$$

where $H_n = \sum_{k=1}^{n} \frac{1}{k}$ is the n-th harmonic number.

In addition to this, two algorithms (for graph compression and graph structure) are presented to ensure that the average codeword length does not exceed the corresponding entropy by more than 2 bits.

The starting point is the use of the general relation between $H(G_t)$ and $H(S_t)$, previously proved only for the Barabási-Albert model in [80]:

Lemma 4.9 (Generalization of Lemma 1 in [80])**.** *For any random graph model in which every two graphs with the same structure and non-zero probability of being generated from the model have exactly the same probability, it holds that*

$$H(G_t) - H(S_t) = \mathbb{E}[\log_2 |\Gamma(G_t)|] - \mathbb{E}[\log_2 |\mathrm{Aut}(G_t)|],$$

where $\Gamma(G_t)$ is the set of possible permutations of the labels of graph G_t for which the resulting graph has non-zero generation probability, while $\mathrm{Aut}(G_t)$ is the set of automorphisms of the graph.

As mentioned earlier, the structure of the graph G_t can be written in terms of a vector $(C_{i,t})_{i=1}^{t_0}$, where the random variable $C_{i,t}$ denotes the number of vertices that are copies (including copies, copies of copies, etc.) of the i-th vertex from the graph G_{t_0}. It turns out that the probability of structures is described by the so-called multinomial Dirichlet distribution. This allows, using the fact that the distributions of all $C_{i,t}$ are identical and that the Euler beta function $B(t, t_0)$ (appearing in the definition of this distribution) has a well-known asymptotic expansion, to obtain the value of $H(S_t)$.

The computation of $|\Gamma(G_t)|$ consists in observing that all permutations of the labels preserving the contents of each class, determined by the initial vertices, are admissible. This leads to the observation that

$$|\Gamma(G_t)| = t! \prod_{i=1}^{t_0} C_{i,t} = t! \; t_0 \; C_t,$$

where C_t is a variable having a beta-binomial distribution (as a marginal probability distribution for the Dirichlet multinomial distribution) shifted by 1, i.e. defined by the formula

$$\Pr[C_t = k + 1] = (t_0 - 1)\binom{t}{k} B(k + 1, t + t_0 - k - 1).$$

Determining the asymptotic value of $\mathbb{E}[\log_2 |\Gamma(G_t)|]$ thus requires only a translation in terms of C_t, followed by the reuse of asymptotic function expansions. Interestingly, establishing the exact probability distribution of the variable C_t improves at the same time on the approximation used in the [30], which suggested (without any formal proof) that it could be replaced by a continuous probability density function of the form $f(x) = \exp\left(-\frac{x}{\mathbb{E}[C_t]}\right)$.

The final part is to calculate the growth rate of the function $\mathbb{E}[\log_2 |\mathrm{Aut}(G_t)|]$. The key step is to note that, assuming that the initial graph was asymmetric, it is true that

$$\mathbb{E}[\log_2 |\mathrm{Aut}(G_t)|] = t_0 \mathbb{E}[\log_2 C_t!],$$

since the automorphism can be any function mapping vertices within particular classes, determined by the initial vertices. Once such a result is obtained, it suffices to estimate with sufficient accuracy the subsequent expressions of the series from the Stirling expansion, i.e. $\mathbb{E}[(X + 1)\log_2(X + 1)]$, $\mathbb{E}[X]$ and $\mathbb{E}[\log_2(X + 1)]$ for the variable X having any beta-dummy distribution. This estimation was proved using the properties of analytic functions such as the Euler beta function or the Euler digamma function, as well as using probabilistic tools such as Chernoff bound.

The compression algorithms developed for both types of input (graphs and their structures) were based on the idea of arithmetic coding. From the above considerations, the exact formulae for the joint, boundary and conditional probabilities for individual vectors $(C_{i,t})_{i=1}^{t_0}$ are known. One can therefore propose to map the graph structure stored in the form of such a vector to some number in the range $[0, 1)$, whose appropriately clipped binary expansion constitutes a codeword with good compression guarantees. A similar approach can be used analogously for the description of a graph (with labels) mentioned at the beginning as a sequence of numbers describing its model vertices from the initial graph. It follows directly from the properties of the arithmetic coding itself that the average length of the codewords does not exceed the corresponding entropies augmented by two bits of [34].

It is worth noting that in this model there is a significant difference between the asymptotic growth rates of the two entropies – unlike, e.g., for the Barabási-Albert model, for which we have $H(G_t) = H(S_t) = \Theta(t \log t)$ [80].

Open problems The knowledge of the entropy formula and the optimal (within a fixed number of bits) algorithm for the special case of $\mathtt{DD}(t, 1, 0)$ directly suggests a further question whether there exists a similar construction for the more general model $\mathtt{DD}(t, p, r)$. Note, however, that in the general case not only is there no simple interchangeability of vertices, allowing them to be counted in the variables $C_{i,t}$, but even the conditions of Lemma 4.9 are not satisfied, since different graphs with the same structure may have different generation probabilities.

At the same time, as shown in [27, 80], the results in the papers [A1]-[A3] may provide a promising starting point for work on the entropy of the probability distribution of the graphs generated by this model, and thus also for the construction of good compression algorithms. In particular, constraints on the maximum degree and degree distribution of a graph may allow the identification of a relatively small set of probabilistic graphs, and their correspondingly parsimonious representation.

5. Other scientific results

[B1] Abram Magner, Krzysztof Turowski, Wojciech Szpankowski, Lossless Compression of Binary Trees with Correlated Vertex Names, *IEEE Transactions on Information Theory* 64(9) (2018), pages 6070-6080.

- conference version: Abram Magner, Krzysztof Turowski, Wojciech Szpankowski, *Lossless compression of binary trees with correlated vertex names*, IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016. Lecture Notes in Computer Science 13025, pages 1217-1221.

[B2] Jacek Cichoń, Abram Magner, Wojciech Szpankowski, Krzysztof Turowski, *On Symmetries of Non-Plane Trees in a Non-Uniform Model*, Proceedings of the Fourteenth Workshop on Analytic Algorithmics and Combinatorics, ANALCO 2017, Barcelona, Spain, Hotel Porta Fira, January 16-17, 2017, pages 156-163.

[C1] Jithin Sreedharan, Krzysztof Turowski, Wojciech Szpankowski, Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18(3) (2021), pages 836-849.

[C2] Jithin Sreedharan, Krzysztof Turowski, Wojciech Szpankowski, Temporal Ordered Clustering in Dynamic Networks: Unsupervised and Semi-supervised Learning Algorithms, *IEEE Transactions on Network Science and Engineering* 8(2) (2021), pages 1426-1442.

- conference version: Jithin Sreedharan, Krzysztof Turowski, Wojciech Szpankowski, *Temporal Ordered Clustering in Dynamic Networks*, IEEE International Symposium on Information Theory, ISIT 2020, Los Angeles, CA, USA, June 21-26, 2020, pages 1349-1354.

[D1] Krzysztof Turowski, Philippe Jacquet, Wojciech Szpankowski, *Asymptotics of Entropy of the Dirichlet-Multinomial Distribution*, IEEE International Symposium on Information Theory, ISIT 2019, Paris, France, July 7-12, 2019, pages 1517-1521.

[E1] Tytus Pikies, Krzysztof Turowski, Marek Kubale, Scheduling with Complete Multipartite Incompatibility Graph on Parallel Machines, *Artificial Intelligence* 309 (2022), pages 103711.

- conference version: Tytus Pikies, Krzysztof Turowski, Marek Kubale, *Scheduling with Complete Multipartite Incompatibility Graph on Parallel Machines*, Proceedings of the Thirty-First International Conference on Automated Planning and Scheduling, ICAPS 2021, Guangzhou, China (virtual), August 2-13, 2021, pages 262-270 [Best Runner-Up Paper Award].

[F1] Krzysztof Turowski, Optimal backbone colouring of split graphs with matching backbones, *Discussiones Mathematicae Graph Theory* 35(1) (2015), pages 157-169.

[F2] Robert Janczewski, Krzysztof Turowski, The backbone colouring problem for bipartite backbones, *Graphs and Combinatorics* 31(5) (2015), pages 1487-1496.

[F3] Robert Janczewski, Krzysztof Turowski, The computational complexity of the backbone colouring problem for planar graphs with connected backbones, *Discrete Applied Mathematics* 184 (2015), pages 237-242.

[F4] Robert Janczewski, Krzysztof Turowski, The computational complexity of the backbone colouring problem for bounded-degree graphs with connected backbones, *Information Processing Letters* 115(2) (2015), pages 232-236.

[F5] Krzysztof Michalik, Krzysztof Turowski, On λ-backbone colouring of cliques with tree backbones in linear time, arXiv:2107.05772, 2021.

[F6] Robert Janczewski, Krzysztof Turowski, An $O(n \log n)$ algorithm for finding edge span of cacti, *Journal of Combinatorial Optimization* 31(4) (2016), pages 1373-1382.

[F7] Robert Janczewski, Krzysztof Turowski, On the hardness of computing span of subcubic graphs, *Information Processing Letters* 116(1) (2016), pages 26-32.

[F8] Robert Janczewski, Anna Maria Trzaskowska, Krzysztof Turowski, T-colourings, divisibility and the circular chromatic number, *Discussiones Mathematicae Graph Theory*, 41(2) (2021), pages 441-450.

[G1] Robert Janczewski, Paweł Obszarski, Krzysztof Turowski, 2-colouring number revisited, *Theoretical Computer Science* 796 (2019), pages 187-195.

[G2] Robert Janczewski, Paweł Obszarski, Krzysztof Turowski, Bartłomiej Wróblewski, Infinite chromatic games, *Discrete Applied Mathematics* 309 (2022), pages 138-146.

[H1] Robert Janczewski, Paweł Obszarski, Krzysztof Turowski, Weighted 2-sections and hypergraph reconstruction, *Theoretical Computer Science* 915 (2022), pages 11-25.

The papers [F1], [F2], [F3] and [F4] were published before obtaining PhD.

5.1. Compression of random trees

The problems of compression of random trees occupy a special place within the area between information theory and graph theory, due to the use of trees in, among other things, mathematical phylogenetics.

The paper [B1] is devoted to the compression of rooted random trees with correlated vertices labels. The model under study has four parameters: a target number of vertices n, a label length m, an alphabet $\mathcal{A}$ and a matrix of letter substitution probabilities P (with a corresponding stationary distribution π). The generation of the tree proceeds as follows: starting with a tree T consisting of a single vertex v (the root of the tree) with a label $l(v)$ generated at random according to the distribution of π[5] we repeat the operations sequentially:

1. select a leaf u in the tree T, add two children w_1 and w_2 to it,
2. for each w_i $(i = 1, 2)$ create a label $l(w_i)$ by independently drawing letters according to the transition probabilities from the matrix P and the corresponding letters from the label $l(u)$.

