Autoreferat

dr inż. Krzysztof Turowski

27 września 2022

1. Posiadane tytuły zawodowe i stopnie naukowe

- **Magister telekomunikacji**
  Przyznany w 2010 roku na podstawie pracy magisterskiej *„Projekt akustyczny wnętrza w oparciu o systemy modelowania akustycznego"* pod kierunkiem prof. dr. hab. inż. Bożeny Kostek na Wydziale Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej na kierunku telekomunikacja, specjalność: inżynieria dźwięku i obrazu.
- **Magister informatyki**
  Przyznany w 2011 roku na podstawie pracy magisterskiej *„Szkieletowe kolorowanie grafów"* pod kierunkiem prof. dr. hab. inż. Marka Kubalego na Wydziale Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej na kierunku informatyka, specjalność: technologie internetowe i algorytmy.
- **Magister filozofii**
  Przyznany w 2016 roku na podstawie pracy magisterskiej *„Wartości i wolna wola w etyce Nicolaia Hartmanna"* pod kierunkiem prof. dr. hab. Stanisława Judyckiego na Wydziale Nauk Społecznych Uniwersytetu Gdańskiego na kierunku filozofia.
- **Doktor nauk technicznych w zakresie informatyki**
  Przyznany 16 czerwca 2015 r. przez Radę Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej na podstawie rozprawy doktorskiej pt. *„Analiza właściwości algorytmicznych problemu szkieletowego kolorowania grafów"* pod kierunkiem prof. dr. hab. inż. Marka Kubalego.

2. Informacje o dotychczasowym zatrudnieniu

- Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska, ul. Narutowicza 11/12, 80-233 Gdańsk, Polska:
  - od 1 listopada 2010 do 30 września 2011 – asystent,
  - od 1 października 2011 do 30 września 2015 – wykładowca,
  - od 1 października 2015 do 29 lutego 2016 – adiunkt.
- Center for Science of Information, Purdue University, 250 N University St, West Lafayette, Indiana 47907, Stany Zjednoczone
  - od 1 sierpnia 2015 do 30 listopada 2016 – staż podoktorski.
  - od 6 maja 2018 do 30 września 2019 – staż podoktorski.
- Google Poland Sp. z.o.o., ul. Emilii Plater 53, 00-113 Warszawa, Polska:
  - od 10 marca 2016 do 30 kwietnia 2018 – inżynier oprogramowania, Google Cloud Platform/Compute Engine.
- Wydział Matematyki i Informatyki, Uniwersytet Jagielloński, ul. Łojasiewicza 6, 30-348 Kraków, Polska:
  - od 1 października 2019 (do dziś) – adiunkt.

3. Wskazanie osiągnięć naukowych podlegających ocenie

3.1. Tytuł osiągnięcia naukowego

Wybranym osiągnięciem naukowym jest cykl publikacji pt. Analiza strukturalna i kompresja dla duplikacyjnych modeli grafów losowych.

3.2. Wykaz publikacji dokumentujących osiągnięcie

[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, s. 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, s. 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, s. 16:1-16:14.

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

- wersja konferencyjna: 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, s. 1089-1096.

4. Omówienie celu naukowego prac i osiągniętych wyników w ramach wskazanego osiągnięcia

4.1. Wstęp

Teoria grafów znalazła zastosowanie w opisie wielu złożonych układów występujących w świecie, takich jak sieci biologiczne czy sieci społecznościowe. Takie podejście wydaje się naturalne, gdy zależy nam na opisie systemu z użyciem pewnych podstawowych elementów (reprezentowanych przez wierzchołki grafu) oraz ich interakcji (reprezentowanych przez krawędzie grafu). Pozwala to na analizę tych struktur zarówno w ujęciu statycznym, czyli opisie ich własności w pewnym ustalonym stanie, jak i dynamicznym, czyli ujęciu ewolucji sieci oraz zmian jej parametrów w czasie.

Zagadnienia analizy strukturalnej i kompresji grafów losowych są naturalnym rozwinięciem analiz dokonywanych w klasycznej teorii grafów i teorii informacji. Po pierwsze, wyrastają one z badań nad grafami losowymi i ich własnościami zapoczątkowanym przez Paula Erdősa i Alfréda Rényiego [40] w 1959 roku. Ten kierunek prac, kontynuowany przez kolejne dekady, zaowocował szeregiem wyników dotyczących popularnych parametrów grafowych, takich jak rozmiar spójnych składowych, maksymalny stopień grafu, rozmiar maksymalnego skojarzenia, czy liczba chromatyczna (podsumowanie ważniejszych wyników można znaleźć w monografiach [16, 48, 67, 99]).

Po drugie, łączą się one również bezpośrednio z poszukiwaniami odpowiednich modeli grafów losowych, które można dopasować do występujących w świecie struktur. Przykładowo, już od lat 70. XX w. pojawiały się hipotezy, że dynamikę ewolucji sieci interakcji protein można opisać za pomocą prostych reguł duplikacji i mutacji [86, 106]. Co więcej, w ramach badania własności rzeczywistych sieci różnego pochodzenia zwrócono uwagę na szereg ich charakterystycznych cech:

- krótkie ścieżki między dowolnymi wierzchołkami – np. zjawisko *Six Degrees of Separation* w sieciach społecznościowych, spopularyzowane przez socjologa Stanleya Milgrama,
- występowanie lokalnych klastrów – również intuicyjnie dostrzegane w sieciach społecznościowych występowanie mniejszych grup z kompletem połączeń wewnętrznych,
- rozkład stopni wierzchołków – istnienie niewielu wierzchołków dużego stopnia (tzw. hubów) oraz wielu o niewielkim stopniu.

Zainteresowanie badaczy wzbudziły przede wszystkim sieci bezskalowe, formalnie zdefiniowane jako takie, w których odsetek wierzchołków stopni k dąży do ustalonej wartości, proporcjonalnej do $k^{-\gamma}$ dla pewnego ustalonego parametru γ. Popularność badań nad sieciami bezskalowymi były dodatkowo motywowane przez stwierdzenia, że taka charakterystyka bardzo dobrze pasuje do wielu rodzajów sieci biologicznych czy społecznościowych np. interakcji protein, cytowań artykułów naukowych czy hiperłączy internetowych [2, 30]. W ramach teoretycznych poszukiwań sposobów generowania grafów z takimi własnościami powstawały kolejne modele, najbardziej znane autorstwa Barabásiego i Alberta [8] czy Wattsa i Strogattza [102]. Również one stawały się przedmiotami dokładnej analizy probabilistyczno-teoriografowej [48, 85, 99]. Przykładowo, dowiedziono, że grafy generowane w modelu Barabásiego-Alberta rzeczywiście charakteryzują się własnością bezskalowości z parametrem $\gamma = 3$ [17].

W ramach teorii informacji z kolei narastała od pewnego czasu potrzeba wyjścia poza ujęcie zaproponowane przez Shannona w jego pracy „A Mathematical Theory of Communication" z 1948 roku. W tekście programowym „Three Great Challenges for Half Century Old Computer Science" Frederick P. Brooks Jr. zauważył, że wyzwaniem na przyszłość jest zastosowanie klasycznej definicji i miary informacji do analiz strukturalnych [22]. Naturalnym kierunkiem badań wydaje się zastosowane tego podejścia do danych ujętych w postaci grafów i poszukiwania liczbowego ujęcia informacji tworzonej i wyrażanej poprzez powstawanie i ewolucję takich struktur [96].

Powiązana jest z tym motywacja czysto praktyczna: w epoce Big Data jednym z rodzajów przechowywanych i przetwarzanych danych wielkoskalowych są dane grafowe np. sieci społecznościowych. Znalezienie skutecznego sposobu ich oszczędnej reprezentacji mogłoby przynieść wymierne praktyczne korzyści. Nawet opracowanie uniwersalnych, asymptotycznie optymalnych algorytmów kompresji, takich jak algorytmy Lempela-Ziva [108, 109], wcale nie zakończyło badań, ponieważ nadal istnieje dostatecznie dużo miejsca na praktyczne ulepszenia za cenę przyjęcia pewnych dodatkowych założeń o charakterze źródła danych.

W ramach podejścia nazwanego „Shannon spotyka Turinga" [96] można badać przede wszystkim entropię grafów (dla wierzchołków z etykietami) i struktur (dla nieodróżnialnych wierzchołków) zgodnie z definicją Shannona nad przestrzenią probabilistyczną grafów wygenerowanych z określonego modelu. Uzyskane w ten sposób wyniki stanowią dolne oszacowanie możliwości kompresji grafów generowanych zgodnie z przyjętym modelem grafów losowych. Drugą częścią tych badań jest poszukiwanie algorytmów, które potrafiłyby osiągnąć odpowiedni współczynnik kompresji z jak najlepszymi gwarancjami dokładności.

W toku badań okazało się, że istnieje doniosły związek między twierdzeniami o strukturze grafów losowych a twierdzeniami o charakterystykach informacyjnych ich modeli. Przede wszystkim wiedza o rozkładzie liczby automorfizmów – ściślej: wiedza o tym, że wygenerowane grafy są z dużym prawdopodobieństwem asymetryczne – dla modeli grafów losowych Erdősa-Renyíego i Barabásiego-Alberta okazała się podstawą do wyznaczenia wartości entropii oraz entropii strukturalnej dla obu modeli [28, 80]. Dla wielu innych modeli grafów udało się z kolei uzyskać pewne asymptotyczne dolne oszacowania entropii oparte m.in. o dowody probabilistycznego zachowania maksymalnego stopnia grafu lub innych prostych struktur grafowych [27].

W ramach opracowywania sposobów bezstratnej kompresji oprócz algorytmów uniwersalnych i tworzenia skutecznych w praktyce heurystyk (zob. przegląd algorytmów w [10]) ważnym zadaniem jest poszukiwanie algorytmów dających pewne średnie i pesymistyczne gwarancje dla konkretnych modeli grafów losowych. W szczególności dla wspomnianych wyżej modeli Erdősa-Renyíego i Barabásiego-Alberta opracowano algorytmy, które osiągają kompresję zgodną z entropią z dokładnością do pierwszych dwóch czynników wiodących [28, 80].

Duża część zainteresowania badaczy skupiła się na dwóch najpopularniejszych modelach grafów losowych: Erdősa-Renyíego i Barabásiego-Alberta. Jednak mimo ich prostych definicji oraz bardzo ciekawych właściwości strukturalnych wydaje się, że nie opisują one trafnie wielu rodzajów rzeczywistych sieci [33, 92]. W szczególności należy zwrócić uwagę, że np. dla sieci biologicznych często

identyfikuje się występowanie cechy bezskalowości ze współczynnikiem $\gamma < 2$, co nie pasuje do żadnego z powyższych modeli [30]. Co więcej, wydaje się, że dopasowanie tych modeli do istniejących sieci nie mogłoby zostać poparte racjami na rzecz takiego a nie innego mechanizmu powstawania sieci.

Intuicja, zgodnie z którą sieci biologiczne i społecznościowe mogą powstawać wskutek wewnętrznego mechanizmu duplikacji i mutacji (*duplication-divergence*), stanowi kierunek badań, który pozwala odpowiedzieć na te wyzwania. Z jednej strony, motywacja ewolucyjna takich modeli duplikacyjnych jest dobrze zakorzeniona w rozważaniach biologicznych [106]. Również w porównaniach różnych modeli z istniejącymi sieciami np. interakcji protein lub cytowań okazuje się, że modele duplikacyjne wypadają najlepiej jeśli chodzi o dopasowanie stopni wierzchołków w grafie [33] czy rozkładu małych podgrafów [92]. Z drugiej strony, nierygorystyczne obliczenia dla takich modeli pokazywały, że dla pewnych parametrów mogą one generować grafy bezskalowe z odpowiednimi współczynnikami [61, 85, 89], co czyni je równie interesującymi dla teoretyków ceniących właśnie te cechy[1].

Właśnie ta problematyka stała się przedmiotem badań w ramach cyklu publikacji [A1]-[A5] składającego się na przedstawiane osiągnięcie. W poszczególnych pracach rozważane było probabilistyczne zachowanie modeli duplikacyjnych, w szczególności najogólniejszego modelu autorstwa Solégo i Pastora-Satorrasa [87, 93], będącego popularnym modelem stanowiącym zarazem uogólnienie innego ważnego modelu tzw. czystej duplikacji [30]. Badano dokładne i asymptotyczne zachowanie różnych zmiennych losowych, takich jak średni stopień grafu, stopień ustalonego wierzchołka, maksymalny stopień w grafie, czy współczynnik skali γ. Podjęto również problem entropii grafowej i kompresji dla pewnego szczególnego przypadku tego modelu. Dla osiągnięcia tego celu zastosowano szereg podejść kombinatorycznych, probabilistycznych i algorytmicznych, m.in. narzędzia wypracowane w dziedzinie kombinatoryki analitycznej, pozwalające rozwiązać zależności rekurencyjne oraz wyprowadzić z nich twierdzenia o asymptotycznym zachowaniu zmiennych.

4.2. Podstawowe definicje

W tym miejscu i w dalszej części autoreferatu przyjmujemy standardowe oznaczenia teoriografowe np. za [37]: $V(G)$ i $E(G)$ oznaczają zbiory wierzchołków i krawędzi grafu G, $\mathcal{N}_G(u)$ – zbiór sąsiadów wierzchołka u w G, $\deg_G(u) = |\mathcal{N}_G(u)|$ – stopień wierzchołka u w G. Wszystkie rozpatrywane grafy są proste, bez pętli i wielokrotnych krawędzi. Ponieważ często wykorzystywana jest notacja G_t na oznaczenie grafu losowego na t wierzchołkach, dla zwięzłości stosujemy zapis $\deg_t(u)$ zamiast $\deg_{G_t}(u)$, $\mathcal{N}_t(u)$ zamiast $\mathcal{N}_{G_t}(u)$ itp.

Podstawowymi parametrami są *średni stopień* $D(G)$ grafu G, zdefiniowany jako

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

oraz *średni kwadrat stopnia* $D_2(G)$ czyli

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

Każdy model grafów losowych wyznacza pewien rozkład prawdopodobieństwa na zbiorach wszystkich grafów o ustalonym zbiorze wierzchołków. Ponieważ dla zdecydowanej większości grafów losowych liczba wierzchołków należy do wejścia modelu, naturalne jest definiowanie zmiennych losowych nad zbiorem wszystkich grafów o zadanej liczbie wierzchołków t, oznaczanym $\mathcal{G}_t$. Przykładowo, średni stopień i średni kwadrat stopnia dla modeli grafów losowych będą zdefiniowane odpowiednio jako:

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

[1] Należy jednak pamiętać, że obliczenia te nie były dostatecznie ścisłe, przez co przynajmniej część z nich została później obalona i poprawiona np. w [57].

W literaturze zmienne $D(G_t)$ i $D_2(G_t)$ spotykane są również pod nazwą odpowiednio pierwszego i drugiego momentu rozkładu stopni. Analogiczne definicje zmiennych losowych można również sformułować dla dowolnego innego znanego parametru grafowego np. dla stopnia ustalonego wierzchołka $\deg_t(s)$ czy dla maksymalnego stopnia grafu $\Delta(G_t)$.