[5] Since the effect of the randomness of the root label on the final result is negligible, we can also assume that the root label takes some fixed value.

Such a tree generation model (without labels) is known in the literature as the Yule model [38].

The object of study here is a random variable LT_n representing a generated labeled tree on n vertcies, equivalently represented as a pair (T_n, F_n), where T_n corresponds to the tree structure without labels, while F_n is a sequence of labels, e.g. given in preorder. The aim of the study is to estimate the entropy of the variable $H(LT_n)$.

From the chain rule for entropy, we know that $H(LT_n) = H(T_n) + H(F_n|T_n)$. The calculation of the second part is straightfoward:

$$H(F_n|T_n) = 2mh(P)(n-1) + mh(\pi),$$

where $h(P) = -\sum_{a\in\mathcal{A}} \pi(a) \sum_{b\in\mathcal{A}} P(b|a) \log P(b|a)$ is the entropy of a Markov process with transition matrix P, while $h(\pi) = -\sum\limits_{a\in\mathcal{A}} \pi(a) \log \pi(a)$ is the entropy of the stationary distribution π for the matrix P.

To compute $H(T_n)$ we prove that our tree formation model is equivalent to the model in which we insert a random permutation of the numbers $\{1, \ldots, n\}$ into the binary tree. This allows us to determine an exact formula for the probability distribution of drawing trees depending on the degrees of their internal vertices. Using the results presented in [71] we show that:

Theorem 5.1. *The entropy of binary trees with n vertices generated according to the Yule model with labels of length m is*

$$\begin{aligned} H(LT_n) &= \log_2(n-1) + 2n \sum_{k=2}^{n-1} \frac{\log_2(k-1)}{k(k+1)} + 2mh(P)(n-1) + mh(\pi) \\ &= n(1.736\ldots + 2mh(P)) + O(\log n). \end{aligned}$$

The model above assumes that the tree is given with orientation on the plane (*plane tree*), i.e. the order of the children for each vertex matters for distinguishing trees. However, one can consider non-oriented trees (*non-plane tree*) and their respective entropies with labels ($H(LS_n)$) or without labels ($H(S_n)$).

The work of [B2] is an extension of the research on the above-introduced random non-plane tree model. In particular, for the function $Z(T)$ (denoted in the paper as $sym(T)$), defined as the number of internal vertices having isomorphic subtrees, a function is derived to form the

$$F(u, z) = \sum_{n=1}^{\infty} \sum_{T\in\mathcal{T}_n} \Pr[T_n = T] u^{Z(T)} z^{|V(T)|},$$

for which the differential equation is derived

$$\frac{\partial(F(u,z)/z)}{\partial z} = \frac{F^2(u,z)}{z^2} + (u-1)B(u^2, z^2)$$

for

$$B(u, z) = \sum_{n=1}^{\infty} \sum_{T\in\mathcal{T}_n} \Pr^2[T_n = T] u^{Z(T)} z^{|V(T)|-1}.$$

It is further pointed out that each non-plane tree S corresponds to $2^{n-1-Z(S)}$ plane trees, where $Z(S)$ denotes the number of vertices having two identical (i.e. isomorphic) subtrees. For a given tree s and a tree $s \circ s$ consisting of a root with exactly two subtrees s, the following recursive equation can be formulated

$$Z(S_n, s) = [T_n \sim s \circ s] + Z(S_{U_{n-1}}, s) + Z(S_{n-U_{n-1}}, s),$$

where U_k denotes a uniform random variable over $\{1, \ldots, k\}$. From this it follows directly from the equation

$$\mathbb{E}[Z(S_n, s)] = \mathbb{E}[T_n \sim s \circ s] + \frac{2}{n-1} \sum_{k=1}^{n-1} \mathbb{E}[Z(S_k, s)].$$

By solving this formula we obtain that

$$\mathbb{E}[Z(S_n)] = \sum_s \mathbb{E}[Z(S_n, s)] = n \sum_{k=1}^{\lfloor (n+1)/2 \rfloor} \frac{\sum_{s \in \mathcal{T}_k} \Pr^2[T_k = s]}{(2k-1)k(2k+1)} + O(1),$$

and from this it can be calculated that

$$\begin{aligned} H(T_n|S_n) &= n - 1 - \mathbb{E}[Z(S_n)] \approx 0.6275n, \\ H(S_n) &\approx 1.109n. \end{aligned}$$

In the paper [B1] there are also presented three algorithms for random tree compression based on the idea of arithmetic coding. The algorithm for plane trees separately compresses the tree and separately the labels based on the known child-parent relationship in the tree. It runs in $O(n^2 \log^2 n \log\log n)$ time and, for trees generated from the model described above, achieves an average codeword length of no more than $H(LT_n) + 2$ bits.

For non-plane trees, two algorithms are presented: an approximate fast CompressNPTreeand a slower asymptotically optimal one. The first one, running in linear time, consists of canonically ordering the binary tree according to the "smaller subtree on the left" rule. It can be proven that such heuristics produce, on average, a markup over optimal value close to 1%:

Theorem 5.2. *Asymptotically as $n \to \infty$ the average codeword length returned by* CompressNPTree *algorithm does not exceed* $1.013H(S_n)$.

The slower algorithm is based on determining a recursively defined order on the trees, allowing to distinguish cases of non-isomorphic trees with the same number of leaves. This, together with an algorithm that computes for a given order and any plane tree S the probabilities $\Pr[S_n < S]$ and $\Pr[S_n \leq S]$ are sufficient to apply an arithmetic coding scheme for obtaining a codeword with an average length not exceeding $H(S_n) + 2$ bits. The pessimistic running time of this algorithm is bounded by $O(n^3 \log^2 n \log\log n)$, while the average is bounded by $O(n^2 \log^2 n \log\log n)$ – because the accuracy of the computed probabilities requires that multiplication of numbers of length $O(n^2)$ bits be taken into account.

5.2. Inference in a duplicate model of random graphs

While considering the practical relevance of theoretical models of random graphs one can often encounter a question of the efficiency of fitting models to existing graphs. The models under study usually have several free parameters: for example, the aforementioned model $\mathtt{DD}(t, p, r)$ by Solé and Pastor-Satorras has, in addition to a fixed number of vertices t, two such parameters p and r. Intuitively, we would want a given real network to be generated with the highest possible probability according to a given model – and so one should look for a set of parameters that maximise such probability. However, since it turns out that many biological or social networks are characterised by the existence of multiple automorphisms, and for many models (including, but not limited to the Solé and Pastor-Satorras model) different isomorphic graphs correspond to different probabilities of being generated, this problem becomes computationally difficult in practice.

Previous works have circumvented this problem by relying on the assumption that, when fitting networks to models generating scale-free graphs, it is possible to reconstruct the parameter values from the scale factor values for the stationary state of the model [59, 92]. For example, for the Solé and Pastor-Satorras model $\mathtt{DD}(t, p, r)$ the formulae given in the papers [59, 92] were used to calculate p and r:

$$\begin{cases} \gamma &= 1 + \frac{1}{p} - p^{\gamma-2} \\ r &= \left(\frac{1}{2} - p\right) D(G) \text{ dla } p < \frac{1}{2}, \end{cases}$$

where γ denotes the scale factor, while $D(G)$ is the average degree of a vertex in the graph.

In the paper [C1], this approach was challenged on the basis of theoretical arguments and simulations. First, there is still a debate in the literature regarding the occurrence of scale-freeness in nature. Some statistical arguments are pointed out, showing that many types of biological and social networks are not characterised by scale-free property at all [21, 70, 97]. One may also come across arguments criticising standard methods for estimating the scale factor γ (e.g. proposed in [4]) as detecting the scale-free feature in graphs generated from models without this property [32]. Regardless of the weight of these arguments, for the protein interaction networks we investigated, we found that standard methods of calculating γ values make a cutoff of only a few percent of the vertices with the highest degree, and the choice of the cutoff value itself significantly modifies the obtained results.

Second, it is also devoid of good justification to assume that the actual networks are large enough to justify its occurrence with sufficient approximation of the relationships characteristic of the average case in the stationary state. Third, it can be argued that the techniques used to prove these relationships have not been sufficiently rigorous. However, it should be noted that the above equality linking γ and p was proved in the work of [A4] for the model $\mathtt{DD}(n, p, r)$ with parameters $0 < p < \exp(-1)$ and $r = 0$, so Jordan in [68] showed, that the corresponding Markov chain characterising this model for $p > \exp(-1)$ and $r = 0$ is transient, which calls into question the applicability of the formulae in the general case.

Moreover, the empirical distributions of the values of the various properties of the graphs generated from the model with the parameters described by the above formulae differed significantly from the actual networks they were supposed to correspond to. In particular, with such parameters, graphs were generated that were characterised by a completely different order of magnitude of the number of automorphisms – one of the key features especially for real biological networks, e.g. protein interaction networks [11]. Finding a proper fit for this particular graph parameter is all the more important as it has been proven that the most popular random graph models (Erdős-Renyí and Barabási-Albert) generate symmetry-free graphs with high probability for a wide range of parameters.

In our paper [C1] we proposed a new way to improve parameter estimation for duplication-divergence graph and to incorporate considerations about number of automorphisms. First, we used the recurrence (2) for $D(G_n)$ from [A1] and we derived the similar formulae for other graph parameters, such as the number of triangles $C_3(G_n)$ and the number of open triangles $S_2(G_n)$:

$$\begin{aligned}
\mathbb{E}[D(G_{n+1})|G_n] =& D(G_n)\left(1 + \frac{2p-1}{n+1} - \frac{2r}{n(n+1)}\right) + \frac{2r}{n+1},\\
\mathbb{E}[D_2(G_{n+1})|G_n] =& D_2(G_n)\left(1 + \frac{2p+p^2-1}{n+1} - \frac{2r(1+p)}{n(n+1)} + \frac{r^2}{n^2(n+1)}\right)\\
&+ D(G_n)\left(\frac{2p-p^2+2pr+2r}{n+1} - \frac{2r+2r^2}{n(n+1)} + \frac{r^2}{n^2(n+1)}\right) + \frac{2r^2+2r}{n+1} - \frac{r^2}{n(n+1)},\\
\mathbb{E}[C_3(G_{n+1})|G_n] =& C_3(G_n)\left(1 + \frac{3p^2}{n} - \frac{6pr}{n^2} + \frac{3r^2}{n^3}\right) + D_2(G_n)\left(\frac{pr}{n} - \frac{r^2}{n^2}\right) + D(G_n)\frac{r^2}{2n},\\
\mathbb{E}[S_2(G_{n+1})|G_n] =& S_2(G_n)\left(1 + \frac{2p+p^2}{n} - \frac{2(p+1)r}{n^2} + \frac{r^2}{n^3}\right)\\
&+ D(G_n)\left(pr + p + r - \frac{pr+r+r^2}{n} + \frac{r^2}{n^2}\right) + \frac{r^2}{2} - \frac{r^2}{2n}.
\end{aligned}$$

These recurrences, combined with a binary search over set of all possible (p, r), helped to derive from the observed values of $D(G_n)$ and $D(G_{n_0})$ the sets of parameters that fit these values[6] and obtain an intersection of these sets as a set of "good" parameter estimates. It was then checked that for networks with known p and r this method returns parameters close to their true values, thus also matching their number of automorphisms. Then, we re-applied this approach for a collection of real-world biological networks, obtaining new estimates for the parameters p and r. It was then verified that models with these parameters generate graphs which are not only close to the real-world network in terms of $D(G_n)$ and other estimated parameters, but also have a very similar values of $\mathrm{Aut}(G_n)$, which corroborates the validity of this approach.