Innym ważnym parametrem jest liczba oraz odsetek wierzchołków danego stopnia, zdefiniowane następująco:

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

W podobny sposób możemy zdefiniować również entropię grafu zgodnie z klasyczną definicją teorioinformacyjną dla dowolnego dyskretnego rozkładu prawdopodobieństwa:

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

W tej definicji zakładamy, że $\mathcal{G}_t$ jest zbiorem wszystkich parami nieizomorficznych grafów o zbiorze wierzchołków $\{1, \ldots, t\}$ tj. każdy wierzchołek ma przypisaną unikalną etykietę[2]. Odpowiednio, można zdefiniować entropię strukturalną tj. wartość

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

gdzie $\mathcal{S}_t$ jest zbiorem wszystkich nieizomorficznych grafów o t wierzchołkach.

Formalna definicja modelu duplikacyjnego Solégo i Pastora-Satorrasa $\texttt{DD}(t, p, r)$ [87, 93], będącego głównym przedmiotem badań w pracach [A1]-[A5], jest następująca: dla danych parametrów $0 \leq p \leq 1$ and $0 \leq r \leq t_0$ i (często niejawnie[3]) danego grafu początkowego G_{t_0} z $V(G_{t_0}) = \{1, \ldots, t_0\}$ dla każdego $t = t_0, t_0 + 1, \ldots$ tworzymy G_{t+1} z G_t zgodnie z następującą procedurą:

1. dodajemy nowy wierzchołek $t + 1$ do grafu,
2. wybieramy losowo wierzchołek u równomiernie ze zbioru $V(G_t) = \{1, \ldots, t\}$ i oznaczamy u jako $\mathit{parent}(t + 1)$,
3. dla każdego wierzchołka $i \in V(G_t)$:
   (a) jeśli $i \in \mathcal{N}_t(\mathit{parent}(t + 1))$, to dodajemy krawędź pomiędzy i a $t + 1$ z prawdopodobieństwem p,
   (b) jeśli $i \notin \mathcal{N}_t(\mathit{parent}(t + 1))$, to dodajemy krawędź pomiędzy i a $t + 1$ z prawdopodobieństwem $\frac{r}{t}$.

Wszystkie losowania tj. wierzchołka-rodzica ze zbioru istniejących wierzchołków oraz wszystkich decyzji o dodawaniu krawędzi między nowym wierzchołkiem a istniejącymi wierzchołkami są niezależne.

Szczególnym przypadkiem tego modelu, rozpatrywanym w literaturze [11, 29, 30], jest tzw. model czystej duplikacji (*pure duplication*). Zawiera on tylko krawędzie powstałe w wyniku duplikacji wierzchołków, ale nie ma w nim krawędzi łączących nowe wierzchołki z istniejącymi wierzchołkami spoza sąsiedztwa ich rodziców tj. jest to model $\texttt{DD}(t, p, 0)$.

[2] Zwróćmy uwagę, że dla parametrów takich jak D czy Δ nie ma znaczenia, czy operujemy na zbiorze grafów z etykietami, czy bez etykiet.

[3] Typowo zakłada się, że G_{t_0} jest grafem pełnym, co jest wyborem tyleż wygodnym z teoretycznego punktu widzenia, co różnie ocenianym z punktu widzenia zastosowań. Zob. również [59].

4.3. Analiza zachowania rozkładu stopni dla duplikacyjnych modeli grafów losowych

Punktem wyjścia do prac nad duplikacyjnymi modelami grafów losowych, w szczególności modelem autorstwa Solégo i Pastora-Satorrasa była obserwacja poczyniona dla wielu innych modeli grafów losowych w co najmniej dwóch niezależnych pracach [27, 80], zgodnie z którą zachowanie zmiennych losowych $\deg_t(s)$ i $\Delta(G_t)$ stanowiło podstawę opracowania oszacowań entropii dla poszczególnych modeli. Ponieważ jednak sam model Solégo i Pastora-Satorrasa stanowił głównie obiekt badań na bazie symulacji komputerowych oraz obliczeń przybliżeń, konieczne było opracowanie rygorystycznej teorii zachowania poszczególnych zmiennych w grafach w duchu klasycznego podejścia teoriografowego np. Bollobása [17]. Praca [A1] zawiera wyniki opisujące zachowanie średniej i wariancji podstawowych zmiennych (średniego stopnia w grafie i stopnia ustalonego wierzchołka), natomiast prace [A2] i [A3] rozszerzają rezultaty na twierdzenia o koncentracji rozkładu tych zmiennych oraz maksymalnego stopnia w grafie wokół wartości średniej.

W pracy [A1] zawarte zostały wyniki badań nad właściwościami rozkładu dwóch zmiennych losowych zdefiniowanych dla modelu $\texttt{DD}(t,p,r)$: stopnia $\deg_t(s)$ danego wierzchołka s w G_t oraz średniego stopnia grafu $D(G_t)$. Pierwszym krokiem było szukanie pierwszych momentów odpowiednich zmiennych:

Problem [A1]: jakie jest dokładne asymptotyczne tempo wzrostu wartości funkcji $\mathbb{E}[\deg_t(s)]$, $\mathbb{E}[D(G_t)]$ oraz $\mathrm{Var}[\deg_t(s)]$ i $\mathrm{Var}[D(G_t)]$?

Punktem wyjścia dowodu była obserwacja, że dla odpowiednich zmiennych można skonstruować równania rekurencyjne. Przykładowo, dla wartości oczekiwanych z definicji modelu można wyprowadzić zależności:

$$\mathbb{E}[\deg_{t+1}(s)] = \mathbb{E}[\deg_t(s)]\left(1+\frac{p}{t}-\frac{r}{t^2}\right)+\frac{r}{t}, \tag{4.1}$$

$$\mathbb{E}[D(G_{t+1})] = \mathbb{E}[D(G_t)]\left(1+\frac{2p-1}{t+1}-\frac{2r}{t(t+1)}\right)+\frac{2r}{t+1}. \tag{4.2}$$

Następnie wykazano, że obie zmienne w tym modelu są blisko powiązane poprzez zależność wartości początkowej dla pierwszej rekurencji $\mathbb{E}[\deg_s(s)]$ od średniego stopnia grafu w poprzedniej iteracji $\mathbb{E}[D(G_{s-1})]$. Z definicji modelu mamy stopień ostatniego wierzchołka zdefiniowany jako

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

tj. $\deg_{t+1}(t+1)$ jest zmienną losowa będącą sumą dwóch zmiennych dwumianowych o odpowiednich parametrach. Wyprowadzona stąd została następująca zależność między wartościami oczekiwanymi obu zmiennych

$$\mathbb{E}[\deg_{t+1}(t+1)] = \left(p-\frac{r}{t}\right)\mathbb{E}[D(G_t)]+r. \tag{4.3}$$

Z zależności (4.3) wynika, że wyznaczenie tempa wzrostu $\mathbb{E}[D(G_t)]$ prowadzi wprost do tempa wzrostu dla warunku początkowego $\mathbb{E}[\deg_s(s)]$ dla równania (4.1) w przypadku gdy $s\to\infty$.

Techniki opracowane w pracy do rozwiązywania powyższych równań służą właściwie do wyznaczania asymptotycznego zachowania na podstawie ogólnych zależności rekurencyjnych postaci

$$\mathbb{E}[f(G_{n+1}) \mid G_n] = f(G_n)g_1(n)+g_2(n), \tag{4.4}$$

dla funkcji $g_1(n)=\frac{W_1(n)}{W_2(n)}$ wyrażalnej jako iloraz wielomianów W_1 i W_2 tego samego stopnia. Jak widać, równania (4.1) i (4.2) są szczególnymi przypadkami formy (4.4).

Po pierwsze, pokazano, że dla rozwiązania powyższego równania rekurencyjnego zapisanego w postaci

$$\mathbb{E}[f(G_n)] = \prod_{k=n_0}^{n-1} g_1(k)\left(f(G_{n_0})+\sum_{j=n_0}^{n-1} g_2(j)\prod_{k=n_0}^{j}\frac{1}{g_1(k)}\right) \tag{4.5}$$

można wykazać, że

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

gdzie Γ jest funkcją gamma Eulera, d jest stopniem wielomianów W_1 i W_2, natomiast a_i i b_i (dla $i = 1, \ldots, d$) są, odpowiednio, pierwiastkami (zespolonymi) wielomianów W_1 i W_2.

W połączeniu ze poniższym lematem określającym tempo wzrostu ilorazu funkcji Γ pozwoliło to na ustalenie asymptotycznego tempa wzrostu[4] pierwszego z czynników w rozwiązaniu rekurencji.

Lemat 4.1 (Abramowitz, Stegun [1])**.** *Dla dowolnych $a, b \in \mathbb{R}$ i $n \to \infty$ zachodzi asymptotycznie*

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

Kolejnym krokiem było wykorzystanie obserwacji, że jeśli funkcję $g_2(n)$ można przedstawić jako iloraz funkcji gamma Eulera, to aby znaleźć asymptotyczne tempo wzrostu formuł postaci (4.5) wystarczy znaleźć asymptotyczne rozwinięcie wyrażeń typu

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

Stosując podobne narzędzia do opisanych powyżej, włączając również teorię funkcji hipergeometrycznych (również obecną w [1]), można otrzymać następujące zależności:

Lemat 4.2. *Niech $a_i, b_i \in \mathbb{R}$ dla $i = 1, 2, \ldots, k$ ($k \in \mathbb{N}$) oraz $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$. Asymptotycznie gdy $n \to \infty$ zachodzi*

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

Lemat 4.3. *Niech $a_i, b_i \in \mathbb{R}$ dla $i = 1, 2, \ldots, k$ ($k \in \mathbb{N}$) oraz $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$. Dla każdego $n \in \mathbb{N}_+$ zachodzi*

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

gdzie $c_3 = p + \sqrt{p^2+2r}$, $c_4 = p - \sqrt{p^2+2r}$, natomiast ${}_pF_q[\begin{smallmatrix}\mathbf{a}\\ \mathbf{b}\end{smallmatrix}; z]$ jest uogólnioną funkcją hipergeometryczną (zob. [1]). Ponadto, asymptotycznie gdy $n \to \infty$ zachodzi

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

Dzięki powyższym lematom można znajdować asymptotyczne tempa wzrostu dla funkcji opisywanych przez równania rekurencyjne postaci (4.5), o ile funkcje $g_1(n)$ i $g_2(n)$ są pewnej szczególnej postaci. Ponieważ oba te warunki te spełnione są przez równanie (4.2) opisujące $D(G_t)$, to dowiedziono, że

Twierdzenie 4.1. *Asymptotycznie dla $t \to \infty$ zachodzi*

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

[4] Ściślej, pełna wersja poniższego lematu umożliwia ustalenie nie tylko głównego czynnika asymptotycznego tempa wzrostu, ale też rozwinięcia z dowolną żądaną dokładnością.

gdzie $D(G_{t_0})$ *to średni stopień grafu początkowego* G_{t_0} *oraz*

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

dla funkcji Pochhammera $(a)_l = a\ (a+1)\dots(a+l-1)$, $(a)_0 = 1$.

Stosując odpowiednie wzory dla $\deg_t(s)$ otrzymano, że

Twierdzenie 4.2. *Asymptotycznie dla* $t \to \infty$ *zachodzi*
(i) *gdy* $s = O(1)$

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

(ii) *gdy* $s = \omega(1)$ *i* $s = o(t)$

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

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

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

Ściślej, w pracy [A1] zostały przedstawione nie tylko oszacowania asymptotyczne, ale też dokładne wzory ze skomplikowanymi współczynnikami wiodącymi (zależnymi od s, p, r) dla poszczególnych przypadków.

Analogiczne dowody przeprowadzono również dla wariancji wyżej omawianych zmiennych otrzymując, że:

Twierdzenie 4.3. *Asymptotycznie dla* $t \to \infty$ *zachodzi*

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

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

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

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

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

Praca [A2] poszerza badania zaprezentowane w pracy [A1] o wyniki asymptotycznego tempa wzrostu wartości zmiennych losowych dla ogonów rozkładu zmiennych $D(G_t)$ oraz $\deg_t(s)$ (dla $s = O(1)$):

Problem [A2]: dla jakich funkcji $f_l(t)$, $f_h(t)$ (odpowiednio, $g_l(t)$, $g_h(t)$) zachodzi $\Pr[D(G_t) \notin [f_l(t), f_h(t)]] = O(t^{-A})$ (odpowiednio, $\Pr[\deg_t(s) \notin [g_l(t), g_h(t)]] = O(t^{-A})$) dla dowolnego $A > 0$?

Czy można pokazać, że powyższa zależność zachodzi dla funkcji $f_l(t), f_h(t) \in \widetilde{\Theta}(\mathbb{E}[D(G_t)])$ (odpowiednio, $g_l(t), g_h(t) \in \widetilde{\Theta}(\mathbb{E}[\deg_t(s)])$)?

Pierwszym głównym wynikiem pracy [A2] jest dowód dolnego ograniczenia:

Twierdzenie 4.4. *Asymptotycznie gdy $t \to \infty$ dla $G_t \sim \texttt{DD}(t, p, r)$ zachodzi*

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

dla pewnej stałej $C > 0$ oraz dla dowolnego $A > 0$.

Dowód tego twierdzenia, tak jak pozostałych wyników w pracy [A2], oparty jest o ostrożne szacowanie zachowania funkcji tworzącej momenty (*moment-generating function*). Uogólniając analizę modelu prowadzącą do równania (4.2) można pokazać, że dla zmiennej $D(G_t)$ zachodzi

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

Wynika z tego, że dla dowolnego ciągu parametrów $\lambda_t \to 0$ zachodzi

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

Dobierając odpowiednie wartości λ_k dla $k = t_0, \ldots, t-1, t$ i przyjmując $\varepsilon_t \geq \lambda_k$ dla wszystkich $k \leq t$ otrzymano

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

Wykorzystując nierówności Czernowa wynika z tego, że

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

co dla odpowiednio dobranych wartości ε_t, λ_t oraz α kończy dowód Twierdzenia 4.4.

Analogiczne podejście umożliwia pokazanie analogicznej zależności również dla zmiennej losowej $\deg_t(s)$:

Twierdzenie 4.5. *Asymptotycznie gdy $t \to \infty$ dla $G_t \sim \texttt{DD}(t, p, r)$ i $s = O(1)$ zachodzi*

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

dla pewnej stałej $C > 0$ oraz dla dowolnego $A > 0$.

Oszacowanie lewej strony rozkładu zmiennych $D(G_t)$ i $\deg_t(s)$ jest wykonywane podobnymi, nieco bardziej złożonymi technikami. Ostatecznie otrzymujemy, że:

Twierdzenie 4.6. *Asymptotycznie gdy $t \to \infty$ dla $G_t \sim \texttt{DD}(t,p,r)$ zachodzi*

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

dla pewnej stałej $C > 0$ oraz dla dowolnych $\varepsilon, A > 0$.

Analogiczne wyniki zostały opracowane dla zmiennej $\deg_t(s)$ z $s = O(1)$:

Twierdzenie 4.7. *Asymptotycznie gdy $t \to \infty$ dla $G_t \sim \texttt{DD}(t,p,r)$ i $s = O(1)$ zachodzi*

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

dla pewnej stałej $C > 0$ oraz dla dowolnych $\varepsilon, A > 0$.

Po dowodzie koncentracji zmiennych $D(G_t)$ i $\deg_t(s)$ wokół odpowiednich wartości średnich można zadać pytanie o analogiczne zachowanie maksymalnego stopnia grafu $\Delta(G_t)$ – i odpowiedzi na to pytanie (dla $\frac{1}{2} < p \leq 1$) poświęcona jest praca [A3]. Rzecz jasna, Twierdzenie 4.7 wprost implikuje identyczne ograniczenie dolne dla zmiennej $\Delta(G_t)$, jednak podobny wynik z ograniczenia górnego (tj. Twierdzenia 4.5) nie przenosi się.

Problem [A3]: dla jakich funkcji $f_l(t)$ i $f_h(t)$ zachodzi $\Pr[\Delta(G_t) \notin [f_l(t), f_h(t)]] = O(t^{-A})$ dla dowolnego $A > 0$?

Głównym wynikiem pracy [A3] jest twierdzenie pokazujące koncentrację maksymalnego stopnia w tym modelu:

Twierdzenie 4.8. *Dla $\frac{1}{2} < p < 1$ asymptotycznie gdy $t \to \infty$ dla $G_t \sim \texttt{DD}(t,p,r)$ zachodzi*

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

dla dowolnych stałych $\varepsilon, A > 0$.

Dowód Twierdzenia 4.8 został podzielony na trzy części: osobno podano (a) dowód ograniczenia dolnego, (b) ograniczenia górnego dla wierzchołków wczesnych i (c) dla wierzchołków późniejszych.

Główna idea dowodu części (b) jest następująca: szukamy takich wartości $(t_i)_{i=0}^k$ oraz takiego ciągu $(X_{t_i})_{i=0}^k$, że

1. $\deg_{t_0}(s) \leq X_{t_0}$ zachodzi z dużym prawdopodobieństwem dla $1 \leq s \leq t_0$,
2. $\deg_{t_{i+1}}(s) - \deg_{t_i}(s) \leq X_{t_{i+1}} - X_{t_i}$ zachodzi z dużym prawdopodobieństwem dla każdego $i = 0, \ldots, k-1$,
3. $t_k \approx t$ oraz $X_{t_k} = \widetilde{O}(t^p)$.

Jak łatwo zauważyć, znalezienie odpowiednich ciągów $(t_i)_{i=0}^k$ i takiego ciągu $(X_{t_i})_{i=0}^k$ spełniających powyższe warunki gwarantuje nam, że z dużym prawdopodobieństwem zachodzi $\deg_t(s) = \widetilde{O}(t^p)$. Punktem wyjścia dowodu jest definicja form funkcyjnych ciągów dla parametrów p, α, β_i ($i = 0, 1, \ldots, k$) i ϕ:

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

Okazuje się, że dla tak zdefiniowanego ciągu istnieje proste ograniczenie dolne bliskie poszukiwanemu:

Lemat 4.4. *Załóżmy, że zachodzi $\phi \geq \log^2 t$, $\alpha \leq \sqrt{\phi}$ i $\beta_i \geq \alpha(p-\delta)$ dla pewnego $\delta \in [0, p)$. Wówczas asymptotycznie dla $t \to \infty$ dla wszystkich $i = 0, 1, \ldots, k$ zachodzi $X_{t_i} \geq t_i^{p-\delta}$.*

Powyższą zależność można dowieść indukcyjnie przez wykorzystanie odpowiednich nierówności z rozwinięć w szereg Taylora oraz definicji parametru α:

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

To dolne ograniczenie z Lematu 4.4 jest następnie wykorzystane w dowodzie silniejszego ograniczenia górnego:

Lemat 4.5. *Załóżmy, że zachodzi $\phi \ge \log^3 t$, $\alpha(p-\delta) \le \beta_i \le \alpha p + \frac{\alpha}{2\log t_i}$ dla pewnego $\delta \in [0,p)$. Wówczas asymptotycznie dla $t \to \infty$ zachodzi $X_{t_i} \le \phi^{1-p} t_i^p \log t_i$ dla wszystkich $i = 0,1,\ldots,k$.*

W szczególności Lemat 4.5 gwarantuje, że ciągi $(t_i)_{i=0}^k$ oraz $(X_{t_i})_{i=0}^k$ zachowują się asymptotycznie zgodnie z naszymi wymaganiami tj. w szczególności $X_{t_i} = O(t_i^p \operatorname{polylog}(t))$, o ile zapewnimy, że $\phi = O(\operatorname{polylog}(t))$.

Drugi warunek tj. zachodzenie nierówności $\deg_{t_{i+1}}(s) - \deg_{t_i}(s) \le X_{t_{i+1}} - X_{t_i}$ z dużym prawdopodobieństwem, osiągany jest z odpowiednio dobranego ograniczenia Czernowa na zwiększanie się stopnia wierzchołka w czasie względem kolejnych wartości ciągu X_τ:

Lemat 4.6. *Niech $1 \le s \le \tau \le t$. Niech $X_\tau \ge 0$, $\varepsilon \in (0,1)$ będą wartościami takimi, że dla dowolnego $A > 0$, zachodzi $\deg_\tau(s) \le X_\tau$ oraz $3A\tau \log t \le \varepsilon^3 X_\tau(pX_\tau + r)$. Wówczas dla dowolnego $h \in \left[\frac{3A\tau\log t}{\varepsilon^2(pX_\tau+r)}, \varepsilon X_\tau\right]$ zachodzi*

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

Dobierając dokładne wartości $\alpha = \Theta(\log^2 t)$, $\beta_i = \alpha p + \frac{\alpha}{2\log t_i}$ i $\phi = \Theta(\log^4 t)$ można pokazać, że założenia obu ostatnich lematów będą spełnione. Pierwszy warunek, o zachodzeniu $\deg_{t_0}(s) \le X_{t_0}$, jest spełniony w sposób oczywisty z samej wartości $X_{t_0} = t_0$, a zatem dla każdego wierzchołka $1 \le s \le t_0 = \phi$ z dużym prawdopodobieństwem stopień w każdym momencie t_i nie przekracza X_{t_i}.

Dowód części (c) głównego twierdzenia tj. dla wierzchołków $s \in [t_i, t_{i+1}]$ dla kolejnych $i = 0,1,\ldots,k-1$ polega na wykorzystaniu obserwacji, że gdy $p < 1$, to z dużym prawdopodobieństwem stopień żadnego z poprzednich wierzchołków nie przekracza $X_{t_{i+1}}$, a zatem $\deg_s(s)$ nie przekracza $(1+\varepsilon)(pX_{t_{i+1}})$. Co więcej, można dowieść, że stopień wierzchołka $\deg_{t_{i+1}}(s)$ z dużym prawdopodobieństwem nie przekracza $X_{t_{i+1}}$ – ponieważ $\deg_s(s)$ jest dostatecznie mały a przyrost stopnia między momentem s a t_{i+1} nie może tego nadrobić. Ostatecznie, okazuje się, że wszystkie wierzchołki z danego przedziału spełniają z dużym prawdopodobieństwem zależność $\deg_s(s) \le X_{t_{i+1}}$, a więc można zastosować dla nich lemat 4.6 aby otrzymać ograniczenie na ich stopnień w momencie t.

W dowodzie dolnego ograniczenia (tj. części (a)) dla Twierdzenia 4.8 również używamy ciągów $(t_i)_{i=0}^k$ oraz $(X_{t_i})_{i=0}^k$ zdefiniowanych powyżej. Również w tym przypadku można dowieść odpowiednią nierówność typu Czernowa:

Lemat 4.7. *Niech $1 \le s \le \tau \le t$. Niech $X_\tau \ge 0$, $\varepsilon \in \left(0, \frac{1}{3}\right)$ będą wartościami takimi, że dla dowolnego $A > 0$ zachodzi $\deg_\tau(s) \le \tau$ and $3A\log t \le \varepsilon^3 pX_\tau$. Wówczas dla dowolnego $h \in \left[\frac{3A\log t}{\varepsilon^2 pX_\tau}, \varepsilon\tau\right]$ zachodzi*

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

Przyjmując odpowiednie wartości α, β_i oraz ϕ można dowieść, że ograniczenie $\deg_{t_i}(s) \ge X_{t_i} - \phi + 1$ zachodzi z dużym prawdopodobieństwem dla kolejnych $i = 0,1,\ldots,k$.

Problemy otwarte Ostatecznie, badania w pracach [A1]-[A3] pozwoliły pokazać, że zarówno rozkłady stopni poszczególnych wierzchołków, jak i maksymalnego stopnia w grafie są skoncentrowane wokół swoich wartości średnich. Można postawić pytanie, czy istniejące oszacowania są najlepszymi tj. np. czy funkcje $f_l(t)$ i $f_h(t)$ są najlepsze możliwe. Przykładowo, można próbować dowieść wyników analogicznych jak dla modelu Barabásiego-Alberta, gdzie pokazano w [46], że z dużym prawdopodobieństwem zachodzi $\Delta(G_t) \in [\sqrt{t}/f(t), \sqrt{t}f(t)]$ dla pewnej wolno rosnącej funkcji $f(t)$ – ale nie jest to prawdą dla żadnej funkcji stałej.

Z pewnością ważnym dalszym krokiem w analizie strukturalnej tego modelu grafów losowych, zwłaszcza z punktu widzenia kompresji takich grafów, może być zidentyfikowanie asymptotycznego pełnego rozkładu stopni w grafie tj. rozkładu zmiennych $F_k(G_t)$ oraz $f_k(G_t)$ dla wszystkich $k = 0, 1, \ldots$.

Innym problemem w ramach badania strukturalnych własności duplikacyjnych grafów losowych, pobocznym z punktu widzenia prac [A1]-[A3], ale interesującym społeczność w przypadku innych modeli [16, 99] może być poszukiwanie asymptotycznego zachowania takich parametrów grafu jak liczba klikowa, rozmiar największego zbioru niezależnego, rozmiar największego skojarzenia w grafie, czy też liczba chromatyczna grafu.

4.4. Badania nad parametrem skali dla duplikacyjnych modeli grafów losowych

Przekonanie o zachodzeniu własności bezskalowości dla grafów generowanych z modeli duplikacyjnych było jednym z istotnych bodźców do prac nad tymi modelami. W szczególności twierdzono m.in. w [30], że różne sieci biologiczne i społeczne charakteryzują się odpowiednio współczynnikami skali z zakresu $(1, 2)$ i $(2, 3)$, co miałoby być zgodne z szacunkowym współczynnikiem skali dla grafów generowanych przez model duplikacyjny $\texttt{DD}(t, p, 0)$ – w odróżnieniu np. od modelu Erdősa-Renyíego (dla którego własność bezskalowości nie zachodzi) lub Barabásiego-Alberta (dla którego $\gamma = 3$) [17].

Bardziej szczegółowe badania w [57] wykazały, że w grafach generowanych z modelu $\texttt{DD}(t, p, 0)$ zachodzi

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

dla pewnej stałej $c(p) \in (0, 1]$ (w szczególności $c(p) = 1$ dla $p < 0.5671\ldots$). To oznacza, że asymptotycznie prawie wszystkie wierzchołki w grafach generowanych w danym modelu są izolowane i nie może zachodzić $\lim_{k\to\infty} \lim_{t\to\infty} f_k(G_t) \sim k^{-\gamma}$.

Tak przeważająca obecność wierzchołków izolowanych w tych grafach, rzadko spotykana w rzeczywistych sieciach, zasugerowała badaczom przedefiniowanie problemu poprzez ograniczenie się do badań (jedynej) spójnej składowej takiego grafu tj. do pominięcia wierzchołków izolowanych:

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

W [68] podjęto formalne rozważania nad tak zdefiniowanym problemem. W szczególności wyprowadzono z zależności rekurencyjnej równanie dla funkcji tworzącej $A(z) = \sum_{k=0}^{\infty} a_k z^k$:

$$A(pz + 1 - p) = 2pA(z) + pz\ (1 - z)A'(z) + A(1 - p), \tag{4.6}$$

a następnie dowiedziono następujące twierdzenie

Twierdzenie 4.9 ([68, Twierdzenie 2.1(3)]). *Dla $0 < p < \exp(-1)$ niech $\beta(p) > 2$ będzie rozwiązaniem równania $p^{\beta-2} + \beta - 3 = 0$. Wówczas dla asymptotycznego rozkładu stopni spójnej składowej $(a_k)_{k=0}^{\infty}$ w modelu $\texttt{DD}(t, p, 0)$ zachodzi dla $k \to \infty$, że*

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

Praca [A4] poświęcona jest poprawieniu wyniku z Twierdzenia 4.9, w szczególności pokazania zachowania rozkładu dla brakującego przypadku gdy $q = \beta(p)$:

Problem [A4]: czy współczynnik $\beta(p)$ jest współczynnikiem skali w grafach wygenerowanych jako spójne składowe modelu $\mathtt{DD}(t, p, 0)$? Jeśli tak, to do jakiej liczby rzeczywistej dąży iloraz $\frac{a_k}{k^{\beta(p)}}$ gdy $k \to \infty$?

Twierdzenie 4.10. *Dla $0 < p < \exp(-1)$ niech $\beta(p) > 2$ będzie rozwiązaniem równania $p^{\beta-2} + \beta - 3 = 0$. Wówczas dla asymptotycznego rozkładu stopni spójnej składowej $(a_k)_{k=0}^{\infty}$ w modelu $\mathtt{DD}(t, p, 0)$ zachodzi dla $k \to \infty$, że*

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

dla funkcji gamma Eulera $\Gamma(s)$ oraz

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

Dowód Twierdzenia 4.10 polegał na przekształceniu poprzez odpowiednie podstawienia w równaniu (4.6) w celu otrzymania równania

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

z warunkami brzegowymi $C(1) = A(0) = 0$ and $\lim_{w\to\infty} C(w) = A(1) = 1$.

Ponieważ nie jest znane zachowanie $C(w)$ dla $w \to 0$, nie można wprost wyznaczyć pasa fundamentalnego (*fundamental strip*) funkcji C i zastosować transformacji Mellina, dobrze znanego narzędzia analitycznego służącego do obliczania rozwinięć asymptotycznych [45, 95]. Zamiast tego zastosowana została procedura odwrotna: najpierw zdefiniowano funkcję

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

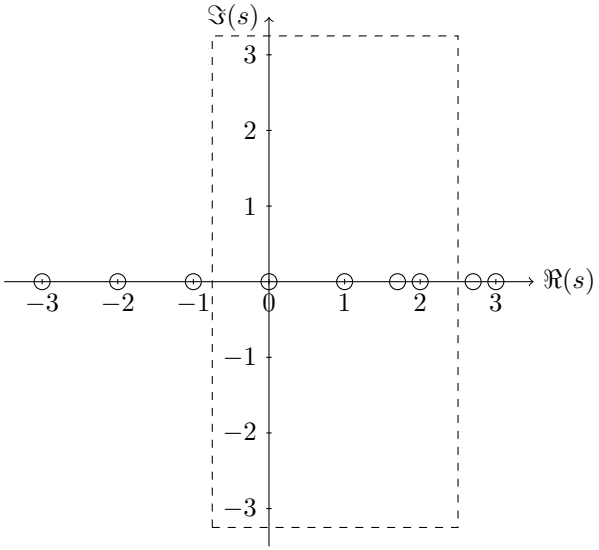
taką, że dla funkcji spełniającej równanie $E(\frac{w}{p}) = 2pE(w) + p(w-1)E'(w) + K$ dla pewnego stałego K zachodzi następujący lemat:

Lemat 4.8. *Funkcja $E^*(s)$ jest transformatą Mellina funkcji $E(w)$ z pasem fundamentalnym $\{s\colon \Re(s) \in (-1, 0)\}$.*

Do obliczenia asymptotycznego rozwinięcia funkcji $E(w)$ w szereg względem zmiennej w zastosowano standardowe podejście (zob. [95]), zgodnie z którym całkę po pewnej prostej pionowej w pasie fundamentalnym $\Re(s) \in (-1, 0)$ zamieniamy na całkę po odpowiednim prostokącie (zob. Rysunek 1) korzystając z tego, że przy pewnych własnościach funkcji E górny i dolny bok prostokąta wnoszą pomijalne małe wielkości (co wynika z tzw. *smallness property* transformaty Mellina) a prawy bok prostokąta dostarcza tylko pewien czynnik asymptotyczny proporcjonalny do w^{-M}.