The procedure sketched above turned out also to be very practical, as it runs in linear time. In comparison, our benchmark MLE approach also led to similar estimations for small graphs, but

[6] In fact, we assumed that the observed values can deviate somewhat from the theoretical average ones, thus enlarging the set.

it required $\Theta(n^3)$ operations, which made it impractical for real-world networks with hundreds of thousands of vertices.

The justification of the correspondence of the network to the model and the estimation of the relevant parameters is the starting point for addressing further problems present in the literature. One of them is the so-called *network archaeology* problem, i.e. reconstructing the evolution of a network over time and retrieving the temporal order of vertices in real-world networks. This problem, previously considered in the context of other random graph models [84, 105], is an important research issue leading to answers to a number of practical questions, such as inferring the structural and functional evolution of the brain [94].

In accordance with the formal definition proposed in the analysis of the Barabási-Albert model [94], it was assumed that the objective was to find a partial order σ defined over the vertices of a given graph, which optimizes two criteria:

- *density*: the number of distinguishable pairs in the partial order σ, normalized by the maximum number of distinguishable pairs:
$$\delta(\sigma) = \frac{K(\sigma)}{\binom{n}{2}},$$
where $K(\sigma) = |\{(u, v) : u <_\sigma v\}|$,
- *precision*: the expected number of correctly identified pairs of vertices from the original temporal order π, expressed as a fraction of all distinguishable pairs in the partial order σ:
$$\theta(\sigma) = \mathbb{E}\left[\frac{|\{u, v \in \{1, \ldots, n\} : u <_\sigma v, \pi^{-1}(u) < \pi^{-1}(v)\}|}{K(\sigma)}\right].$$

In this approach, for each fixed value of $\varepsilon \in (0, 1]$ the order σ with density $\delta(\sigma) \geq \varepsilon$ maximising $\theta(\sigma)$ shows the theoretical limits of the reconstructability of the order i.e. the expected quality of the prediction when requiring that at least a percentage of ε vertex pairs are recognisable. These values also provide a target for temporal order reconstruction algorithms in networks.

The whole paper [C2] has been devoted precisely to the problem of network archaeology for a wide class of random graph models with a particular focus on duplicate models. Specifically, the assumption characterising the Barabási-Albert model of equal probability of generating two isomorphic graphs was abandoned, since duplication models in general do not have such property.

The first result was to propose two integer programs and their relaxations to linear programs for finding order in dynamical networks based only on their snapshot at some single fixed point in time. Both programs in their general form depended only on parameters describing the probability $p_{u,v}$ of the order of all pairs of vertices u and v based on the structure of the graph and the assumed model of its evolution in time (see Table 1). The second step was to propose a suitable sampling procedure (so-called *sequential importance sampling*) to estimate the value of $p_{u,v}$ for any random graph model that can be written in the form of a time-inhomogeneous Markov chain, e.g. describing an evolution involving sequential random addition of vertices and edges. It has been proved that this procedure defines an asymptotic strong consistency i.e. an almost sure convergence to the true values of $p_{u,v}$.

Next, the above approach was applied to the model $\mathtt{DD}(t, p, r)$. In particular, detailed formulae have been derived from the model definition to allow the sampling procedure to be used to compute the values of $p_{u,v}$ and to determine, according to linear programming, the optimal values of θ and δ (so called *theoretical bounds*) for example graphs generated from the model for example parameter values. Since the sampling procedure leaves some freedom in the choice of the weight functions, a number of experiments were also performed to illustrate the convergence to the true values of $p_{u,v}$ with the number of samples. Ultimately, we proposed two methods with different weights: proportional HIGH-PROB-SAMPLING and equal LOCAL-UNIF-SAMPLING, and surprisingly the latter one had visibly faster rate of convergence in practice (see Figure 2).

Since the $p_{u,v}$ values themselves already give some information about the order sought and the determination of $p_{u,v}$ in practice turns out to be much faster than the use of linear programming, two algorithms, named $p_{u,v}$-THRESHOLD and SORT-BY-$p_{u,v}$-SUM, based on the ordering of vertices into groups according to the $p_{u,v}$ values, were also proposed in this paper. To test the quality of the selected algorithms, simulations were performed on graphs generated from the Solé and

LP-clusters		LP-partial-order	
IP	**LP approximation**	**IP**	**LP approximation**
$\max_{z} \frac{\sum_{\substack{1\le u\neq v\le n\\ 1\le i<j\le n}} p_{u,v}\, z_{u,i,v,j}}{\sum_{\substack{1\le k<l\le n\\ 1\le w\neq w'\le n}} z_{w,k,w',l}}$	$\max_{z'} \sum_{\substack{1\le u\neq v\le n\\ 1\le i<j\le n}} p_{u,v}\, z'_{u,i,v,j}$	$\max_{y} \frac{\sum_{1\le u\neq v\le n} p_{u,v}\, y_{u,v}}{\sum_{1\le u\neq v\le n} y_{u,v}}$	$\max_{y'} \sum_{1\le u\neq v\le n} p_{u,v}\, y'_{u,v}$
subject to	subject to	subject to	subject to
$\forall_{u,i,v,j\in[n]}\ z_{u,i,v,j} \in \{0,1\},$	$\forall_{u,i,v,j\in[n]}\ z'_{u,i,v,j} \in [0, 1/\epsilon\binom{n}{2}],$	$\forall_{u,v\in[n]}\ y_{u,v} \in \{0,1\},$	$\forall_{u,v\in[n]}\ y'_{u,v} \in [0, 1/\epsilon\binom{n}{2}]$
$\sum_{\substack{1\le u\neq v\le n\\ 1\le i<j\le n}} z_{u,i,v,j} \ge \epsilon\binom{n}{2},$	$\sum_{\substack{1\le u\neq v\le n\\ 1\le i<j\le n}} z'_{u,i,v,j} = 1,$	$\sum_{1\le u\neq v\le n} y_{u,v} \ge \epsilon\binom{n}{2}$	$\sum_{1\le u\neq v\le n} y'_{u,v} = 1$
$\sum_{i\in[n]} z_{u,i,u,i} = 1,$	$\sum_{i\in[n]} z'_{u,i,u,i} \le 1/\epsilon\binom{n}{2},$	$y_{u,v} + y_{v,u} \le 1,$	$y'_{u,v} + y'_{v,u} \le 1/\epsilon\binom{n}{2}$
$z_{u,i,v,j} = z_{v,j,u,i},$ $\sum_{i\in[n]} z_{u,i,v,j} = z_{v,j,v,j}.$	$z'_{u,i,v,j} = z'_{v,j,u,i},$ $\sum_{i\in[n]} z'_{u,i,v,j} = z'_{v,j,v,j}.$	$y_{u,v} + y_{v,w} - y_{u,w} \le 1.$	$y'_{u,v} + y'_{v,w} - y'_{u,w} \le 1/\epsilon\binom{n}{2}.$

Table 1: Two proposed integer programs and their respective LP relaxations

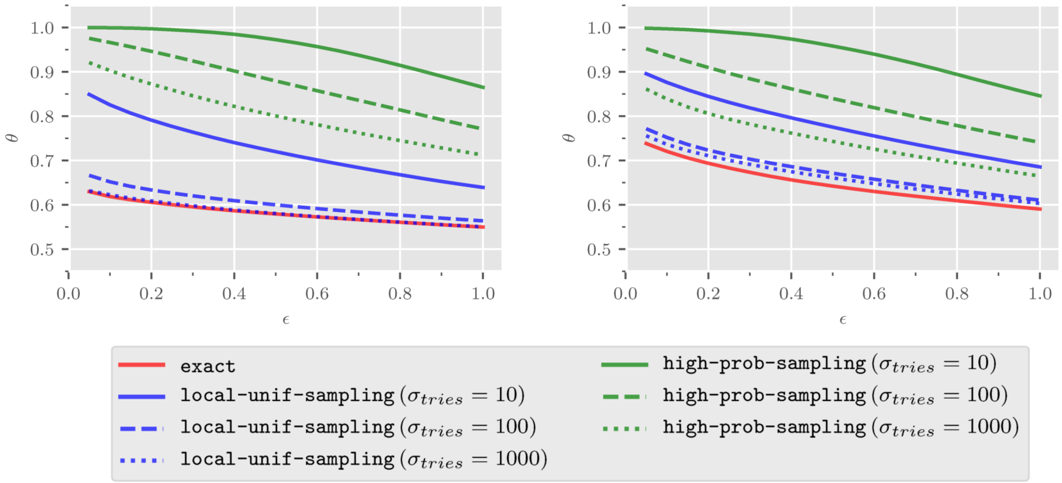

Figure 2: Results on the convergence to the exact precision (θ)-minimum density (ε) curve: $G_n \sim \mathtt{DD}(13, p, 1.0, G_{n_0})$ for $p = 0.3$ (left) and 0.6 (right), averaged over 100 graphs. G_{n_0} is generated from Erdős-Renyi graph with $n_0 = 4$ and $p_0 = 0.6$.

Pastor-Satorras model with fixed parameters. The results showed that the quality of the solutions returned by the proposed algorithms does not deviate significantly from the upper bound determined from the corresponding linear programming, at least for sufficiently small graphs for which the computation of these constraints was possible. The results were also compared with simple heuristics based on properties of the graph, such as sorting by degree or vertex neighbourhood containment relations. These heuristics, however proposed for other models, fail: as we know, for example, from considerations in [A1], the expected degree of a vertex $\mathbb{E}[\deg_t(u)]$ for a fixed t increases with u for certain ranges of parameters p and r. And indeed, it turns out that the results obtained with algorithms based on $p_{u,v}$ values are noticeably better and, moreover, their parameters can be tuned to obtain a desired precision-density tradeoff (see Figure 3).