Zgodnie z twierdzeniem Cauchy'ego o residuach można zamienić tę całkę na sumę wartości odpowiednich residuów. Po bliższym przyjrzeniu się funkcji $E^*(s)$ dla $0 < p < \exp(-1)$ można zauważyć, że ma ona tylko 3 typy prostych, izolowanych biegunów:

- dla $s = 0, -1, -2, \ldots$, wprowadzonych przez funkcję $\Gamma(s)$,
- dla $s = 1, 2, 3, \ldots$, wprowadzonych przez funkcję $\frac{1}{D(s)}$,
- dla $s = s^* + 1, s^* + 2, s^* + 3, \ldots$, wprowadzonych przez funkcję $\frac{1}{D(s)}$,



Rysunek 1: Przykładowy obszar całkowania dla $E^*(s)$ i $E(w)$ z $s^* = 0.7$ i $M = 2.5$.

gdzie s^* jest nietrywialnym (tj. różnym od zera) rozwiązaniem równania $p^s + s - 1 = 0$. Obliczenie residuów dla biegunów znajdujących się wewnątrz prostokąta pozwala zatem to na obliczenie rozwinięcia $E(w)$ w szereg względem w z dokładnością do $O(w^{-M})$ dla dowolnego $M > 0$. Przekształcenie asymptotycznego rozwinięcia $E(w)$ na $C(w)$ pozostaje kwestią prostego liniowego skalowania.

Na koniec wystarczy jedynie zauważyć, że rozwinięcie $C(w)$ w szereg $w^{-\alpha}$ jest równoważne rozwinięciu $A(z)$ w szereg $(1-z)^\alpha$ – a tempo wzrostu tego ostatniego jest determinowane przez wyraz z najmniejszym $\alpha \in \mathbb{R}_+ \setminus \mathbb{N}$, ponieważ z klasycznych twierdzeń Flajoleta i Odlyzki o przenoszeniu tempa wzrostu z [44] wiemy, że

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

Słowem, ponieważ s^*+1 jest najmniejszym niecałkowitym biegunem funkcji $E^*(s)$, to $a_k = [z^k]A(z)$ będzie asymptotycznie proporcjonalne do k^{-s^*-2} z odpowiednią znaną stałą.

Ponieważ dla $p \in (0, \exp(-1))$ wiadomo, że $s^* \in (0,1)$, to można powiedzieć również, że głównym wynikiem pracy [A4] jest wykazanie, że dla pewnych wartości parametru p model $\mathtt{DD}(t,p,0)$ generuje grafy bezskalowe ze współczynnikiem z zakresu $(2,3)$, a więc pasującym do oszacowań dla rzeczywistych sieci pewnych typów [30].

Problemy otwarte Naturalnie narzucającym się dalszym problemem jest rozstrzygnięcie, czy grafy generowane według tego modelu wykazują własność bezskalowości również dla przypadku $p \geq \exp(-1)$. Możliwość istnienia przejścia fazowego została zasugerowana już w [68], gdzie pokazano metodami opartymi o nieskończone łańcuchy Markowa, że odpowiedni łańcuch sprzężony z oryginalnym procesem generowania grafu jest chwilowy (*transient*), a zatem nie ma rozkładu stacjonarnego – co może sugerować, że to samo zachodzi również dla samego modelu $\mathtt{DD}(t,p,0)$.

Podobnym kierunkiem badań byłoby poszukiwanie odpowiedzi na to samo pytanie dla ogólniejszego modelu $\mathtt{DD}(t,p,r)$ i rozstrzygnięcie, czy wprowadzenie dodatkowych krawędzi zachowuje współczynnik skali lub chociaż własność bezskalowości dla $0 < p < \exp(-1)$.

4.5. Kompresja dla duplikacyjnych modeli grafów losowych

W pracy [A5] podjęto zagadnienie kompresji dla grafów generowanych z modelu $\mathtt{DD}(t,1,0)$, czyli zgodnie z tzw. modelem pełnej duplikacji (*full duplication*) [11, 30].

Punktem wyjścia prac nad powyższym problemem była obserwacja, że w takim szczególnym przypadku każdy nowo dodawany wierzchołek będzie dokładną kopią jednego z istniejących wierzchołków – a zatem również kopią jednego z wierzchołków grafu początkowego. Jeśli przyjmiemy, że zaczynamy od grafu G_{t_0} na t_0 wierzchołkach możemy zatem reprezentować graf w postaci $t - t_0$ liczb długości $\log_2 t_0$ opisujących etykiety odpowiednich wierzchołków grafu początkowego. Również strukturę grafu można opisać jako sekwencję t_0 liczb długości $\log_2 t$ opisujących liczbę wierzchołków będących (również pośrednimi) kopiami kolejnych wierzchołków z grafu początkowego. W ten sposób otrzymujemy również proste oszacowania entropii i entropii strukturalnej dla tego modelu:

$$H(G_t) \leq (t - t_0) \log_2 t_0,$$
$$H(S_t) \leq t_0 \log_2 t.$$

Naturalnie, można zadać pytanie, czy można poprawić te oszacowania, a zatem jakie są właściwe dolne ograniczenia informacyjne na możliwości kompresji takich grafów:

Problem [A5]: ile wynoszą dokładne asymptotyczne tempa wzrostu entropii ($H(G_t)$) i entropii strukturalnej ($H(S_t)$) dla grafów wygenerowanych z modelu $\texttt{DD}(t, 1, 0)$?

Głównym wynikiem pracy jest ustalenie z dokładnością do $o(1)$ asymptotycznego tempa wzrostu obu entropii:

Twierdzenie 4.11. *Asymptotycznie wraz z $t \to \infty$ zachodzi*

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

gdzie $H_n = \sum_{k=1}^{n} \frac{1}{k}$ jest n-tą liczbą harmoniczną.

Oprócz tego zaprezentowano dwa algorytmy (dla kompresji grafu i jego struktury) zapewniające, że średnia długość słowa kodowego nie przekracza odpowiedniej entropii plus co najwyżej 2 bity.

Punktem wyjścia jest użycie ogólnej zależności między $H(G_t)$ a $H(S_t)$, dowiedzionej wcześniej wyłącznie dla modelu Barabásiego-Alberta w [80]:

Lemat 4.9 (Ugólnienie lematu 1 w [80])**.** *Dla dowolnego modelu grafów losowych, w którym każde dwa grafy o tej samej strukturze i niezerowym prawdopodobieństwie wygenerowania z modelu mają dokładnie to samo prawdopodobieństwo, zachodzi*

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

gdzie $\Gamma(G_t)$ jest zbiorem możliwych permutacji etykiet grafu G_t, dla których otrzymany graf ma niezerowe prawdopodobieństwo wygenerowania, natomiast $\mathrm{Aut}(G_t)$ jest zbiorem automorfizmów grafu.

Tak jak wspomniano wcześniej, strukturę grafu G_t można zapisać w postaci wektora $(C_{i,t})_{i=1}^{t_0}$, gdzie zmienna losowa $C_{i,t}$ oznacza liczbę wierzchołków będącymi kopiami (wliczając kopie kopii itd.) i-tego wierzchołka z grafu G_{t_0}. Okazuje się, że prawdopodobieństwo struktur jest opisywane tzw. rozkładem wielomianowym Dirichleta. Pozwala to, wykorzystując fakt identyczności rozkładów wszystkich $C_{i,t}$ oraz wiedzę o asymptotycznym rozwinięciu funkcji beta $B(t, t_0)$ występującej w definicji tego rozkładu, na uzyskanie wartości $H(S_t)$.

Obliczenie $|\Gamma(G_t)|$ polega na zaobserwowaniu, że dopuszczalne są wszystkie permutacje etykiet zachowujące zawartość poszczególnych klas, wyznaczonych przez wierzchołki początkowe. Prowadzi to do stwierdzenia, że

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

gdzie C_t jest zmienną mającą rozkład beta-dwumianowy (jako rozkład prawdopodobieństwa brzegowego dla rozkładu wielomianowego Dirichleta) przesunięty o 1, tj. określony wzorem

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

Wyznaczenie asymptotycznej wartości $\mathbb{E}[\log_2 |\Gamma(G_t)|]$ sprowadza się zatem do odpowiedniego podstawienia, a następnie do ponownego użycia asymptotycznych rozwinięć funkcji. Co ciekawe, ustalenie dokładnego rozkładu prawdopodobieństwa zmiennej C_t poprawia zarazem przybliżenie zastosowane w [30], któe sugerowało (bez formalnego dowodu), że można je zastąpić przez ciągłą funkcję gęstości prawdopodobieństwa postaci $f(x) = \exp\left(-\frac{x}{\mathbb{E}[C_t]}\right)$.

Ostatnią częścią jest obliczenie tempa wzrostu funkcji $\mathbb{E}[\log_2 |\mathrm{Aut}(G_t)|]$. Kluczowy krok polega na zauważeniu, że przy założeniu, że graf początkowy był asymetryczny, zachodzi równość

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

ponieważ automorfizmem może być dowolna funkcja odwzorowująca wierzchołki w ramach poszczególnych klas, wyznaczonych przez wierzchołki początkowe. Po otrzymaniu takiego wyniku wystarczy oszacować z odpowiednią dokładnością kolejne wyrażenia szeregu z rozwinięcia Stirlinga tj. $\mathbb{E}[(X+1)\log_2(X+1)]$, $\mathbb{E}[X]$ i $\mathbb{E}[\log_2(X+1)]$ dla zmiennej X mającej dowolny rozkład beta-dwumianowy. Oszacowanie to zostało dowiedzione z wykorzystaniem własności funkcji analitycznych takich jak funkcja beta czy funkcja digamma, jak również przy użyciu narzędzi probabilistycznych np. nierówności Czernowa.

Opracowane algorytmy kompresji dla obu typów wejścia (grafów i ich struktur) zostały oparte o ideę kodowania arytmetycznego. Z powyższych rozważań znane są dokładne wzory na prawdopodobieństwa łączne, brzegowe i warunkowe dla poszczególnych wektorów $(C_{i,t})_{i=1}^{t_0}$. Można zatem zaproponować odwzorowanie struktury grafu zapisanej w postaci takiego wektora na pewną liczbę z zakresu $[0, 1)$, której odpowiednio przycięte rozwinięcie binarne stanowi słowo kodowe z dobrymi gwarancjami kompresji. Podobne podejście można zastosować analogicznie dla wspomnianego na początku opisu grafu (z etykietami) jako ciągu liczb opisujących jego wierzchołki wzorcowe z grafu początkowego. Z właściwości samego kodowania arytmetycznego wprost wynika, że średnia długość słów kodowych nie przekracza odpowiednich entropii powiększonych o dwa bity [34].

Warto zauważyć, że w tym modelu zachodzi znacząca różnica między asymptotycznymi tempami wzrostu obu entropii – inaczej niż np. dla modelu Barabásiego-Alberta, dla którego $H(G_t) = H(S_t) = \Theta(t \log t)$ [80].

Problemy otwarte Znajomość wzoru na entropię oraz optymalnego (z dokładnością do stałej liczby bitów) algorytmu dla szczególnego przypadku $\mathtt{DD}(t, 1, 0)$ narzuca wprost dalsze pytanie o podobną konstrukcję dla ogólniejszego modelu $\mathtt{DD}(t, p, r)$. Należy jednak zwrócić uwagę, że w ogólnym przypadku nie tylko nie istnieje prosta wymienność wierzchołków, umożliwiająca zliczanie ich w zmiennych $C_{i,t}$, ale też nawet nie są spełnione warunki Lematu 4.9, ponieważ różne grafy o tej samej strukturze mogą mieć różne prawdopodobieństwo wygenerowania.

Zarazem, zgodnie z tym, co pokazano w [27, 80], to wyniki zawarte w pracach [A1]-[A3] mogą stanowić obiecujący punkt wyjścia do prac nad entropią rozkładu prawdopodobieństwa grafów generowanych przez ten model, a zatem również do konstrukcji dobrych algorytmów kompresji. W szczególności, ograniczenia co do maksymalnego stopnia i rozkładu stopni w grafie mogą pozwolić na identyfikację stosunkowo małego zbioru grafów o dużym prawdopodobieństwie, a także na ich odpowiednio oszczędną reprezentację.

5. Omówienie celu naukowego prac i osiągniętych pozostałych wyników

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

- wersja konferencyjna: 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, s. 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, s. 156-163.

[C1] Jithin Sreedharan, Wojciech Szpankowski, Krzysztof Turowski, Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18(3) (2021), s. 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), s. 1426-1442.

- wersja konferencyjna: 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, s. 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, s. 1517-1521.

[E1] Tytus Pikies, Krzysztof Turowski, Marek Kubale, Scheduling with complete multipartite incompatibility graph on parallel machines: Complexity and algorithms, *Artificial Intelligence* 309 (2022), s. 103711.

- wersja konferencyjna: 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, s. 262-270 [Best Runner-Up Paper Award].

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

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

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

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

[F5] Krzysztof Michalik, Krzysztof Turowski, On λ-backbone coloring 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), s. 1373-1382.

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

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

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

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

[H1] Robert Janczewski, Paweł Obszarski, Krzysztof Turowski, 2-Coloring number revisited, *Theoretical Computer Science* 796 (2019), s. 187-195.

Prace [F1], [F2], [F3] oraz [F4] zostały opublikowane przed uzyskaniem stopnia doktora.

5.1. Kompresja drzew losowych

Obok kompresji grafów losowych rozważa się również zagadnienie kompresji szczególnych klas grafów. Szczególne miejsce, z uwagi na wykorzystanie drzew m.in. w filogenetyce matematycznej, zajmują problemy kompresji drzew losowych.

Praca [B1] jest poświęcona kompresji ukorzenionych drzew losowych ze skorelowanymi etykietami w wierzchołkach. Badany model ma 4 parametry: docelową liczbę wierzchołków n, długość etykiet m, alfabet $\mathcal{A}$ oraz macierz prawdopodobieństw zamiany liter P (z odpowiednim rozkładem stacjonarnym π), Generowanie drzewa przebiega w następujący sposób: rozpoczynając od drzewa T składającego się pojedynczego wierzchołka (korzenia drzewa) v z etykietą $l(v)$ losowaną zgodnie z rozkładem π[5] powtarzamy kolejno operacje

1. wybierz liść u w drzewie T, dodaj do niego dwoje dzieci w_1 i w_2,
2. dla każdego w_i $(i = 1, 2)$ stwórz etykietę $l(w_i)$ poprzez niezależne losowanie liter wg prawdopodobieństw przejścia z macierzy P i odpowiednich liter z etykiety $l(u)$.

Taki model generowania drzewa (bez etykiet) jest znany w literaturze jako model Yule'a [38].