The final step in the paper [C2] was to improve the proposed algorithms based on $p_{u,v}$ values to be able to take into account external knowledge in the form of so-called perfect pairs as certain (external) knowledge about the order of certain vertices in the true order π.

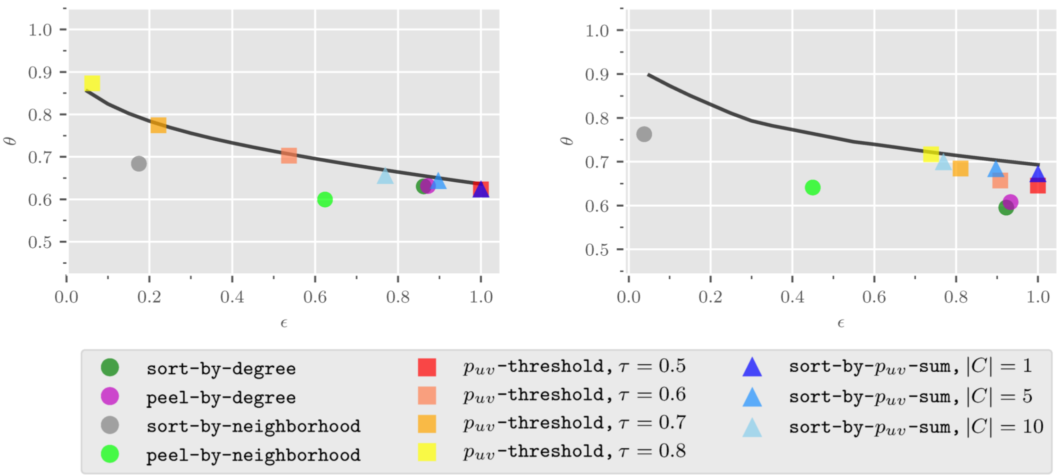

Figure 3: Results on greedy and $p_{u,v}$-based algorithms: $G_n \sim \mathtt{DD}(50, p, 1.0, G_{n_0})$ for $p = 0.3$ (left) and 0.6 (right), averaged over 100 graphs. $p_{u,v}$-based algorithms use $\sigma_{\text{tries}} = 100{,}000$. G_{n_0} is generated from Erdős-Renyi model with $n_0 = 10$ and $p_0 = 0.6$.

5.3. Entropy estimation of probability distributions

In the information theory community, one important research problem is the calculation of exact and asymptotic values of entropy for common probability distributions. Formally, for the discrete distribution described by $\Pr[X_n = k]$ for a given n and k we are looking for

$$H(X_n) = -\sum_k \Pr[X_n = k] \log_2 \Pr[X_n = k].$$

Previous work has led to the respective solutions for the binomial distribution [43, 62], the negative binomial distribution [31] and the Poisson distribution [76].

There were also formulae describing the exact entropy values for many other distributions, e.g. beta-dinomial and hypergeometric [26], were also derived. However, the results of these works were only formulae containing very complicated functions that were difficult to estimate asymptotically.

In the paper [D1], we answer an analogous question for the Dirichlet multinomial distribution, which was crucial for the analyses in the paper [A5]. At the same time, this distribution is directly related to the urn model of Pólya: for a given m of urns, each containing α_i $(i = 1, \ldots, m)$ balls, and a process of drawing balls with uniform probability with the addition of a ball (instead of subtraction, as with the basic urn model) to a chosen urn, the distribution of balls in the urn tends towards this distribution. In particular, for the variable $\bar{X}$ from the Dirichlet multinomial distribution $DM(n, \bar{\alpha})$ the probability of taking the value $\bar{x} = (x_1, \ldots, x_m)$ is equal to

$$\Pr[\bar{X} = \bar{x}] = \frac{\Gamma(n+1)\Gamma(\alpha_0)}{\Gamma(n+\alpha_0)} \prod_{k=1}^{m} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(x_k + 1)\Gamma(\alpha_k)},$$

where $\Gamma(x)$ is the Euler gamma function, while $\alpha_0 = \sum_{k=1}^{m} \alpha_k$.

The proof consisted in reducing the expression $H(\bar{X})$ to a sum of factors containing expressions of the form $\mathbb{E}[\Gamma(X_k + l)]$ for X_k having a beta-binomial distribution with parameters n, α_k and $\alpha_0 - \alpha_k$, and some constant values of l. Then, the fact that for the variable $X \sim BBin(n, \alpha, \beta)$ we have

$$\mathbb{E}[f(X)] = \int_0^1 \pi(p, \alpha, \beta) \mathbb{E}[f(X_p)] \, \mathrm{d}p$$

for the beta distribution probability function $\pi(p, \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$ (with Euler beta function $B(\alpha, \beta)$) and variable $X_p \sim Bin(n, p)$.

Each expression $\mathbb{E}[f(X_p + l)]$ was then decomposed into a Taylor series around the mean np and the resulting expressions were appropriately asymptotically estimated using hypergeometric functions and their expansions. Finally, it was obtained that

Theorem 5.3. *For the random variable $\bar{X} \sim DM(n, \bar{\alpha})$ it holds that*

$$\begin{aligned} H(\bar{X}) =& (m-1)\log n - \log\Gamma(\alpha_0) + \sum_{k=1}^{m} \log\Gamma(\alpha_k) + \log e \sum_{k=1}^{m} (\alpha_k - 1)(\psi(\alpha_k) - \psi(\alpha_0)) \\ &+ \sum_{s=1}^{\lceil \min\{\alpha_i\} \rceil - 1} e_s n^{-s} + O\left(\frac{\operatorname{polylog}(n)}{n^{\min\{\alpha_i\}}}\right), \end{aligned}$$

for explicitly computable coefficients e_s.

5.4. Task scheduling with incompatibility graph

Task scheduling is one of the most important areas of interest within operations research, as it includes many problems with various constraints and optimisation criteria [3]. It is indicated that this field has some practical significance in, among others, modelling optimisation issues e.g. for cloud computing [5, 25, 98].

The basic task scheduling problem, closely related to the set partitioning problem, denoted as $P||C_{max}$[7] is defined as follows:

Definition 5.1 ($P||C_{max}$)**.** *For a given set of tasks $J = \{j_1, \ldots, j_n\}$, a set of machines $M = \{m_1, \ldots, m_m\}$ and a processing time function $p\colon J \to \mathbb{N}_+$, find a schedule for allocating tasks to machines in which:*

1. *each task is executed on a certain machine i.e. the schedule assigns to each task j_k a pair of machine m_j and an interval $(t, t + p(j_k))$,*
2. *if machine m_j executes task j_k at time $[t, t + p(j_k)]$, then it does not execute any other task within this interval,*
3. *the maximum task completion time (C_{max}) over all the machines is as small as possible.*

This problem has a number of generalisations and a variety of variants, taking into account, for example, the different processing speed for different machines or the existence of dependencies between tasks. Their definition and overview can be found, for example, in [12, 23].

An example of such a broad class of problems is the task scheduling with incompatibility graphs, introduced in [15]. In these problems, it is assumed that an input is given in addition to the so called *incompatibility graph* G with $V(G) = J$ such that if $\{j_k, j_l\} \in E(G)$, then the corresponding tasks cannot be executed on the same machine at any time. This problem is a generalisation of other problems defined in task scheduling terms, e.g. Bounded Independent Sets [14] and Mutual Exclusion Scheduling [7, 49, 65].

In the past, there was shown a polynomial algorithm for $P|\chi(G) = k|C_{\max}$ i.e. when the incompatibility graph is k-colourable for a fixed k, as well as a 2-approximation algorithm for $P|G = \mathit{bipartite}|C_{\max}$ and a fully polynomial time approximation scheme (FPTAS) for graphs of bounded treewidth i.e. $P|tw(G) = k|C_{\max}$ [15]. There were investigated also incompatibility graphs that are unions of clique [36, 53, 66], for which there was developed a polynomial time approximation scheme (PTAS) for the case of identical machines and $(\log n)^{1/4-\varepsilon}$-non-aproximability (unless $\mathsf{P} = \mathsf{NP}$) for arbitrary machines.

As it can be seen, it is typically assumed that the class of constraint graphs has a relatively simple characteristic. This is justified by the direct connection of this class of serialisation problems with the colouring problems of the corresponding class of graphs, and hence with the transferability of results, e.g. about the NP-hardness of finding $O(n^{1-\varepsilon})$-approximations for the chromatic number of the graph [110].

In the work of [E1], the problem of task scheduling with a complete k-partite incompatibility graphs was addressed. Two variants of the problem were considered: when the number of partitions of the constraint graph is a parameter of the problem (denoted as *k-partite*) or when it is part of the input data (*multipartite*).

[7] We use here the so-called Lawler three-field notation [78], according to which the type of machines or problem is placed in the first field, e.g. P – identical machines, Q – uniform, R – arbitrary; in the second field constraints e.g. on task lengths, class of constraint graphs; in the third field an optimisation criterion, typically C_{max} or C_j.

Problem	**Approximation**	**Complexity**
$P\|G = \textit{complete multipartite}\|\sum C_j$	exact	$O(mn + n \log n)$
$Q\|G = \textit{complete k-partite}, p_j = 1\|C_{\max}$	exact	$O(mn^{k+1} \log(mn))$
$Q\|G = \textit{complete k-partite}, p_j = 1\|\sum C_j$	exact	$O(mn^{k+1})$
$Q\|G = \textit{complete k-partite}\|\sum C_j$	$1 + \varepsilon$ (PTAS)	P
$Q\|G = \textit{complete multipartite}, p_j = 1\|C_{\max}$	Strongly NP-hard	
	2	$O(mn \log(mn))$
	$1 + \varepsilon$ (PTAS)	P
$Q\|G = \textit{complete multipartite}, p_j = 1\|\sum C_j$	Strongly NP-hard	
	4	$O(m^2 n^3 \log m)$
$R\|G = \textit{complete 2-partite}, p_j \in \{p_1, p_2\}\|C_{\max}$	$O(n^b s_{\max}^{1-c})$-inapproximability for $b, c > 0$	
$R\|G = \textit{complete multipartite}\|C_{\max}$	$(1 + \epsilon)s_{\max}$	P
$R\|G = \textit{complete 2-partite}, p_j \in \{p_1, p_2\}\|\sum C_j$	$O(n^b s_{\max}^{1-c})$-inapproximability for $b, c > 0$	
$R\|G = \textit{complete multipartite}\|\sum C_j$	$s_{\max}$	$O(mn + n \log n)$

Table 2: A list of results from [E1].

For identical machines, it was previously known that $P|G = \textit{complete multipartite}|C_{\max}$ is strongly NP-hard, but at the same time there is a PTAS [15] for it. In the paper [E1] it is proved that if we change the criterion to $\sum C_j$, there exists an algorithm that returns an exact solution to the problem in time $O(mn + n \log n)$. This is a greedy algorithm, allocating machines to sets of tasks in such a way as to locally reduce the present value of the solution as much as possible. Within a fixed allocation of machines to sets of tasks, it is sufficient to compute the appropriate number of subproblems $P||\sum C_j$ to obtain the order of tasks on each machine.

For uniform machines, the starting point is to establish a strong NP-hardness for the problems $Q|G = \textit{complete multipartite}, p_j = 1|C_{\max}$ and $Q|G = \textit{complete multipartite}, p_j = 1|\sum C_j$. In both cases, a suitable reduction can be performed from the well-known problem 3-Partition [51]: it suffices that

- sizes will be directly mapped to machine speeds,
- the constraint on the sum of items in one subset will be mapped directly to the size of each partition J_i,
- the number of target subsets will be directly mapped to the number of partitions in the graph.

It then turns out that there is a required set partitioning for an instance of the 3-Partition problem if and only if there is a schedule with the appropriate $C_{\max}$ or $\sum C_j$ constraint.

For analogous problems in which the number of partitions of the constraint graph is not part of the input but is a part of the problem definition, there are suitable algorithms based on the idea of dynamic programming combined with binary search for an estimated value of $C_{\max}$ contained in the interval $[1, mn]$.

A PTAS based on a combination of three ideas has been developed for the problem $Q|G = \textit{complete k-partite}|\sum C_j$: rounding the speed of machines to the form $(1 + \varepsilon)^i$, exhaustive search for the allocation of the fastest machines for each partition of the constraint graph, and linear programming to determine the allocation of the remaining machines. Appropriate programme conditions ensure that all tasks have machines allocated to them, and an additional re-granulation procedure based on the idea of flow allows a solution in which individual tasks have been allocated to different machines to be converted into an acceptable schedule without increasing the total cost.

The next step was to present algorithms for the aforementioned strongly NP-hard problems: 2-approximation algorithm for $Q|G = \textit{complete multipartite}, p_j = 1|C_{\max}$ and 4-approximation algorithm for $Q|G = \textit{complete multipartite}, p_j = 1|\sum C_j$. Both algorithms were based on the idea of binary searching the length of the schedule, calculating according to this for each machine the capacity (i.e. the number of possible jobs to be processed), and then allocating the parts (sorted non-incrementally by size) to the machines (sorted non-incrementally by capacity) greedily until 50% coverage is reached. For the second problem $Q|G = \textit{complete multipartite}, p_j = 1|C_{\max}$ we found a PTAS, which combined a number of algorithmic techniques such as:

- guessing (binary search) of the upper bound on the value of $C_{\max}$,
- rounding of machine speeds to form $(1+\varepsilon)^i$,
- partitioning the machines by speed into very small, small, medium and large,
- partitioning of the constraint graph into ranges according to similar abundance and iterative processing of subsequent ranges,
- appropriate allocation of machines of different types within one step of the dynamic programme.

The crux of the proof is to show that, in the successive sets of state vectors for successive ranges, at least one vector will be *good* i.e., after rounding, it will provide a $(1+\varepsilon)$ approximation – and thus, if a suitable schedule exists for a given constraint $C_{\max}$, then the set of final state vectors will be non-empty and the approximate solution will be sought in it.

Complementing these results, it is shown that for arbitrary machines and the criterion $C_{\max}$ (respectively, $\sum C_j$) there exists a simple s_{max}-approximation algorithm (respectively, $((1+\varepsilon)s_{max})$-approximation algorithm) i.e. applying the algorithm to identical machines yields just such an approximation factor. On the other hand, it is shown that as long as $\mathsf{P} \neq \mathsf{NP}$ exists, there is no polynomial $O(n^b s_{max}^{1-c})$approximate algorithm for any $b, c > 0$ for both criteria $C_{\max}$ and $\sum C_j$ even if we are restricted to bipartite incompatibility graphs.

5.5. Backbone graph colouring and its generalisations

This line of research is a continuation of a PhD thesis investigating the issue of backbone graph colouring:

Definition 5.2 (λ-backbone graph colouring)**.** *For a given graph G and its spanning subgraph H, the function $c\colon V(G) \to \mathbb{N}_+$ is a* λ-backbone colouring of graph G with backbone H *when:*

- *for each edge $u, v \in E(G)$ it holds that $|c(u) - c(v)| \geq 1$,*
- *for each edge $u, v \in E(H)$ it holds that $|c(u) - c(v)| \geq \lambda$.*

Definition 5.3 (The λ-backbone graph colouring problem)**.** *For a given graph G and its spanning subgraph H, find the λ-backbone chromatic number $BBC_\lambda(G, H)$ i.e. the smallest number $k \in \mathbb{N}_+$ such that there exists a λ-backbone colouring c of graph G with backbone H satisfying $\max_{u \in V(G)} c(u) \leq k$.*

This problem, introduced in 2003 by Hajo Broersma [18] was a subject of a number of detailed studies both for different classes of graphs e.g. split graphs [20] or planar graphs [56], as well as for different classes of backbones e.g. matchings and disjoint stars [20] or forests [56]. It was also the subject of work within the series comprising the author's doctoral dissertation [F1],[F2],[F3],[F4]. Of particular interest was the case of $\lambda = 2$ for different classes of graphs and backbones [6, 19, 82, 83].

The paper [F5] is dedicated to the backbone colouring problem for cliques with backbone forests. Previously, it was proved in [63] that there is a 2-approximation algorithm for complete graphs with a bipartite backbone and a $\frac{3}{2}$-approximation algorithm for complete graphs with a consistent bipartite backbone. Both algorithms run in linear time. The first part of [F5] improves this result for forest backbones.

Theorem 5.4. *For a forest F on n vertices and $\lambda \geq 2$, it holds that $BBC_\lambda(K_n, F) \leq \max\{n, 2\lambda\} + \Delta^2(F)\lceil \log n \rceil$.*

In addition, there is an algorithm running in time $O(n)$ finding the appropriate λ-backbone colouring of the graph.

Specifically, the implication is that there is a linear algorithm with an additive error not exceeding $\Delta^2(F)\lceil \log n \rceil$ – which improves the previous restriction for $\Delta(F) = o(\sqrt{n}/\log n)$.

The proof is based on the idea of a red-blue-yellow (k, l)-decomposition, i.e., splitting the vertices of the tree into three independent sets R, B, Y such that R and B are independent sets with $||R| - |B|| \leq k$, while Y is a set with $|Y| \leq l = \lceil \log n \rceil$. This decomposition is applied recursively, according to the division of the tree into subtrees of at most half the size of the parent tree (which can be done in $O(n)$ time). By appropriately allocating colours to the vertices of

Y (smallest and largest), and ensuring that there are appropriate differences between the colour ranges allocated to each set of decompositions, it can be proved that the resulting colouring satisfies the given constraints.

The second part of the paper is the construction of an infinite family of trees of bounded degree ($\Delta(T_r) = 3$) such that $BBC_\lambda(K_n, T_r) \geq \max\{n, 2\lambda\} + \Omega(\log n)$. These trees, named Fibonacci trees because the construction of the r-th Fibonacci tree T_r involves combining the $(r-1)$-th and $(r-2)$-th Fibonacci trees with three additional vertices into a new tree.

It turns out that finding a colouring or equivalent red-blue-yellow (k, l)-decomposition of the T_r-tree for $l = o(\log n)$ reduces to the problem of finding such numbers $a_i \in \{-1, 0, 1\}$ such that $\sum_{i=0}^{r} |a_i| \leq l + 1$ and $\sum_{i=0}^{r} a_i F_i = \frac{F_r}{2} + o(\log n)$ occurs, where F_i denotes the i-th Fibonacci number. Intuitively, Y determines our division of T_r into $2|Y|+1$ trees[8], while the sign a_i determines whether there are more red or blue vertices in a given subtreeThe construction implies that the i-th Fibonacci tree coloured into 2 colours has exactly F_i more vertices in one of the colours. The component $o(\log n)$ is responsible for the possible modifications related to the selection of vertices for the set Y.

There exists a theorem by Zeckendorf which states the decomposition of any number into the sum of Fibonacci numbers, we know that $\frac{F_r}{2}$ can be decomposed into the sum of at least $\frac{r}{3}$ Fibonacci numbers [77]. Generalizing this to allow both sums and differences of Fibonacci numbers, we proved that a tree with F_r vertices requires decomposition into at least $\Omega(r)$ red-blue subtrees – and thus it requires that $|Y| = \Omega(r)$. Or, equivalently, there exists n-vertex tree requiring $|Y| = \Omega(\log n)$.

A generalization of the λ-backbone colouring problem and at the same time a certain formalisation of the frequency assignment problem [39, 55] is the so-called ξ-colouring problem: for each pair of transmitting stations, we require that their frequencies are spaced by certain bands (depending on the position and strength of the transmitters) so that they do not interfere with each other.

Definition 5.4 (ξ-colouring)**.** *For a given graph G and a function $\xi \colon E(G) \to \mathbb{N}_+$ a function $c \colon V(G) \to \mathbb{N}_+$ is a* ξ-colouring *of graph G when for each edge $\{u, v\} \in E(G)$ it holds that $|c(u) - c(v)| \geq \xi(\{u, v\})$.*

This problem is also a generalization of the $L(p, q)$-labelling problem [54], in which we require that neighbours in a graph receive colours that differ by at least p, while vertices within distance 2 receive colours that differ by at least q (see also a review of the $L(p, q)$-labelling results in [24, 104]).

Definition 5.5 (The problem of the minimum ξ-span of a graph)**.** *For a given graph G and a function $\xi \colon E(G) \to \mathbb{N}_+$, find* $\mathrm{sp}(G, \xi)$ *as the smallest number $k \in \mathbb{N}_+$ such that there exists a ξ-colouring c of G satisfying* $\max_{u \in V(G)} c(u) \leq k$.

In addition to the span, which is a concept generalizing the chromatic number, the so-called edge span is also considered in the literature for the aforementioned colouring models as a local optimisation criterion [60, 103]

Definition 5.6 (The problem of minimum edge ξ-span of a graph)**.** *For a given graph G and a function $\xi \colon E(G) \to \mathbb{N}_+$, find* $\mathrm{esp}(G, \xi)$ *as the smallest number $k \in \mathbb{N}_+$ such that there exists a ξ-colouring c of G satisfying* $\max_{\{u,v\} \in \mathbb{E}(G)} |c(u) - c(v)| \leq k$.

The paper [F6] presents a number of results related to edge ξ-spanning graphs. In particular, by polynomial reduction from the NP-complete set partitioning problem, it was shown that this problem remains difficult even for subcubic outerplanar graphs. It is also proved that for cacti, i.e. for graphs not containing edges contained in multiple cycles, there exists an algorithm that returns an optimal colouring, achieving $\mathrm{esp}(G, \xi)$ and running in time $O(n \log n)$.