Przedmiotem badania jest tu zmienna losowa LT_n reprezentująca drzewo z etykietami, równoważnie reprezentowane jako para (T_n, F_n), gdzie T_n odpowiada samemu drzewu bez etykiet, natomiast F_n to sekwencja etykiet np. w kolejności preorder. Celem badań jest oszacowanie entropii zmiennej $H(LT_n)$.

Z reguły łańcuchowej dla entropii wiemy, że $H(LT_n) = H(T_n) + H(F_n|T_n)$. Obliczenie drugiego składnika nie stanowi problemu:

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

gdzie $h(P) = -\sum_{a\in\mathcal{A}} \pi(a) \sum_{b\in\mathcal{A}} P(b|a) \log P(b|a)$ to entropia procesu Markowa z macierzą przejścia P, natomiast $h(\pi) = -\sum\limits_{a\in\mathcal{A}} \pi(a) \log \pi(a)$ to entropia rozkładu stacjonarnego π dla macierzy P.

Aby obliczyć $H(T_n)$ dowodzimy, że nasz model tworzenia drzewa jest równoważny modelowi, w którym wstawiamy do drzewa binarnego losową permutację liczb $\{1, \ldots, n\}$. Dzięki temu możemy wyznaczyć dokładny wzór na rozkład prawdopodobieństwa wylosowania drzew T w zależności od stopni ich wierzchołków wewnętrznych. Na tej podstawie, wykorzystując wyniki przedstawione w [71] można dowieść, że:

Twierdzenie 5.1. *Entropia drzew binarnych o n wierzchołkach generowanych zgodnie z modelem Yule'a z etykietami długości m wynosi*

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

W powyższym modelu zakładano, że drzewo jest dane wraz z orientacją na płaszczyźnie (*plane tree*), tj. kolejność dzieci dla każdego wierzchołka ma znaczenie. Można jednak rozważyć drzewa niezorientowane (*non-plane tree*) i ich odpowiednie entropie z etykietami ($H(LS_n)$) lub bez etykiet ($H(S_n)$).

[5] Ponieważ wpływ losowości etykiety korzenia na wynik końcowy jest znikomy, można również przyjąć, że etykieta korzenia przyjmuje pewną ustaloną wartość.

Praca [B2] stanowi rozwinięcie badań nad powyżej wprowadzonym modelem losowych drzew niezorientowanych. W szczególności dla funkcji $Z(T)$ (w artykule opisanej jako $sym(T)$), zdefiniowanej jako liczba wewnętrznych wierzchołków mających izomorficzne poddrzewa wyprowadzono funkcję tworzącą

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

dla której wyprowadzono równanie różniczkowe

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

gdzie

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

Dalej wykazano, że każdemu drzewu niezorientowanemu S odpowiada $2^{n-1-Z(S)}$ drzew zorientowanych, gdzie $Z(S)$ oznacza liczbę wierzchołków mających dwa identyczne (tj. izomorficzne) poddrzewa. Dla danego drzewa s i drzewa $s \circ s$ jako korzenia mającego dwa identyczne poddrzewa s można stworzyć następujące równanie rekurencyjne:

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

gdzie U_k oznacza zmienną losową z rozkładem równomiernym o wartościach $\{1, \ldots, k\}$. Z tego wprost wynika równanie opisujące wartość oczekiwaną

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

Rozwiązując je otrzymujemy, że

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

a z tego można policzyć pierwsze elementy sumy otrzymując, że

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

W pracy [B1] przedstawiono również trzy algorytmy kompresji drzew losowych oparte o ideę kodowania arytmetycznego. Algorytm dla drzew zorientowanych osobno kompresuje drzewo i osobno etykiety na bazie znanej relacji dziecko-rodzic w drzewie. Działa on w czasie $O(n^2 \log^2 n \log\log n)$ i dla drzew generowanych z modelu opisanego powyżej osiąga średnią długość słowa kodowego nie większą niż $H(LT_n) + 2$ bitów.

Dla drzew niezorientowanych przedstawiono dwa algorytmy: przybliżony szybki CompressNPTree oraz wolniejszy asymptotycznie optymalny. Pierwszy, działający w czasie liniowym, polega na kanonicznym uporządkowaniu drzewa binarnego zgodnie z regułą „mniejsze poddrzewo po lewej". Można dowieść, że taka heurystyka jest algorytmem przybliżonym z nadmiarowością 1% względem algorytmu optymalnego:

Twierdzenie 5.2. *Asymptotycznie gdy $n \to \infty$ średnia długość słowa kodowego w algorytmie* CompressNPTree *nie przekracza* $1.013H(S_n)$.

Wolniejszy algorytm oparty jest z kolei o wyznaczenie rekurencyjnie zdefiniowanego porządku na drzewach, umożliwiającego rozróżnienie przypadków nieizomorficznych drzew o tej samej liczbie liści. To, wraz z algorytmem obliczającym dla danego porządku i dowolnego drzewa niezorientowanego S prawdopodobieństwa $\Pr[S_n < S]$ oraz $\Pr[S_n \leq S]$ wystarczy do zastosowania schematu kodowania arytmetycznego dla otrzymania słowa kodowego o średniej długości nieprzekraczającej $H(S_n) + 2$ bitów. Pesymistyczny czas działania tego algorytmu jest ograniczony przez $O(n^3 \log^2 n \log\log n)$, natomiast średni przez $O(n^2 \log^2 n \log\log n)$ – ponieważ dokładność obliczanych prawdopodobieństw wymaga, by uwzględnić mnożenie liczb o długości $O(n^2)$ bitów.

5.2. Wnioskowanie w modelu duplikacyjnym grafów losowych

Rozważanie znaczenia praktycznego teoretycznych modeli grafów losowych niewątpliwie zaczyna się od pytania o skuteczność dopasowywania modeli do istniejących grafów. Badane modele mają zwykle kilka wolnych parametrów: przykładowo wspomniany wcześniej model Solégo i Pastora-Satorrasa $\texttt{DD}(t, p, r)$ ma, oprócz ustalonej liczby wierzchołków t, dwa takie parametry p i r. Intuicyjnie zależałoby nam, by dana rzeczywista sieć mogła być wygenerowana z jak największym prawdopodobieństwem zgodnie z danego modelu – a zatem należy szukać zestawu parametrów maksymalizujących takie prawdopodobieństwo. Ponieważ jednak okazuje się, że wiele sieci biologicznych czy społecznościowych charakteryzuje się istnieniem wielu automorfizmów, a dla wielu modeli (w tym m.in. dla modelu Solégo i Pastora-Satorrasa) poszczególnym izomorficznym grafom odpowiadają różne prawdopodobieństwa ich wygenerowania, problem ten staje się trudny obliczeniowo w praktyce.

Wcześniejsze prace omijały ten problem opierając się na założeniu, że w przypadku dopasowania sieci do modeli generujących grafy bezskalowe możliwe jest odtworzenie wartości parametrów z wartości współczynnika skali dla stanu stacjonarnego modelu [59, 92]. Przykładowo, dla modelu Solégo i Pastora-Satorrasa $\texttt{DD}(t, p, r)$ do obliczenia p i r korzystano z wzorów podanych w pracach [59, 92]:

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

gdzie γ oznacza współczynnik skali, natomiast $D(G)$ to średni stopień wierzchołka w grafie.

W pracy [C1] to podejście zostało zakwestionowane na bazie argumentów teoretycznych i symulacji. Po pierwsze, w literaturze przedmiotu nadal trwa dyskusja dotycząca występowania zjawiska bezskalowości w przyrodzie. Wskazuje się na pewne argumenty statystyczne, świadczące o tym, że wiele typów sieci biologicznych i społecznościowych wcale nie charakteryzuje się bezskalowością [21, 70, 97]. Można natrafić również na argumenty krytykujące standardowe metody szacowania współczynnika skali γ (np. zaproponowane w [4]) jako wykrywające cechę bezskalowości w grafach generowanych z modeli bez tej właściwości [32]. Niezależnie od wagi tych argumentów, dla badanych przez nas sieci interakcji protein okazało się, że standardowe metody obliczania wartości γ dokonują odcięcia na poziomie zaledwie kilku procent wierzchołków o największym stopniu, a sam dobór wartości progu odcięcia znacząco modyfikuje otrzymywany wynik.

Po drugie, pozbawione dobrego uzasadnienia jest również założenie, że rzeczywiste sieci są na tyle duże, że uzasadnia to zachodzenie z dostatecznym przybliżeniem zależności charakterystycznych dla średniego przypadku w stanie stacjonarnym. Po trzecie, można argumentować, że techniki użyte do dowodzenia tych zależności nie były dostatecznie rygorystycznie dowiedzione. Jakkolwiek trzeba zauważyć, że powyższa równość wiążąca γ i p została udowodniona w pracy [A4] dla modelu $\texttt{DD}(t, p, r)$ z parametrami $0 < p < \exp(-1)$ i $r = 0$, tak Jordan w [68] pokazał, że odpowiedni łańcuch Markowa charakteryzujący ten model dla $p \geq \exp(-1)$ i $r = 0$ jest chwilowy (*transient*), co stawia pod znakiem zapytania możliwość stosowania wzorów w przypadku ogólnym.

Co więcej, empiryczne rozkłady wartości różnych własności grafów generowanych z modelu z parametrami opisanymi powyższymi wzorami różniły się znacząco od rzeczywistych sieci, którym miały odpowiadać. W szczególności przy takich parametrach generowano grafy, które charakteryzowały się zupełnie innym rzędem wielkości liczby automorfizmów – jednej z kluczowych cech zwłaszcza dla rzeczywistych sieci biologicznych np. sieci interakcji protein [11]. Dopasowanie tego współczynnika jest tym ważniejsze, że dowiedziono, że najpopularniejsze modele grafów losowych (Erdősa-Rényi'ego i Barabásiego-Alberta) generują z dużym prawdopodobieństwem dla szerokiego zakresu parametrów grafy pozbawione symetrii.

W naszej pracy [C1] przedstawiliśmy nowy sposób estymacji parametrów dla grafów duplikacyjnych, uwzględnieniający również rozważania o dopasowaniu liczby automorfizmów. Najpierw skorzystaliśmy z rekurencji (4.2) dla $D(G_n)$ z [A1] oraz wyprowadziliśmy podobne równania dla innych parametrów grafu, takich jak liczba trójkątów $C_3(G_n)$ i liczba "otwartych trójkątów" (tj.

gwiazd o dwóch liściach) $S_2(G_n)$:

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

Te wzory, w połączeniu z przeszukiwaniem binarnym zbioru wszystkich możliwych (p, r), pozwoliły znaleźć na podstawie obserwowanych wartości $D(G_n)$ i $D(G_{n_0})$ zbiory parametrów, które pasują do tych wartości[6] i otrzymać iloczyn tych zbiorów jako zbiór dobrych parametrów wyjściowych. Następnie sprawdzono, że dla grafów o znanych p i r metoda ta zwraca rzeczywiście parametry bliskie ich prawdziwym wartościom, a więc również pasujące do ich liczby automorfizmów. Kolejnym krokiem było zastosowanie tego podejścia dla wybranych rzeczywistych sieci biologicznych, w wyniku czego uzyskaliśmy nowe oszacowania parametrów p i r. Okazało się, że modele z tak obliczonymi parametrami generują grafy, które nie tylko mają bliskie sieciom rzeczywistym wartości $D(G_n)$ i innych zmiennych użytych w powyższych wzorach, ale także mają bardzo podobne wartości $\mathrm{Aut}(G_n)$, co dodatkowo potwierdza słuszność takiego podejścia.

Zarysowana powyżej procedura okazała się również bardzo praktyczna, gdyż działa w czasie liniowym. Dla porównania, nasze referencyjne podejście zgodne z metodą największej wiarygodności (*maximum likelihood estimation*) również prowadziło do podobnych oszacowań dla małych grafów, ale wymagało operacji $Theta(n^3)$, co czyniło je niepraktycznym dla sieci rzeczywistych o setkach tysięcy wierzchołków.

Uzasadnienie zgodności sieci z modelem oraz oszacowanie odpowiednich parametrów stanowi punkt wyjścia do podjęcia kolejnych obecnych w literaturze problemów. Jednym z nich jest tzw. problem archeologii sieci (*network archaeology*), czyli rekonstrukcji ewolucji sieci w czasie i odtworzenia porządku czasowego wierzchołków w sieciach rzeczywistych. Problem ten, rozważany wcześniej w kontekście innych modeli grafów losowych [84, 105], stanowi ważne zagadnienie badawcze prowadzące do odpowiedzi na szereg pytań praktycznych, takich jak identyfikacja dawnych składowych grafu, czy wnioskowanie o strukturalnej i funkcjonalnej ewolucji mózgu [94].

Zgodnie z formalną definicją zaproponowaną w analizie modelu Barabásiego-Alberta [94] przyjęto, że celem jest znalezienie częściowego porządku σ, zdefiniowanego nad wierzchołkami zadanego grafu, optymalizującego dwa kryteria:

- *gęstość*: miarę liczności częściowego porządku σ jako znormalizowaną liczbę rozróżnialnych par w porządku σ:

$$\delta(\sigma) = \frac{K(\sigma)}{\binom{n}{2}},$$

  gdzie $K(\sigma) = |\{(u, v) \colon u <_\sigma v\}|$,
- *precyzję*: miarę oczekiwanej liczby poprawnie zidentyfikowanych par wierzchołków z oryginalnego porządku czasowego π, wyrażoną jako ich odsetek z wszystkich rozróżnialnych par w porządku σ:

$$\theta(\sigma) = \mathbb{E}\left[\frac{|\{u, v \in \{1, \ldots, n\} \colon u <_\sigma v, \pi^{-1}(u) < \pi^{-1}(v)\}|}{K(\sigma)}\right].$$

[6] W rzeczywistości założyliśmy, że obserwowane wartości mogą nieco odbiegać od teoretycznych średnich, powiększając tym samym zbiory sensownych parametrów.

W tym ujęciu dla każdej ustalonej wartości $\varepsilon \in (0,1]$ porządek σ o gęstości $\delta(\sigma) \geq \varepsilon$ maksymalizujący $\theta(\sigma)$ pokazuje teoretyczne granice rekonstruowalności porządku tj. oczekiwaną jakość predykcji przy wymaganiu aby co najmniej odsetek ε par wierzchołków był rozpoznawalny. Wartości te stanowią zarazem cel dla algorytmów odtwarzania czasowego porządku w sieciach.

Praca [C2] została poświęcona właśnie problemowi archeologii sieci dla szerokiej klasy modeli grafów losowych ze szczególnym uwzględnieniem modeli duplikacyjnych. W szczególności, zrezygnowano z założenia charakteryzującego model Barabásiego-Alberta o równym prawdopodobieństwie wygenerowania dwóch izomorficznych grafów, ponieważ modele duplikacyjne w ogólności nie charakteryzują się takimi własnościami.