The work of [F7] contains results on ξ-graph spanning for subcubic graphs. It shows that the problem even in this restricted case remains strongly NP-hard, in contrast to the classical graph colouring problem and even the λ-backbone colouring problem, which can be solved in polynomial time [64]. The proof consists in a reduction from the well-known NP-hard problem Not-All-Equal 3-SAT [91]. Moreover, it turns out that, assuming $P \neq NP$, there cannot exist a $(\frac{3}{2} - \varepsilon)$-approximation algorithm for the ξ-spanning problem of subcubic graphs with the function ξ taking at most two values.

[8] There are at most $|Y| + 1$ trees, but we allow some further local operations.

If the function ξ takes at most two values, then we proved that there is a $\frac{3}{2}$-approximation algorithm for the above problem running in time $O(n+m)$. And for any function ξ, in turn, there exists a 2-approximation algorithm, also with time complexity $O(n+m)$. Interestingly, if we assume that the graph induced by edges with maximum weights ξ form a spanning subgraph, we show in the paper that there exists an exact algorithm running in time $O(n^2)$ and a $\frac{4}{3}$-approximate algorithm with time complexity $O(n+m)$, respectively.

Another related problem to backbone colouring and ξ-colouring is T-colouring, introduced in [55]:

Definition 5.7 (T-colouring)**.** *For a given graph G and a finite set $T \subseteq \mathbb{N}$ $(0 \in T)$, the function $c\colon V(G) \to \mathbb{N}_+$ is a* T-colouring of graph G for a set T *when, for each edge $\{u,v\} \in E(G)$ it is true that $|c(u) - c(v)| \notin T$.*

Also, for this problem, the minimum T-span $\mathrm{sp}(G,T)$ and the minimum edge T-span $\mathrm{esp}(G,T)$ can be defined [35]. Previous research on $\mathrm{esp}(G,T)$ has mainly consisted of looking for the relationship between $\mathrm{esp}(G,T)$ and $\mathrm{sp}(G,T)$ for selected classes of graphs and sets T [69, 90, 107].

The paper [F8] was devoted to the relationship between edge T-span and circular chromatic number, as defined in [100]. To this end, the operation $\odot$ is introduced as $d \odot T := \{0 \leq t \leq d(\max T + 1)\colon d|t \Rightarrow t/d \in T\}$ for the number $d \in \mathbb{N}_+$ and the set $T \subseteq \mathbb{N}$.

The paper presents a number of results concerning $\mathrm{esp}(G, d \odot T)$, $\mathrm{esp}(G,T)$, and $\mathrm{sp}(G,T)$. The main result concerns $\mathrm{esp}(G,T)$ for the set $T = d \odot \{0\} = \{0, 1, \ldots, d-1\}$. Note that colouring any graph G with the set $d \odot \{0\}$ is a special case of ξ-colouring with $\xi(e) = d-1$ for all $e \in E(G)$. For this case it is proved that

Theorem 5.5. *For any graph G and any number $d \in E(G)$, it holds that*

$$\chi_c(G) = 1 + \inf\{\mathrm{esp}_{d\odot\{0\}}(G)/d\colon d \geq 1\}.$$

If $\chi_c(G) = k/d$ for some $d \in \{1, 2, \ldots, k\}$ then $\chi_c(G) = 1 + \mathrm{esp}_{d\odot\{0\}}(G)/d$.

This relation allowed us to resolve the open problem concerning the T-colouring of powers of cycles C_n^p posed in [107]:

Theorem 5.6. *For any $n, d, p \in \mathbb{N}_+$, it holds that*

$$\mathrm{esp}_{d\odot\{0\}}(C_n^p) = pd + \lceil rd/q \rceil\,.$$

The authors of the hypothesis also proved in [107] that the above theorem is true when $p \geq (q - p\lfloor q/p \rfloor)d$. Using Theorem 5.5 and proving that it always holds $\chi_c(C_n^p) = n/q$ we showed in [F8] that the above theorem also holds in the general case.

5.6. Chromatic games

The basic chromatic game, proposed in [50] and reintroduced in [13] in the formal context of graph theory, is a popular research topic that has seen many variants for different classes of graphs and rules of motion (see overview of results in [9]). In the basic version, two players, Alice and Bob, are provided with a set of colours $\{1, \ldots, k\}$ and make alternating moves. Each move consists of allocating to yet uncoloured vertex one of the colours from the pool so as to maintain correct (partial) colouring i.e. that no two neighbouring coloured vertices have the same colour. Alice's goal is to colour the entire graph, while Bob's goal is to enforce a situation when a player cannot make a legal move. We denote by the *game chromatic number* $\chi_g(G)$ the smallest number k for which Alice has a winning strategy, i.e., regardless of Bob's strategy, a proper colouring of the entire graph is always obtained.

In estimates related to chromatic games, the notion of the so-called colouring number (*colouring number*) and its generalizations sometimes appear:

Definition 5.8. *Let G be a graph and $\prec$ a linear ordering on $V(G)$. Let $N_G^-(v,\prec) = \{u \in V(G)\colon \{u,v\} \in E(G) \wedge u \prec v\}$ denote the backward neighbourhood of v in G. Then*

$$\mathit{col}(G) = \min\{k\colon \exists_{\prec}\forall_{v \in V(G)} |N_G^-(v,\prec)| \leq k-1\}.$$

Definition 5.9. *Let G be a graph and $\prec$ a linear ordering on $V(G)$. For any $r \in \mathbb{N}_+$ let*

$N_G^-(v, r, \prec) = \{u \in V(G) \colon \exists_{w_1,\ldots,w_{r-1}} \{uw_1, w_1w_2, \ldots, w_{r-1}v\} \subseteq E(G) \wedge u \prec v \wedge \forall_{i=1,\ldots,r-1} v \prec w_i\}$

denote the k-th backward neighbourhood of v in G. Then

$$col_r(G) = \min\{k \colon \exists_\prec \forall_{v \in V(G)} |N_G^-(v, r, \prec)| \leq k - 1\}.$$

For example, in the paper [74] it is shown that for any planar graph G there exists $\chi_g(G) \leq 4\,\mathrm{col}_2(G) + 1$. For general graphs it is proved that the inequality $\chi_g(G) \leq \chi(G)(\mathrm{col}_2(G) + 1)$ [9] holds. It can also be pointed out that the parameter $\mathrm{col}_2(G)$ appears in the proofs and constraints of such graph parameters as the acyclic chromatic number [75] or the oriented game chromatic number [73].

In the paper [G1] the properties of the parameter $\mathrm{col}_2(G)$ were investigated. First, we improved for the case $r = 2$ the restriction $\mathrm{col}_r(G)$ presented by Kierstead and Kostochka in [72] proving that:

Theorem 5.7. *For any graph G, it holds that $\mathrm{col}_2(G) \leq \frac{1}{2}\Delta(G)(\Delta(G) - 1) + 2$.*

Second, it is shown that there is a large class of graphs for which the relation $\mathrm{col}_2(G) = \Theta(\Delta(G)^2)$ does indeed hold:

Theorem 5.8. *For any regular graph G containing neither C_3 nor C_4, it holds that*

$$\mathrm{col}_2(G) \geq \frac{\Delta(G)^2}{8} + \frac{\Delta(G)}{4} + 1.$$

The proof is based on a precise count of triples of the form (j, i, k) such that $j < i < k$ and $\{v_i, v_j\}, \{v_j, v_k\} \in E(G)$. This way, it became possible to estimate from below the sum of all "second backward degrees" (that is, the neighbourhood sizes $N_G^-(v, 2, \prec)$) for any ordering $\prec$ as some function of the number of edges of the graph G – and hence the corresponding estimate of the maximum. It is worth noting that this implies that for such graphs $\mathrm{col}_2(G)$ is a bad estimate of $\chi_g(G)$, since it is well-known that $\chi_g(G) \leq \Delta(G) + 1$ for any graph G.

Third, a polynomial algorithm has been proposed to compute the value of $\mathrm{col}_2(G)$ for subcubic graphs i.e. with $\Delta(G) \leq 3$. The algorithm is based on recursively finding and removing small subgraphs from the graph in a way that preserves their values of $\mathrm{col}_2(G)$ – or until it encounters a small subgraph H that guarantees that $\mathrm{col}_2(G) = \mathrm{col}_2(H)$.

The work of [G2] was devoted to a game called *infinite graph colouring game.* It differs from the usual chromatic game in the assumption that the game sides do not have a fixed finite set of colours, but rather an infinite one. As in the basic variant, Alice aims to colour the graph with as few colours as possible, while Bob does the opposite, trying to maximise the number of colours in use. Correspondingly, the *infinite game chromatic number* $\chi_g^\infty(G)$ is defined as the number of colours used by the players when they use optimal strategies.

An obvious bound is $\chi_g^\infty(G) \geq 1 + \lfloor \frac{n(G)}{2} \rfloor$, since Alice starts and Bob can use a new colour in each move. First, we proved in [G2] that

Theorem 5.9. *For all graphs G it is true that*

$$\chi_g^\infty(G) \leq \min\left\{\left\lfloor \frac{1}{2} n(G) \right\rfloor + \chi(G), n(G) + 1 - \left\lceil \frac{1}{2}\alpha(G) \right\rceil \right\}.$$

It is also shown that the problem is not at all trivial from the point of view of strategy there is an infinite family of graphs for which Bob's optimal strategy is not at all to use a new colour every time – but sometimes a well-placed repetition of an already used colour is more profitable for him.

Next, a number of results for particular classes of graphs are presented.

Theorem 5.10. *For a graph G satisfying $\Delta(G) \leq \frac{1}{3}(n(G) - 1)$ it is true that $\chi_g^\infty(G) = \lfloor \frac{1}{2} n(G) \rfloor + 1$.*

This allowed us to obtain complete results for subcubic graphs.

Theorem 5.11. *For graphs G with $\Delta(G) \leq 3$, the following holds:*

$$\chi_g^\infty(G) = \begin{cases} 3 & \text{for } G = K_3, \\ 4 & \text{for } G \in \{C_4, K_4 - e, K_4\}, \\ \lfloor \frac{1}{2} n(G) \rfloor + 1 & \text{otherwise.} \end{cases}$$

Complete results were also obtained for k-partite complete graphs:

Theorem 5.12. *Let l be the number of odd numbers in the multiset $\{r_1, r_2, \ldots, r_k\}$. Then*

$$\chi_g^\infty(K_{r_1,r_2,\ldots,r_k}) = \begin{cases} \lfloor \frac{1}{2} n(G) \rfloor + \frac{l+1}{2} & \text{for } l \text{ odd,} \\ \lfloor \frac{1}{2} n(G) \rfloor + k - \frac{l}{2} & \text{for } l \text{ even.} \end{cases}$$

The most important result of the paper is to show that there is a strategy for Bob proving that for all graphs it is true that $\chi_g^\infty(G) \leq n(G) - \alpha'(\bar{G})$ and a strategy for Alice proving that for all graphs with an odd number of vertices it holds that $\chi_g^\infty(G) \leq n(G) - \alpha'(\bar{G})$.

5.7. Reconstruction of hypergraphs

One of the operations studied in the context of hypergraphs is the so-called 2-subsection of a hypergraph. This operation consists of replacing each hyperedge in the graph by a clique and additionally removing the resulting duplicated edges. Analyses of graphs obtained through such operations are claimed to have applications in computational biology [42], language theory [52] and compiler optimisation [88], among others.

However, since this approach implies the loss of information about the original hypergraph, one can modify this problem by leaving multiple edges or, equivalently, natural weights on the edges, and define the following hypergraph reconstruction problem:

Definition 5.10 (Hypergraph reconstruction problem)**.** *For a given graph G and a weight function $w\colon E(G) \to \mathbb{N}_+$, is there a hypergraph H such that (G, w) is its weighted 2-subsection?*

For the optimisation version of this problem, i.e. when we are looking for a suitable hypergraph with the minimum number of hypergraphs, it is proved that it belongs to the class FPTwhen it is parameterised by the output size [41]. If, on the other hand, we have an optimisation problem but without a given weight function, it has been shown that the problem is NP-hard for planar graphs [79], K_4-free graphs [81] and split graphs [101], but there are FPT algorithms for planar graphs and graphs without K_4 when the output size [47] is parameterised.

In the paper [H1] it was shown that the decision problem is also NP-complete for complete graphs even when the edge weights belong to the set $\{1, 2\}$. The proof consisted of a reduction from the well-known NP-hard edge problem of 3-colouring of connected cubic graphs [58].

Moreover, we showed in the same paper that it is possible to find algorithms that solve the problem and find a suitable hypergraph in polynomial time for partial 2-trees, 2-degenerate graphs, and graphs with $\Delta(G) \leq 4$.

6. Presentation of teaching and organizational achievements

6.1. Teaching

6.1.1. Courses fully prepared by applicant

- String algorithms (lecture, tutorial) at Jagiellonian University.
- Programming language: C++ (lecture, laboratory) at Jagiellonian University.
- Programming language: Python (lecture, laboratory) at Jagiellonian University.
- Mobile programming (laboratory) at Jagiellonian University.
- Programming languages (lecture, laboratory) at Gdańsk University of Technology.

6.1.2. Other teaching activities

- Concurrent programming (laboratory) at Jagiellonian University.
- Distributed systems (laboratory) at Jagiellonian University.
- Principles of programming (laboratory) at Gdańsk University of Technology.
- Algorithms and data structures (laboratory) at Gdańsk University of Technology.
- Discrete optimization algorithms (laboratory) at Gdańsk University of Technology.
- Analysis of algorithms (tutorial) at Gdańsk University of Technology.
- Elements of bioinformatics (laboratory) at Gdańsk University of Technology.
- Bioinformatics (laboratory) at Gdańsk University of Technology.
- Modelling and systems simulation (laboratory) at Gdańsk University of Technology.

6.2. Supervised theses

6.2.1. Supervised master theses at Jagiellonian University

1. Adrian Siwiec, *Perfect graph recognition and colouring*, 2020.
2. Paweł Palenica, *Random graph compression algorithms*, 2020.
3. Wojciech Grabis, *Implementation of selected macroeconomic DSGE models*, 2022.
4. Michał Stobierski, *Fast computation of maximum flows using combinatorial methods*, 2022.

6.2.2. Supervised bachelor theses at Jagiellonian University

1. Mateusz Górski, *The generalization of Turán number*, 2020.
2. Marcin Serwin, *Overview of cograph related algorithms*, 2020.
3. Krzysztof Michalik, *On λ-backbone colouring of cliques with tree backbones in linear time*, 2021.
4. Mikołaj Twaróg, *Polynomial time approximation schemes for NP-hard problems on planar graphs*, 2021.
5. Mateusz Pach, *Distributed consensus protocols*, 2022.
6. Inka Sokołowska, *Task scheduling for block-type conflict graphs*, 2022.

7. Obtained grants

I am the principal investigator of the NCN SONATA grant for the project: „Towards Estimation of Information Content for Graph Structures" (no. 2020/39/D/ST6/00419).

References

[1] Milton Abramowitz and Irene Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Dover Publications, 1972.

[2] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 171–180, 2000.

[3] Ali Allahverdi, Chi To Ng, T.C. Edwin Cheng, and Mikhail Kovalyov. A survey of scheduling problems with setup times or costs. *European Journal of Operational Research*, 187(3):985–1032, 2008.

[4] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.

[5] Aida Amini Motlagh, Ali Movaghar, and Amir Masoud Rahmani. Task scheduling mechanisms in cloud computing: A systematic review. *International Journal of Communication Systems*, 33(6):e4302, 2020.

[6] Camila Araujo, Julio Araujo, Ana Silva, and Alexandre Cezar. Backbone coloring of graphs with galaxy backbones. *Electronic Notes in Theoretical Computer Science*, 346:53–64, 2019.

[7] Brenda Baker and Edward Coffman Jr. Mutual exclusion scheduling. *Theoretical Computer Science*, 162(2):225–243, 1996.

[8] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[9] Tomasz Bartnicki, Jarosław Grytczuk, Hal Kierstead, and Xuding Zhu. The map-coloring game. *The American Mathematical Monthly*, 114(9):793–803, 2007.

[10] Maciej Besta and Torsten Hoefler. Survey and taxonomy of lossless graph compression and space-efficient graph representations. *arXiv preprint arXiv:1806.01799*, 2018.

[11] Ashish Bhan, David Galas, and T Gregory Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

[12] Jacek Błażewicz, Klaus Ecker, Erwin Pesch, Günter Schmidt, and Jan Węglarz. *Handbook on scheduling: from theory to applications*. Springer, 2019.

[13] Hans Bodlaender. On the complexity of some coloring games. *International Journal of Foundations of Computer Science*, 2(02):133–147, 1991.

[14] Hans Bodlaender and Klaus Jansen. On the complexity of scheduling incompatible jobs with unit-times. In *International Symposium on Mathematical Foundations of Computer Science*, pages 291–300. Springer, 1993.

[15] Hans Bodlaender, Klaus Jansen, and Gerhard Woeginger. Scheduling with incompatible jobs. *Discrete Applied Mathematics*, 55(3):219–232, 1994.

[16] Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[17] Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The degree sequence of a scale-free random graph process. In *The Structure and Dynamics of Networks*, pages 384–395. Princeton University Press, 2011.

[18] Hajo Broersma. A general framework for coloring problems: old results, new results, and open problems. In *Indonesia-Japan Joint Conference on Combinatorial Geometry and Graph Theory*, pages 65–79. Springer, 2003.

[19] Hajo Broersma, Fedor Fomin, Petr Golovach, and Gerhard Woeginger. Backbone colorings for graphs: tree and path backbones. *Journal of Graph Theory*, 55(2):137–152, 2007.

[20] Hajo Broersma, Bert Marchal, Daniël Paulusma, and A.N.M. Salman. Backbone colorings along stars and matchings in split graphs: their span is close to the chromatic number. *Discussiones Mathematicae Graph Theory*, 29(1):143–162, 2009.

[21] Anna Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1):1–10, 2019.

[22] Frederick Brooks Jr. Three great challenges for half-century-old computer science. *Journal of the ACM*, 50(1):25–26, 2003.

[23] Peter Brucker. *Scheduling Algorithms.* Springer-Verlag, 2007.

[24] Tiziana Calamoneri. The $L(h, k)$-labelling problem: an updated survey and annotated bibliography. *The Computer Journal*, 54(8):1344–1371, 2011.

[25] Weihong Chen, Guoqi Xie, Renfa Li, Yang Bai, Chunnian Fan, and Keqin Li. Efficient task scheduling for budget constrained parallel applications on heterogeneous cloud computing systems. *Future Generation Computer Systems*, 74:1–11, 2017.

[26] Mahdi Cheraghchi. Expressions for the entropy of basic discrete distributions. *IEEE Transactions on Information Theory*, 65(7):3999–4009, 2019.

[27] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Alessandro Panconesi, and Prabhakar Raghavan. Models for the compressible web. *SIAM Journal on Computing*, 42(5):1777–1802, 2013.

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

[29] Fan Chung and Linyuan Lu. *Complex graphs and networks.* CBMS Regional Conference Series in Mathematics. American Mathematical Society, 2006.

[30] Fan Chung, Linyuan Lu, T. Gregory Dewey, and David Galas. Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687, 2003.

[31] Jacek Cichoń and Zbigniew Gołębiewski. On Bernoulli Sums and Bernstein Polynomials. In *23rd International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms*, pages 179–190. Discrete Mathematics and Theoretical Computer Science, 2012.

[32] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[33] Recep Colak, Fereydoun Hormozdiari, Flavia Moser, Alexander Schönhuth, J Holman, Martin Ester, and Süleyman Cenk Sahinalp. Dense graphlet statistics of protein interaction and random networks. In *Biocomputing 2009*, pages 178–189. World Scientific Publishing, Singapore, 2009.

[34] Thomas Cover and Joy Thomas. *Elements of Information Theory.* John Wiley & Sons, 2006.

[35] Margaret Cozzens and Fred Roberts. T-colorings of graphs and the channel assignment problem. *Congressus Numerantium*, 35(b):191–208, 1982.

[36] Syamantak Das and Andreas Wiese. On minimizing the makespan when some jobs cannot be assigned on the same machine. In *25th Annual European Symposium on Algorithms (ESA 2017)*, volume 87 of *LIPIcs*, pages 31:1–31:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[37] Reinhard Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics.* Springer, 2006.

[38] Michael Drmota. *Random trees: an interplay between combinatorics and probability.* Springer Science & Business Media, 2009.

[39] Andreas Eisenblätter, Martin Grötschel, and Arie Koster. Frequency planning and ramifications of coloring. *Discussiones Mathematicae Graph Theory*, 1(22):51–88, 2002.

[40] Paul Erdős and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae*, 6:290–297, 1959.

[41] Andreas Emil Feldmann, Davis Issac, and Ashutosh Rai. Fixed-parameter tractability of the weighted edge clique partition problem. In *15th International Symposium on Parameterized and Exact Computation (IPEC 2020)*, volume 180, pages 17:1–16. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[42] Andres Figueroa, James Borneman, and Tao Jiang. Clustering binary fingerprint vectors with missing values for DNA array data analysis. *Journal of Computational biology*, 11(5):887–901, 2004.

[43] Philippe Flajolet. Singularity analysis and asymptotics of Bernoulli sums. *Theoretical Computer Science*, 215(1):371–381, 1999.