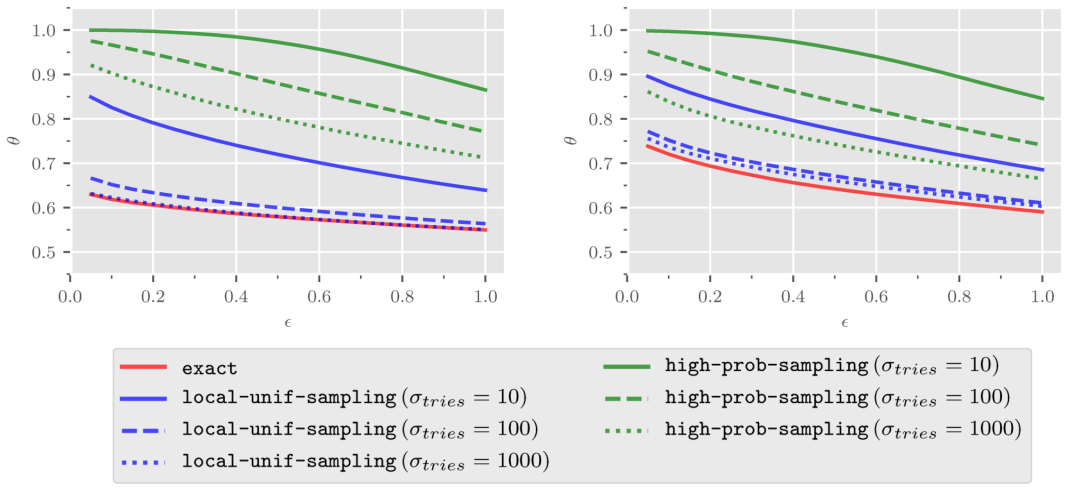
Pierwszym wynikiem było zaproponowanie dwóch programów całkowitoliczbowych i ich relaksacji do programów liniowych dla znajdowania porządku w sieciach dynamicznych tylko na podstawie ich stanu w pewnym ustalonym momencie. Oba programy w ogólnej postaci zależały tylko od parametrów opisujących prawdopodobieństwo $p_{u,v}$ kolejności wszystkich par wierzchołków u i v na podstawie struktury grafu i przyjętego modelu jego ewolucji w czasie (patrz Table 1). Drugim krokiem było zaproponowanie odpowiedniej procedury próbkowania (tzw. *sequential importance sampling*) do oszacowania wartości $p_{u,v}$ dla dowolnego modelu grafów losowych, który można zapisać w postaci niehomogenicznego w czasie łańcucha Markowa np. opisującego ewolucję polegającą na sekwencyjnym losowym dodawaniu wierzchołków i krawędzi. Dowiedziono, że procedura ta definiuje asymptotyczną silną zgodność (*strong consistency*) tj. prawie pewną zbieżność (*almost sure convergence*) do prawdziwych wartości $p_{u,v}$.

LP-CLUSTERS		LP-PARTIAL-ORDER	
IP	**Relaksacja LP**	**IP**	**Relaksacja LP**
$\max_{z} \frac{\sum_{\substack{1\leq u\neq v\leq n \\ 1\leq i\ <j\leq n}} p_{u,v}\, z_{u,i,v,j}}{\sum_{\substack{1\leq k<l\leq n \\ 1\leq w\ \neq w'\leq n}} z_{w,k,w',l}}$	$\max_{z'} \sum_{\substack{1\leq u\neq v\leq n \\ 1\leq i\ <j\leq n}} p_{u,v}\, z'_{u,i,v,j}$	$\max_{y} \frac{\sum_{1\leq u\neq v\leq n} p_{u,v}\, y_{u,v}}{\sum_{1\leq u\neq v\leq n} y_{u,v}}$	$\max_{y'} \sum_{1\leq u\neq v\leq n} p_{u,v}\, y'_{u,v}$
przy warunkach	przy warunkach	przy warunkach	przy warunkach
$\forall_{u,i,v,j\in[n]}\ z_{u,i,v,j} \in \{0,1\},$	$\forall_{u,i,v,j\in[n]}\ z'_{u,i,v,j} \in [0, 1/\epsilon\binom{n}{2}],$	$\forall_{u,v\in[n]}\ y_{u,v} \in \{0,1\},$	$\forall_{u,v\in[n]}\ y'_{u,v} \in [0, 1/\epsilon\binom{n}{2}]$
$\sum_{\substack{1\leq u\neq v\leq n \\ 1\leq i\ <j\leq n}} z_{u,i,v,j} \geq \epsilon\binom{n}{2},$	$\sum_{\substack{1\leq u\neq v\leq n \\ 1\leq i\ <j\leq n}} z'_{u,i,v,j} = 1,$	$\sum_{1\leq u\neq v\leq n} y_{u,v} \geq \epsilon\binom{n}{2}$	$\sum_{1\leq u\neq v\leq n} y'_{u,v} = 1$
$\sum_{i\in[n]} z_{u,i,u,i} = 1,$	$\sum_{i\in[n]} z'_{u,i,u,i} \leq 1/\epsilon\binom{n}{2},$	$y_{u,v} + y_{v,u} \leq 1,$	$y'_{u,v} + y'_{v,u} \leq 1/\epsilon\binom{n}{2}$
$z_{u,i,v,j} = z_{v,j,u,i},$ $\sum_{i\in[n]} z_{u,i,v,j} = z_{v,j,v,j}.$	$z'_{u,i,v,j} = z'_{v,j,u,i},$ $\sum_{i\in[n]} z'_{u,i,v,j} = z'_{v,j,v,j}.$	$y_{u,v} + y_{v,w} - y_{u,w} \leq 1.$	$y'_{u,v} + y'_{v,w} - y'_{u,w} \leq 1/\epsilon\binom{n}{2}.$

Tablica 1: Dwa zaproponowane programy całkowitoliczbowe i ich relaksacje

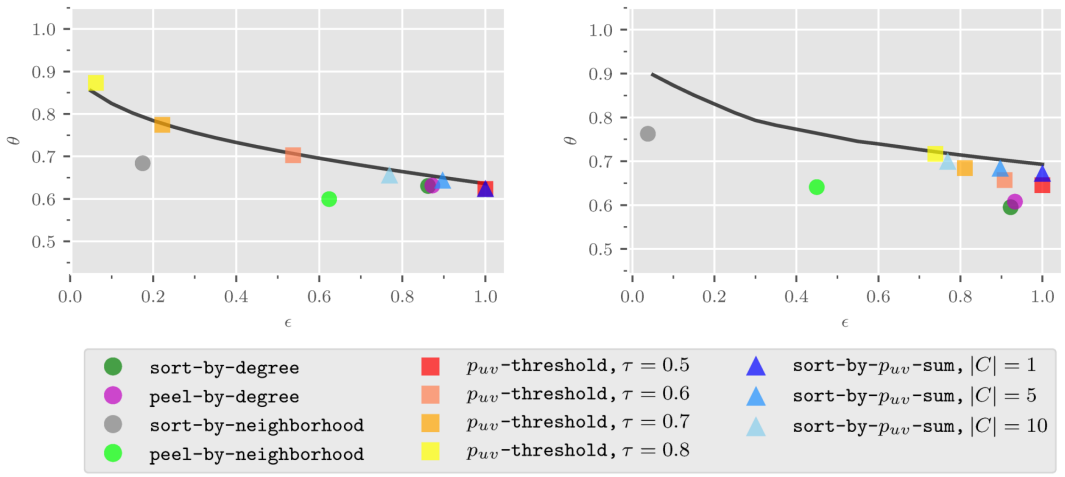
Powyższe pomysły zostały zastosowane do modelu $\mathtt{DD}(t,p,r)$. W szczególności z definicji modelu zostały wyprowadzone szczegółowe wzory umożliwiające zastosowanie procedury próbkowania do obliczenia wartości $p_{u,v}$ oraz wyznaczenia na ich podstawie według programowania liniowego optymalnych wartości θ i δ dla przykładowych grafów wygenerowanych z modelu dla przykładowo ustalonych wartości parametrów. Ponieważ procedura próbkowania zostawia pewną dowolność wyboru funkcji wag, wykonano również szereg eksperymentów ilustrujących zbieżność do prawdziwych wartości $p_{u,v}$ wraz z liczbą próbek. Ostatecznie, zaproponowane zostały dwie metody próbkowania z różnymi wagami: proporcjonalnymi HIGH-PROB-SAMPLING i identycznymi LOCAL-UNIF-SAMPLING. Okazało się, że ta druga miała wyraźnie szybsze tempo zbieżności w praktyce (patrz Figure 2).

Skoro już same wartości $p_{u,v}$ dają pewną informację o poszukiwanym porządku a wyznaczanie $p_{u,v}$ w praktyce okazuje się dużo szybsze niż zastosowanie programowania liniowego, w pracy również zaproponowano dwa algorytmy, nazwane $p_{u,v}$-THRESHOLD i SORT-BY-$p_{u,v}$-SUM, oparte na uporządkowaniu wierzchołków w grupy zgodnie z wartościami $p_{u,v}$. Aby sprawdzić jakość wybranych algorytmów przeprowadzono symulacje na grafach wygenerowanych z modelu Solégo i Pastora-Satorrasa z ustalonymi parametrami. Wyniki pokazały, że jakość rozwiązań zwracanych przez zaproponowane algorytmy nie odbiega znacząco od górnej granicy wyznaczonej z odpowiedniego pro-



Rysunek 2: Wyniki zbieżności do dokładnej krzywej precyzji (θ)-minimalnej gęstości (ε): $G_n \sim \texttt{DD}(13, p, 1.0, G_{n_0})$ dla $p = 0.3$ (po lewej) i 0.6 (po prawej), uśrednione po 100 grafach. G_{n_0} zostało wygenerowane według modelu Erdősa-Renyí'ego z $n_0 = 4$ i $p_0 = 0.6$.

gramowania liniowego, przynajmniej dla dostatecznie małych grafów, dla których obliczenie tych ograniczeń było możliwe. Porównano również wyniki z prostymi heurystykami opartymi o właściwości grafu takie jak sortowanie według stopni czy relacji zawierania się sąsiedztw wierzchołków. Heurystyki te, jakkolwiek proponowane dla innych modeli, zawodzą: jak wiemy np. z rozważań w [A1] oczekiwany stopień wierzchołka $\mathbb{E}[\deg_t(u)]$ dla ustalonego t rośnie wraz z u. I rzeczywiście okazuje się, że wyniki otrzymane dzięki algorytmom opartym na wartościach $p_{u,v}$ są nieznacznie, ale jednak zauważalnie lepsze (zobacz Figure 3).



Rysunek 3: Wyniki dla algorytmów zachłannych i opartych o wartości $p_{u,v}$: $G_n \sim \texttt{DD}(50, p, 1.0, G_{n_0})$ dla $p = 0.3$ (po lewej) i 0.6 (po prawej), uśrednione dla 100 grafów. Oszacowania $p_{u,v}$ wykonane na bazie $\sigma_{\text{tries}} = 100{,}000$ ścieżek. G_{n_0} zostało wygenerowane według modelu Erdősa-Renyí'ego z $n_0 = 10$ i $p_0 = 0.6$.

Ostatnim krokiem było ulepszenie zaproponowanych algorytmów opartych na wartościach $p_{u,v}$, aby potrafiły uwzględniać wiedzę zewnętrzną w postaci tzw. par doskonałych jako wiedzy pewnej (zewnętrznej) o kolejności określonych wierzchołków w prawdziwym porządku π.

5.3. Szacowanie entropii rozkładów prawdopodobieństwa

W społeczności zajmującej się teorią informacji jednym z ważnych problemów badawczych jest obliczanie dokładnych i asymptotycznych wartości funkcji entropii dla popularnych rozkładów prawdopodobieństwa. Formalnie, dla rozkładu dyskretnego opisanego przez $\Pr[X_n = k]$ dla danych n i k szukamy wartości

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

Wcześniejsze prace doprowadziły do ustalenia odpowiednich wartości dla rozkładu dwumianowego [43, 62], ujemnego rozkładu dwumianowego [31] i rozkładu Poissona [76].

Wyprowadzono również wzory opisujące dokładne wartości entropii dla wielu innych rozkładów np. beta-dwumianowego i hipergeometrycznego [26]. Wynikami tych prac były jednak tylko wzory zawierające bardzo skomplikowane funkcje, trudne do asymptotycznego oszacowania.

W pracy [D1] odpowiadamy na analogiczne pytanie dla rozkładu wielomianowego Dirichleta, który okazał się kluczowy dla analiz z pracy [A5]. Jest on zarazem wprost związany z modelem urn Pólyi: dla danych m urn, każdej zawierającej α_i $(i = 1, \ldots, m)$ kul oraz procesu losowania urn z równomiernym prawdopodobieństwem wraz z dorzucaniem dodatkowej kuli do tak wybranej urny, rozkład kul w urnach dąży właśnie do tego rozkładu. W szczególności, dla zmiennej $\bar{X}$ z rozkładu wielomianowego Dirichleta $DM(n, \bar{\alpha})$ prawdopodobieństwo przyjęcia wartości $\bar{x} = (x_1, \ldots, x_m)$ jest równe

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

gdzie $\Gamma(x)$ jest funkcją gamma Eulera, natomiast $\alpha_0 = \sum_{k=1}^{m} \alpha_k$.

Dowód polegał na sprowadzeniu wyrażenia $H(\bar{X})$ do sumy czynników zawierających wyrażenia postaci $\mathbb{E}[\Gamma(X_k + l)]$ dla X_k mającego rozkład beta-dwumianowy z parametrami n, α_k i $\alpha_0 - \alpha_k$, oraz pewnych stałych wartości l. Następnie wykorzystano fakt, że dla zmiennej $X \sim BBin(n, \alpha, \beta)$ zachodzi

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

dla funkcji prawdopodobieństwa rozkładu beta $\pi(p, \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$ (z funkcją beta $B(\alpha, \beta)$) i zmiennej $X_p \sim Bin(n, p)$.

Każde z wyrażeń $\mathbb{E}[\Gamma(X_p + l)]$ zostało następnie rozłożone w szereg Taylora wokół średniej np a powstałe wyrażenia zostały w odpowiedni sposób asymptotycznie oszacowane z użyciem funkcji hipergeometrycznych i ich rozwinięć. Ostatecznie otrzymano, że

Twierdzenie 5.3. *Dla zmiennej $\bar{X} \sim DM(n, \bar{\alpha})$ zachodzi*

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

dla jawnie obliczonych współczynników e_s.

5.4. Szeregowanie zadań z grafem ograniczeń

Szeregowanie zadań jest jednym z ważniejszych zagadnień w ramach badań operacyjnych, obejmującym wiele problemów z różnymi ograniczeniami i kryteriami optymalizacji [3]. Wskazuje się, że ta dziedzina ma pewne znaczenie praktyczne m.in. w modelowaniu zagadnień optymalizacji dla przetwarzania w chmurze [5, 25, 98].

Podstawowym problemem szeregowania, blisko związanym z zagadnieniem podziału zbioru, jest problem $P||C_{max}$[7], zdefiniowany następująco:

Definicja 5.1 ($P||C_{max}$)**.** *Dla zbioru zadań $J = \{j_1, \ldots, j_n\}$, zbioru maszyn $M = \{m_1, \ldots, m_m\}$ oraz funkcji prędkości wykonania zadań $p\colon J \to \mathbb{N}_+$ należy znaleźć harmonogram przydziału zadań do maszyn, w którym:*

1. *każde zadanie jest całkowicie wykonane na pewnej maszynie tj. harmonogram przyporządkowuje każdemu zadaniu j_k parę maszyna i przedział $[t, t + p(j_k))$,*
2. *jeśli maszyna m_j wykonuje zadanie j_k w czasie $[t, t + p(j_k))$, to nie wykonuje w tym czasie żadnego innego zadania,*
3. *maksymalny czas zakończenia zadań (C_{max}) na maszynach jest jak najmniejszy.*

Ten problem doczekał się szeregu uogólnień i różnorodnych wariantów, uwzględniających np. różną szybkość przetwarzania zadań na maszynach czy istnienie zależności kolejnościowych między zadaniami. Ich definicje i przegląd można znaleźć np. w [12, 23].

Przykładem takiej szerokiej klasy problemów jest szeregowanie zadań z grafami ograniczeń, wprowadzone w [15]. W tych problemach zakłada się, że na wejściu podawany jest dodatkowo tzw. *graf ograniczeń* G z $V(G) = J$ taki, że jeśli $\{j_k, j_l\} \in E(G)$, to odpowiednie zadania nie mogą zostać wykonane na tej samej maszynie. Problem ten stanowi uogólnienie innych zagadnień definiowanych w terminach szeregowania zadań np. Bounded Independent Sets [14] i Mutual Exclusion Scheduling [7, 49, 65].

W przeszłości pokazano m.in. algorytm wielomianowy dla $P|\chi(G) = k|C_{\max}$ tj. gdy graf ograniczeń jest k-kolorowalny dla ustalonego k, a także algorytm 2-przybliżony dla $P|G = bipartite|C_{\max}$ i w pełni wielomianowy schemat aproksymacyjny (*fully polynomial time approximation scheme*, FPTAS) dla grafów o ograniczonej szerokości drzewiastej tj. $P|tw(G) = k|C_{\max}$ [15]. Przedmiotem badań stały się również grafy będące uniami klik [36, 53, 66], dla których opracowano m.in. wielomianowy schemat aproksymacyjny (*polynomial time approximation scheme*, PTAS) dla przypadku maszyn identycznych oraz $(\log n)^{1/4-\varepsilon}$-nieaproksymowalność (o ile $\mathsf{P} \neq \mathsf{NP}$) dla maszyn dowolnych.

Jak widać, typowo przyjmuje się, że klasy grafów ograniczeń ma stosunkowo prostą charakterystykę. Jest to uzasadnione bezpośrednim związkiem tej klasy problemów szeregowania zadań z problemami kolorowania odpowiedniej klasy grafów, a zatem z przenoszeniem się wyników np. o NP-trudności znalezienia $O(n^{1-\varepsilon})$-przybliżenia dla liczby chromatycznej grafu [110].

W pracy [E1] podjęty został problem szeregowania zadań z k-dzielnymi pełnymi grafami ograniczeń. Rozważano dwa warianty problemu: gdy liczba partycji grafu ograniczeń jest parametrem problemu (ograniczenie oznaczane *k-partite*) lub gdy jest ona częścią danych wejściowych (*multipartite*).

Dla maszyn identycznych dowiedziono wcześniej, że $P|G = complete\ multipartite|C_{\max}$ jest problemem silnie NP-trudnym, ale zarazem że istnieje dla niego PTAS [15]. W pracy [E1] dowiedziono, że jeśli zmienimy kryterium na $\sum C_j$, to istnieje algorytm, który zwraca dokładne rozwiązanie problemu w czasie $O(mn + n \log n)$. Jest to algorytm zachłanny, polegający na przydzielaniu maszyn do zbiorów zadań w taki sposób, aby lokalnie jak najbardziej zmniejszać bieżącą wartość rozwiązania. W ramach ustalonego przydziału maszyn do zbiorów zadań wystarczy obliczyć odpowiednią liczbę podproblemów $P||\sum C_j$ aby otrzymać kolejność zadań na poszczególnych maszynach.

Dla maszyn jednorodnych punktem wyjścia jest ustalenie silnej NP-trudności dla problemów $Q|G = complete\ multipartite, p_j = 1|C_{\max}$ i $Q|G = complete\ multipartite, p_j = 1|\sum C_j$. W obu przypadkach można przeprowadzić odpowiednią redukcję z dobrze znanego problemu 3-Partition [51]: wystarczy, że

- rozmiary przedmiotów zostaną wprost odwzorowane na szybkości maszyn,
- ograniczenie na sumę przedmiotów w jednym podzbiorze zostanie wprost odwzorowane na rozmiar każdej partycji J_i,
- liczba docelowych podzbiorów zostanie wprost odwzorowana na liczbę partycji w grafie.

[7] Stosujemy tu tzw. notację trójpolową Lawlera [78], zgodnie z którą w pierwszym polu umieszczany jest typ maszyn lub problemu np. P – maszyny identyczne, Q – jednorodne, R – dowolne; w drugim polu ograniczenia np. na długości zadań, klasę grafów ograniczeń; w trzecim polu kryterium optymalizacji, typowo C_{max} lub $\sum C_j$.

Problem	**Przybliżenie**	**Złożoność**
$P\|G = \textit{complete multipartite}\|\sum C_j$	dokładny	$O(mn + n\log n)$
$Q\|G = \textit{complete k-partite}, p_j = 1\|C_{\max}$	dokładny	$O(mn^{k+1}\log(mn))$
$Q\|G = \textit{complete k-partite}, p_j = 1\|\sum C_j$	dokładny	$O(mn^{k+1})$
$Q\|G = \textit{complete k-partite}\|\sum C_j$	$1+\varepsilon$ (PTAS)	P
$Q\|G = \textit{complete multipartite}, p_j = 1\|C_{\max}$	Silnie NP-trudny	
	2	$O(mn\log(mn))$
	$1+\varepsilon$ (PTAS)	P
$Q\|G = \textit{complete multipartite}, p_j = 1\|\sum C_j$	Silnie NP-trudny	
	4	$O(m^2n^3\log m)$
$R\|G = \textit{complete 2-partite}, p_j \in \{p_1, p_2\}\|C_{\max}$	$O(n^b s_{\max}^{1-c})$-nieaproksymowalny dla $b, c > 0$	
$R\|G = \textit{complete multipartite}\|C_{\max}$	$(1+\epsilon)s_{\max}$	P
$R\|G = \textit{complete 2-partite}, p_j \in \{p_1, p_2\}\|\sum C_j$	$O(n^b s_{\max}^{1-c})$-nieaproksymowalny dla $b, c > 0$	
$R\|G = \textit{complete multipartite}\|\sum C_j$	$s_{\max}$	$O(mn + n\log n)$

Tablica 2: Podsumowanie wyników otrzymanych w pracy [E1].

Wówczas okazuje się, że istnieje wymagany podział zbioru dla instancji problemu 3-Partition wtedy i tylko wtedy, gdy istnieje harmonogram z odpowiednim ograniczeniem $C_{\max}$ lub $\sum C_j$.

Dla analogicznych problemów, w których liczba partycji grafu ograniczeń nie należy do wejścia, ale jest parametrem problemu, istnieją odpowiednie algorytmy oparte o ideę programowania dynamicznego połączonego z przeszukiwaniem binarnego dla szacowanej wartości $C_{\max}$, zawierającej się w przedziale $[1, mn]$.

Dla problemu $Q|G = \textit{complete k-partite}|\sum C_j$ opracowano PTAS oparty o połączenie trzech idei: zaokrąglenia szybkości maszyn do postaci $(1+\varepsilon)^i$, wyczerpującego przeszukiwania dla przydziału najszybszych maszyn dla każdej partycji grafu ograniczeń, oraz programowania liniowego służącego do ustalenia przydziału pozostałych maszyn. Odpowiednie warunki programu zapewniają, że wszystkie zadania mają przydzielone maszyny, a dodatkowa procedura przeogranizowująca oparta na idei przepływu pozwala zamienić rozwiązanie, w którym pojedyncze zadania zostały przyporządkowane do różnych maszyn, na dopuszczalny harmonogram bez zwiększania całkowitego kosztu.

Kolejnym krokiem było przedstawienie dwóch algorytmów przybliżonych dla wspomnianych wcześniej problemów silnie NP-trudnych: 2-przybliżonego dla $Q|G = \textit{complete multipartite}, p_j = 1|C_{\max}$ oraz 4-przybliżonego dla $Q|G = \textit{complete multipartite}, p_j = 1|\sum C_j$. Oba algorytmy zostały oparte na idei binarnego przeszukiwania długości harmonogramu, obliczania według tego dla każdej maszyny pojemności (tj. liczby możliwych zadań do przetworzenia), a następnie przydzielania części (posortowanych nierosnąco według rozmiarów) do maszyn (posortowanych nierosnąco według pojemności) zachłannie aż do osiągnięcia pokrycia 50%. Dla drugiego problemu $Q|G = \textit{complete multipartite}, p_j = 1|C_{\max}$ został również pokazany PTAS, łączący szereg idei, takich jak:

- zgadnięcie (przeszukiwanie binarne) górnego ograniczenia wartości $C_{\max}$,
- zaokrąglenie szybkości maszyn do postaci $(1+\varepsilon)^i$,
- podział maszyn według szybkości na bardzo małe, małe, średnie i duże,
- podział partycji grafu ograniczeń na zakresy wg podobnej liczności i iteracyjne przetwarzanie kolejnych zakresów,
- konstrukcja odpowiedniego programu dynamicznego obliczającego wektory stanów (odpowiednio wybrane częściowe harmonogramy) dla kolejnych zakresów,
- odpowiednie przydzielanie maszyn różnych rodzajów w ramach jednego kroku programu dynamicznego.

Sednem dowodu jest pokazanie, że w kolejnych zbiorach wektorów stanów dla kolejnych zakresów co najmniej jeden wektor będzie „dobry" tj. po zaokrągleniu będzie zapewniał $(1+\varepsilon)$ przybliżenie – a zatem jeśli dla danego ograniczenia $C_{\max}$ istnieje odpowiedni harmonogram, to zbiór końcowych wektorów stanów będzie niepusty i będzie w nim poszukiwane rozwiązanie przybliżone.

Uzupełnieniem tych wyników jest pokazanie, że dla maszyn dowolnych i kryterium $C_{\max}$ (odpowiednio, $\sum C_j$) istnieje prosty algorytm s_{max}-przybliżony (odpowiednio, $((1+\varepsilon)s_{max})$-przybliżony)

tj. zastosowanie algorytmu dla maszyn identycznych daje właśnie taki współczynnik przybliżenia. Z drugiej strony, pokazano, że o ile zachodzi $\mathsf{P} \neq \mathsf{NP}$, to nie istnieje żaden wielomianowy algorytm $O(n^b s_{max}^{1-c})$-przybliżony dla dowolnych $b, c > 0$ dla obu kryteriów $C_{\max}$ i $\sum C_j$ nawet, gdy przyjmiemy, że graf ograniczeń jest dwudzielny.

5.5. Szkieletowe kolorowanie grafów i jego uogólnienia

Niniejsze badania są kontynuacją pracy doktorskiej, badającej zagadnienie szkieletowego kolorowania grafów:

Definicja 5.2 (λ-szkieletowe pokolorowanie grafu)**.** *Dla danego grafu G i jego podgrafu spinającego H funkcja $c\colon V(G) \to \mathbb{N}_+$ jest* λ-szkieletowym pokolorowaniem grafu G ze szkieletem H*, gdy:*

- *dla każdej krawędzi $\{u, v\} \in E(G)$ zachodzi $|c(u) - c(v)| \geq 1$,*
- *dla każdej krawędzi $\{u, v\} \in E(H)$ zachodzi $|c(u) - c(v)| \geq \lambda$.*

Definicja 5.3 (Problem λ-szkieletowego kolorowania grafów)**.** *Dla danego grafu G i jego podgrafu spinającego H należy znaleźć λ-szkieletową liczbę chromatyczną $BBC_\lambda(G, H)$ jako najmniejszą liczbę $k \in \mathbb{N}_+$ taką, że istnieje λ-szkieletowe pokolorowanie c grafu G ze szkieletem H spełniające $\max_{u \in V(G)} c(u) \leq k$.*

Problem ten, wprowadzony w 2003 roku przez Hajo Broersmę [18] doczekał się szeregu szczegółowych badań zarówno dla różnych klas grafów np. split grafów [20] czy grafów planarnych [56], jak i dla różnych klas szkieletów np. skojarzeń i rozłącznych gwiazd [20] czy lasów [56]. Był on również przedmiotem prac w ramach cyklu składającego się na doktorat autora [F1],[F2],[F3],[F4]. Szczególnym zainteresowaniem cieszyły się badania przypadku $\lambda = 2$ dla różnych klas grafów i szkieletów [6, 19, 82, 83].

Praca [F5] jest poświęcona problemowi kolorowania szkieletowego dla klik z lasami w szkielecie. Poprzednio, w [63] dowiedziono, że istnieje 2-przybliżony algorytm dla grafów pełnych z dwudzielnym szkieletem oraz $\frac{3}{2}$-przybliżony algorytm dla grafów pełnych ze spójnym dwudzielnym szkieletem. Oba algorytmy działają w czasie liniowym. Pierwsza część [F5] została poświęcona poprawieniu tego wyniku dla szkieletów będących lasami.

Twierdzenie 5.4. *Dla lasu F na n wierzchołkach i $\lambda \geq 2$ zachodzi nierówność $BBC_\lambda(K_n, F) \leq \max\{n, 2\lambda\} + \Delta^2(F)\lceil \log n \rceil$.*

Dodatkowo, istnieje algorytm działający w czasie $O(n)$ znajdujący odpowiednie λ-szkieletowe pokolorowanie grafu.

W szczególności, implikuje to, że istnieje liniowy algorytm z addytywnym błędem nie przekraczającym $\Delta^2(F)\lceil \log n \rceil$ – co poprawia poprzednie ograniczenia dla $\Delta(F) = o(\sqrt{n}/\log n)$.

Dowód oparty jest o ideę czerwono-niebiesko-żółtej (k, l)-dekompozycji tj. podziału wierzchołków drzewa na trzy zbiory niezależne R, B, Y w taki sposób, aby R i B były zbiorami niezależnymi z $||R| - |B|| \leq k$, natomiast Y był zbiorem o $|Y| \leq l = \lceil \log n \rceil$. Dekompozycja ta jest aplikowana rekurencyjnie, zgodnie z podziałem drzewa na poddrzewa o wielkości co najwyżej połowy drzewa-rodzica (co można wykonać w czasie $O(n)$). Poprzez odpowiedni przydział kolorów wierzchołkom z Y (najmniejsze i największe), oraz zapewnienie odpowiednich różnic między przedziałami kolorów przydzielonymi poszczególnym zbiorom dekompozycji można dowieść, że otrzymane kolorowanie spełnia podane ograniczenia.

Drugą część pracy stanowi konstrukcja nieskończonej rodziny drzew ograniczonego stopnia ($\Delta(T_r) = 3$) takiej, że $BBC_\lambda(K_n, T_r) \geq \max\{n, 2\lambda\} + \Omega(\log n)$. Drzewa te, nazwane drzewami Fibonacciego, ponieważ konstrukcja r-tego drzewa Fibonacciego T_r polega na połączeniu $(r-1)$-ego i $(r-2)$-ego drzewa Fibonacciego z trzema dodatkowymi wierzchołkami w nowe drzewo.