[44] Philippe Flajolet and Andrew Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240, 1990.

[45] Philippe Flajolet and Robert Sedgwick. *Analytic Combinatorics.* Cambridge University Press, 2009.

[46] Abraham Flaxman, Alan Frieze, and Trevor Fenner. High degree vertices and eigenvalues in the preferential attachment graph. *Internet Mathematics*, 2(1):1–19, 2005.

[47] Rudolf Fleischer and Xiaotian Wu. Edge clique partition of K_4-free and planar graphs. In *International Conference on Computational Geometry, Graphs and Applications*, pages 84–95, 2010.

[48] Alan Frieze and Michał Karoński. *Introduction to random graphs.* Cambridge University Press, 2016.

[49] Frédéric Gardi. Mutual exclusion scheduling with interval graphs or related classes: complexity and algorithms. *4OR*, 4(1):87–90, 2006.

[50] Martin Gardner. Mathematical games. *Scientific American*, 222:132–140, 1970.

[51] Michael Garey and David Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, United States of America, 1979.

[52] Floris Geerts, Bart Goethals, and Taneli Mielikäinen. Tiling databases. In *International Conference on Discovery Science*, pages 278–289. Springer, 2004.

[53] Kilian Grage, Klaus Jansen, and Kim-Manuel Klein. An EPTAS for machine scheduling with bag-constraints. In *The 31st ACM Symposium on Parallelism in Algorithms and Architectures*, pages 135–144. ACM, 2019.

[54] Jerrold Griggs and Roger Yeh. Labeling graphs with a condition at distance 2. *SIAM Journal of Discrete Mathematics*, 5:586–595, 1992.

[55] William Hale. Frequency assignment: Theory and applications. *Proceedings of the IEEE*, 68(12):1497–1514, 1980.

[56] Frédéric Havet, Andrew King, Mathieu Liedloff, and Ioan Todinca. (Circular) backbone colouring: Forest backbones in planar graphs. *Discrete Applied Mathematics*, 169:119–134, 2014.

[57] Felix Hermann and Peter Pfaffelhuber. Large-scale behavior of the partial duplication random graph. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 13:687–710, 2016.

[58] Ian Holyer. The NP-completeness of edge-coloring. *SIAM Journal on Computing*, 10(4):718–720, 1981.

[59] Fereydoun Hormozdiari, Petra Berenbrink, Nataša Pržulj, and Süleyman Cenk Sahinalp. Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. *PLoS Computational Biology*, 3(7):e118, 2007.

[60] Shin-Jie Hu, Su-Tzu Juan, and Gerard J Chang. T-colorings and T-edge spans of graphs. *Graphs and Combinatorics*, 15(3):295–301, 1999.

[61] Iaroslav Ispolatov, Paul Krapivsky, and Anton Yuryev. Duplication-divergence model of protein interaction network. *Physical Review E*, 71(6):061911, 2005.

[62] Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theoretical Computer Science*, 201(1-2):1–62, 1998.

[63] Robert Janczewski and Krzysztof Turowski. The backbone coloring problem for bipartite backbones. *Graphs and Combinatorics*, 31(5):1487–1496, 2015.

[64] Robert Janczewski and Krzysztof Turowski. The computational complexity of the backbone coloring problem for bounded-degree graphs with connected backbones. *Information Processing Letters*, 115(2):232–236, 2015.

[65] Klaus Jansen. The mutual exclusion scheduling problem for permutation and comparability graphs. *Information and Computation*, 180(2):71–81, 2003.

[66] Klaus Jansen, Alexandra Lassota, Marten Maack, and Tytus Pikies. Total completion time minimization for scheduling with incompatibility cliques. In Susanne Biundo, Minh Do, Robert Goldman, Michael Katz, Qiang Yang, and Hankz Hankui Zhuo, editors, *Proceedings of the Thirty-First International Conference on Automated Planning and Scheduling, ICAPS 2021, Guangzhou, China (virtual), August 2-13, 2021*, pages 192–200. AAAI Press, 2021.

[67] Svante Janson, Andrzej Ruciński, and Tomasz Łuczak. *Random graphs*. John Wiley & Sons, 2011.

[68] Jonathan Jordan. The connected component of the partial duplication graph. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 15:1431–1445, 2018.

[69] Justie Su-Tzu Juan, I-fan Sun, and Pin-Xian Wu. T-coloring on folded hypercubes. *Taiwanese Journal of Mathematics*, 13(4):1331–1341, 2009.

[70] Raya Khanin and Ernst Wit. How scale-free are biological networks. *Journal of Computational Biology*, 13(3):810–818, 2006.

[71] John Kieffer, En-Hui Yang Yang, and Wojciech Szpankowski. Structural complexity of random binary trees. In *2009 IEEE International Symposium on Information Theory*, pages 635–639. IEEE, 2009.

[72] Hal Kierstead and Alexandr Kostochka. Efficient graph packing via game colouring. *Combinatorics, Probability and Computing*, 18(5):765–774, 2009.

[73] Hal Kierstead, Bojan Mohar, Simon Špacapan, Daqing Yang, and Xuding Zhu. The two-coloring number and degenerate colorings of planar graphs. *SIAM Journal on Discrete Mathematics*, 23(3):1548–1560, 2009.

[74] Hal Kierstead and William Trotter. Planar graph coloring with an uncooperative partner. *Journal of Graph Theory*, 18(6):569–584, 1994.

[75] Hal Kierstead and Daqing Yang. Orderings on graphs and game coloring number. *Order*, 20(3):255–264, 2003.

[76] Charles Knessl. Integral representations and asymptotic expansions for Shannon and Renyi entropies. *Applied Mathematics Letters*, 11(2):69–74, 1998.

[77] Donald Knuth. Fibonacci multiplication. *Applied Mathematics Letters*, 1(1):57–60, 1988.

[78] Eugene Lawler, Jan Karel Lenstra, and Alexander Rinnooy Kan. Recent developments in deterministic sequencing and scheduling: A survey. In Michael Dempster, Jan Karel Lenstra, and Alexander Rinnooy Kan, editors, *Deterministic and Stochastic Scheduling*, volume 84 of *NATO Advanced Study Institutes Series* (*Series C – Mathematical and Physical Sciences*), pages 35–73. Springer, 1982.

[79] Hoang-Oanh Le and Van Bang Le. Constrained representations of map graphs and half-squares. In *44th International Symposium on Mathematical Foundations of Computer Science (MFCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[80] Tomasz Łuczak, Abram Magner, and Wojciech Szpankowski. Asymmetry and structural information in preferential attachment graphs. *Random Structures & Algorithms*, 55(3):696–718, 2019.

[81] S.H. Ma, Walter Wallis, and Julin Wu. The complexity of the clique partition number problem. *Congressium Numerantium*, 67:59–66, 1988.

[82] Jozef Miškuf, Riste Škrekovski, and Martin Tancer. Backbone colorings and generalized mycielski graphs. *SIAM Journal on Discrete Mathematics*, 23(2):1063–1070, 2009.

[83] Jozef Miškuf, Riste Škrekovski, and Martin Tancer. Backbone colorings of graphs with bounded degree. *Discrete Applied Mathematics*, 158(5):534–542, 2010.

[84] Saket Navlakha and Carl Kingsford. Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Computational Biology*, 7(4):e1001119, 2011.

[85] Mark Newman. *Networks*. Oxford University Press, 2018.

[86] Susumu Ohno. *Evolution by gene duplication*. Springer-Verlag, Berlin–Heidelberg, 1970.

[87] Romualdo Pastor-Satorras, Eric Smith, and Ricard V Solé. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222(2):199–210, 2003.

[88] Subramanian Rajagopalan, Manish Vachharajani, and Sharad Malik. Handling irregular ILP within conventional VLIW schedulers using artificial resource constraints. In *Proceedings of the 2000 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems*, pages 157–164. ACM, 2000.

[89] Alpan Raval. Some asymptotic properties of duplication graphs. *Physical Review E*, 68(6):066119, 2003.

[90] Arundhati Raychaudhuri. Further results on T-coloring and frequency assignment problems. *SIAM Journal on Discrete Mathematics*, 7(4):605–613, 1994.

[91] Thomas Schaefer. The complexity of satisfiability problems. In *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing*, pages 216–226, 1978.

[92] Mingyu Shao, Yi Yang, Jihong Guan, and Shuigeng Zhou. Choosing appropriate models for protein–protein interaction networks: a comparison study. *Briefings in Bioinformatics*, 15(5):823–838, 2013.

[93] Ricard Solé, Romualdo Pastor-Satorras, Eric Smith, and Thomas Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(01):43–54, 2002.

[94] Jithin Sreedharan, Abram Magner, Ananth Grama, and Wojciech Szpankowski. Inferring temporal information from a snapshot of a dynamic network. *Nature Scientific Reports*, 9(1):1–10, 2019.

[95] Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York, 2001.

[96] Wojciech Szpankowski and Ananth Grama. Frontiers of science of information: Shannon meets turing. *Computer*, 51(1):28–38, 2018.

[97] Reiko Tanaka, Tau-Mu Yi, and John Doyle. Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, 579(23):5140–5144, 2005.

[98] Chun-Wei Tsai and Joel Rodrigues. Metaheuristic scheduling for cloud: A survey. *IEEE Systems Journal*, 8(1):279–291, 2013.

[99] Remco Van Der Hofstad. *Random Graphs and Complex Networks.* Cambridge University Press, 2016.

[100] Andrew Vince. Star chromatic number. *Journal of Graph Theory*, 12(4):551–559, 1988.

[101] Walter Wallis and Julin Wu. On clique partitions of split graphs. *Discrete Mathematics*, 92(1-3):427–429, 1991.

[102] Duncan Watts and Steven Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998.

[103] Roger Yeh. The edge span of distance two labellings of graphs. *Taiwanese Journal of Mathematics*, 4(4):675–683, 2000.

[104] Roger Yeh. A survey on labeling graphs with a condition at distance two. *Discrete Mathematics*, 306(12):1217–1231, 2006.

[105] Jean-Gabriel Young, Guillaume St-Onge, Edward Laurence, Charles Murphy, Laurent Hébert-Dufresne, and Patrick Desrosiers. Phase transition in the recoverability of network history. *Physical Review X*, 9(4):041056, 2019.

[106] Jianzhi Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, 2003.

[107] Yongqiang Zhao, Wenjie He, and Rongrong Cao. The edge span of T-coloring on graph C_n^d. *Applied mathematics letters*, 19(7):647–651, 2006.

[108] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343, 1977.

[109] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, 24(5):530–536, 1978.

[110] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(1):103–128, 2007.

Krzysztof Turowski

Elektronicznie
podpisany przez
Krzysztof Turowski
Data: 2022.09.27
21:30:54 +02'00'