Okazuje się, że znalezienie kolorowania lub równoważnej mu czerwono-niebiesko-żółtej (k, l)-dekompozycji drzewie T_r dla $l = o(\log n)$ sprowadza się do problemu znalezienia takich liczb $a_i \in \{-1, 0, 1\}$, żeby zachodziło $\sum_{i=0}^{r} |a_i| \leq l + 1$ oraz $\sum_{i=0}^{r} a_i F_i = \frac{F_r}{2} + o(\log n)$, gdzie F_i oznacza i-tą liczbę Fibonacciego. Intuicyjnie Y wyznacza nam podział T_r na $2|Y| + 1$ drzew[8], natomiast

[8] Wprost dzieli nam na co najwyżej $|Y| + 1$ drzew, ale jeszcze dopuszczamy pewne operacje lokalne.

znak a_i określa czy więcej jest w danym poddrzewie wierzchołków czerwonych czy niebieskich[9]. Składnik $o(\log n)$ odpowiada za możliwe modyfikacje związane z wyborem wierzchołków do zbioru Y.

Dzięki twierdzeniu Zeckendorfa określającemu dekompozycję dowolnej liczby na sumę liczb Fibonacciego wiemy, że $\frac{F_r}{2}$ da się rozłożyć na sumą aż $\frac{r}{3}$ liczb Fibonacciego [77]. Uogólniając to na operacje sumy i różnicy otrzymujemy, że drzewo o $\Theta(F_r)$ wierzchołkach wymaga dekompozycji na co najmniej $\Omega(r)$ czerwono-niebieskich poddrzew – a zatem wymagane jest, by $|Y| = \Omega(r)$.

Uogólnieniem problemu λ-szkieletowego kolorowania grafów a zarazem pewną formalizacją zagadnienia przydziału częstotliwości [39, 55] jest tzw. problem ξ-kolorowania grafów: dla każdej pary stacji nadawczych wymagamy, by ich częstotliwości były oddalone o pewne pasma (zależne od położenia i siły nadajników), aby nie interferowały ze sobą.

Definicja 5.4 (ξ-pokolorowanie grafu)**.** *Dla danego grafu G i funkcji $\xi\colon E(G) \to \mathbb{N}_+$ funkcja $c\colon V(G) \to \mathbb{N}_+$ jest* ξ-pokolorowaniem grafu G*, gdy dla każdej krawędzi $\{u, v\} \in E(G)$ zachodzi $|c(u) - c(v)| \geq \xi(\{u, v\})$.*

Problem ten uogólnia również zagadnienie $L(p, q)$-etykietowania grafów [54], w którym wymagamy by sąsiedzi w grafie otrzymali kolory różniące się o co najmniej p, natomiast wierzchołki w odległości 2 otrzymały kolory różniące się o co najmniej q (zob. też przegląd wyników w zakresie $L(p, q)$-etykietowania w [24, 104]).

Definicja 5.5 (Problem minimalnej ξ-rozpiętości grafu)**.** *Dla danego grafu G i funkcji $\xi\colon E(G) \to \mathbb{N}_+$ należy znaleźć $\mathrm{sp}(G, \xi)$ jako najmniejszą liczbę $k \in \mathbb{N}_+$ taką, że istnieje ξ-pokolorowanie c grafu G spełniające $\max_{u \in V(G)} c(u) \leq k$.*

Oprócz rozpiętości, stanowiącej odpowiednik liczby chromatycznej, rozważa się w literaturze dla wyżej wymienionych modeli kolorowania również tzw. rozpiętość krawędziową jako lokalne kryterium optymalizacji [60, 103]

Definicja 5.6 (Problem minimalnej krawędziowej ξ-rozpiętości grafu)**.** *Dla danego grafu G i funkcji $\xi\colon E(G) \to \mathbb{N}_+$ należy znaleźć $\mathrm{esp}(G, \xi)$ jako najmniejszą liczbę $k \in \mathbb{N}_+$ taką, że istnieje ξ-pokolorowanie c grafu G spełniające $\max_{\{u,v\} \in E(G)} |c(u) - c(v)| \leq k$.*

W pracy [F6] przedstawiono szereg wyników związanych z krawędziową ξ-rozpiętością grafu. W szczególności poprzez redukcję wielomianową z NP-zupełnego problemu podziału zbioru pokazano, że problem ten pozostaje trudny nawet dla podkubicznych grafów zewnętrznie planarnych. Dowiedziono również, że dla kaktusów tj. dla grafów niezawierających krawędzi incydentnych do wielu cykli istnieje algorytm zwracający optymalne pokolorowanie, zgodne z $\mathrm{esp}(G, \xi)$ i działający w czasie $O(n \log n)$.

Praca [F7] zawiera wyniki dotyczące ξ-rozpiętości grafu dla grafów podkubicznych. Pokazano w niej, że problem nawet w tym ograniczonym przypadku pozostaje silnie NP-trudny, w odróżnieniu od klasycznego problemu kolorowania grafów, a nawet λ-szkieletowego kolorowania grafów [64], które można rozwiązać w czasie wielomianowym. Dowód polega na redukcji ze znanego NP-trudnego problemu Not-All-Equal 3-SAT [91]. Co więcej, okazuje się, że przy założeniu, że $\mathsf{P} \neq \mathsf{NP}$ nie może istnieć algorytm $(\frac{3}{2} - \varepsilon)$-przybliżony dla problemu ξ-rozpiętości grafów podkubicznych z funkcją ξ przyjmującą co najwyżej dwie wartości.

Jeśli funkcja ξ przyjmuje co najwyżej dwie wartości, to istnieje za to algorytm $\frac{3}{2}$-przybliżony dla powyższego problemu działający w czasie $O(n + m)$. Dla dowolnej funkcji ξ istnieje z kolei algorytm 2-przybliżony, również o złożoności czasowej $O(n+m)$. Co ciekawe, jeśli założymy, że graf indukowany przez krawędzie o maksymalnych wagach ξ tworzą graf spójny, to w pracy pokazano, że odpowiednio istnieje algorytm dokładny, działający w czasie $O(n^2)$ (gdy ξ przyjmuje co najwyżej dwie wartości) i algorytm $\frac{4}{3}$-przybliżony (dla dowolnych ξ) o złożoności czasowej $O(n + m)$.

Innym problemem pokrewnym kolorowaniu szkieletowemu i ξ-kolorowaniu jest kolorowanie kontrastowe, wprowadzone w [55]:

Definicja 5.7 (Pokolorowanie kontrastowe grafu)**.** *Dla danego grafu G i skończonego zbioru $T \subseteq \mathbb{N}$ $(0 \in T)$ funkcja $c\colon V(G) \to \mathbb{N}_+$ jest* pokolorowaniem kontrastowym grafu G dla zbioru T*, gdy dla każdej krawędzi $\{u, v\} \in E(G)$ zachodzi $|c(u) - c(v)| \notin T$.*

[9] Konstrukcja implikuje, że i-te drzewo Fibonacciego pokolorowane na 2 kolory ma dokładnie F_i więcej wierzchołków w jednym z kolorów.

Również dla tego problemu można zdefiniować minimalną rozpiętość kontrastową $\mathrm{sp}(G,T)$ i minimalną krawędziową rozpiętość kontrastową $\mathrm{esp}(G,T)$ [35].

Poprzednie prace badawcze nad $\mathrm{esp}(G,T)$ polegały głównie na poszukiwaniu zależności między $\mathrm{esp}(G,T)$ a $\mathrm{sp}(G,T)$ (rozpiętością krawędziową, zdefiniowaną analogicznie jak dla ξ-kolorowania) dla wybranych klas grafów i zbiorów T [69, 90, 107].

Praca [F8] została poświęcona związkom między krawędziową rozpiętością kontrastową a kolorowaniem cyrkularnym (*circular coloring*), zdefiniowanym w [100]. W tym celu wprowadzono operację $\odot$ dla liczby $d \in \mathbb{N}_+$ i zbioru $T \subseteq \mathbb{N}$ zdefiniowaną jako $d \odot T := \{0 \leq t \leq d(\max T + 1) \colon d|t \Rightarrow t/d \in T\}$. W pracy przedstawiono szereg podstawowych zależności między $\mathrm{esp}(G, d \odot T)$ a $\mathrm{esp}(G,T)$ i $\mathrm{sp}(G,T)$.

Głównym wynikiem pracy [F8] jest są twierdzenia charakteryzujące $\mathrm{esp}(G,T)$ dla zbioru $T = d \odot \{0\} = \{0, 1, \ldots, d-1\}$. Kolorowanie dowolnego grafu G ze zbiorem $d \odot \{0\}$ jest zatem szczególnym przypadkiem ξ-kolorowania z $\xi(e) = d - 1$ dla wszystkich $e \in E(G)$. Dla tego przypadku dowiedziono, że

Twierdzenie 5.5. *Dla dowolnego grafu G i dowolnej liczby $d \in \mathbb{N}_+$ zachodzi*

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

Jeśli $\chi_c(G) = k/d$ dla pewnego $d \in \{1, 2, \ldots, k\}$ to $\chi_c(G) = 1 + \mathrm{esp}_{d\odot\{0\}}(G)/d$.

Ta zależność pozwoliła na rozstrzygnięcie hipotezy dotyczącej kolorowania kontrastowego potęg cykli C_n^p, postawionej w [107]:

Twierdzenie 5.6. *Dla dowolnych $n, d, p \in \mathbb{N}_+$ zachodzi*

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

Autorzy hipotezy dowiedli również w [107], że powyższe twierdzenie jest prawdziwe, gdy $p \geq (q - p\lfloor q/p \rfloor)d$. Wykorzystując Twierdzenie 5.5 oraz dowodząc, że zawsze zachodzi $\chi_c(C_n^p) = n/q$ pokazano w [F8], że twierdzenie to zachodzi również w przypadku ogólnym.

5.6. Gry chromatyczne

Podstawowa gra chromatyczna, zaproponowana w [50] i wprowadzona na nowo w [13] w formalnym kontekście teorii grafów jest popularnym tematem badawczym, który doczekał się wielu wariantów dla różnych klas grafów i reguł ruchu (zob. przegląd wyników w [9]). W podstawowej wersji para graczy, Alicja i Bogdan, ma do dyspozycji pulę kolorów $\{1, \ldots, k\}$ i wykonuje ruchy naprzemiennie. Każdy ruch polega na przydzieleniu dotychczas niepokolorowanemu wierzchołkowi jednego z kolorów z puli tak, aby zachować poprawność kolorowania tj. aby żadne dwa sąsiednie pokolorowane wierzchołki nie miały tego samego koloru. Celem Alicji jest pokolorowanie całego grafu, celem Bogdana zaś doprowadzenie do sytuacji, w której gracz nie będzie mógł wykonać żadnego legalnego ruchu. Rozgrywaną liczbą chromatyczną (*game chromatic number*) $\chi_g(G)$ oznaczamy najmniejszą liczbę k, dla której Alicja ma strategię wygrywającą tj. niezależnie od strategii Bogdana zawsze osiągane jest pokolorowanie całego grafu.

W oszacowaniach związanych z grami chromatycznymi pojawia się czasem pojęcie tzw. liczba kolorowania (*coloring number*) i jego uogólnienia:

Definicja 5.8. *Niech G będzie grafem a $\prec$ porządkiem liniowym na $V(G)$. Niech $N_G^-(v, \prec) = \{u \in V(G) \colon \{u, v\} \in E(G) \wedge u \prec v\}$ oznacza sąsiedztwo wsteczne v w G. Wówczas*

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

Definicja 5.9. *Niech G będzie grafem a $\prec$ porządkiem liniowym na $V(G)$. Dla dowolnego $r \in \mathbb{N}_+$ niech*

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

oznacza k-te sąsiedztwo wsteczne v w G. Wówczas

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

Przykładowo, w pracy [74] pokazano, że dla dowolnego grafu planarnego G zachodzi $\chi_g(G) \leq 4\,\mathrm{col}_2(G)+1$. Dla grafów ogólnych dowiedziono, że zachodzi nierówność $\chi_g(G) \leq \chi(G)(\mathrm{col}_2(G)+1)$ [9]. Można również wskazać, że parametr $\mathrm{col}_2(G)$ pojawia się w dowodach i ograniczeniach takich parametrów grafowych jak acykliczna liczba chromatyczna [75] czy zorientowana rozgrywana liczba chromatyczna [73].

W pracy [G1] badano właściwości parametru $\mathrm{col}_2(G)$. Po pierwsze, poprawiono dla przypadku $r = 2$ ograniczenie $\chi(G) \leq \mathrm{col}_r(G) \leq \Delta(G)(\Delta(G)-1)^{r-1}+1$ przedstawione przez Kiersteada i Kostochkę w [72] dowodząc, że:

Twierdzenie 5.7. *Dla każdego grafu G zachodzi* $\mathrm{col}_2(G) \leq \frac{1}{2}\Delta(G)(\Delta(G)-1)+2$.

Po drugie, pokazano, że istnieje duża klasa grafów, dla których rzeczywiście zachodzi zależność $\mathrm{col}_2(G) = \Theta(\Delta(G)^2)$:

Twierdzenie 5.8. *Dla każdego regularnego grafu G niezawierającego C_3 ani C_4 zachodzi*

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

Dowód polegał na odpowiednio dokładnym zliczaniu trójek postaci (j,i,k) takich, że $j < i < k$ oraz $\{v_i, v_j\}, \{v_j, v_k\} \in E(G)$. W ten sposób możliwe stało się oszacowanie od dołu sumy wszystkich „drugich stopni wstecznych" (wielkości sąsiedztw $N_G^-(v, 2, \prec)$) dla dowolnego porządku $\prec$ jako pewnej funkcji liczby krawędzi grafu G – a zatem odpowiednie oszacowanie maksimum. Warto zwrócić uwagę, że to oznacza, że dla takich grafów $\mathrm{col}_2(G)$ jest złym oszacowaniem $\chi_g(G)$, ponieważ wiadomo, że $\chi_g(G) \leq \Delta(G)+1$ dla każdego grafu G.

Po trzecie, został zaproponowany wielomianowy algorytm obliczania wartości $\mathrm{col}_2(G)$ dla grafów podkubicznych tj. o $\Delta(G) \leq 3$. Algorytm ten opiera się o rekurencyjne znajdowanie i usuwanie małych fragmentów grafu, dla których można dowieść, że ich usunięcie zachowuje wartość współczynnika bez zmian.

Praca [G2] została poświęcona grze nazwanej *nieskończoną grą chromatyczną* (*infinite graph coloring game*). Różni się ona od zwykłej gry chromatycznej założeniem, że strony gry nie mają ustalonego, skończonego zbioru kolorów, ale raczej nieskończony. Tak jak w wariancie podstawowym, Alicja dąży do pokolorowania grafu jak najmniejszą ilością kolorów, natomiast Bogdan przeciwnie. Odpowiednio, *nieskończona rozgrywana liczba chromatyczna* $\chi_g^\infty(G)$ jest zdefiniowana jako liczba użytych przez graczy kolorów, gdy stosują oni strategie optymalne.

Oczywistym ograniczeniem jest $\chi_g^\infty(G) \geq 1 + \lfloor \frac{n(G)}{2} \rfloor$, ponieważ zaczyna Alicja a Bob może w każdym ruchu użyć nowego koloru. Z podstawowych ograniczeń górnych dowiedziono, że

Twierdzenie 5.9. *Dla każdego grafu G zachodzi*

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

Pokazano również, że problem nie jest wcale trywialny z punktu widzenia strategii istnieje nieskończona rodzina grafów, dla której optymalną strategią Bogdana wcale nie jest każdorazowe używanie nowego koloru – ale czasami bardziej opłaca się dobrze umiejscowione powtórzenie już wcześniej wykorzystanego koloru.

Następnie, przedstawiono szereg wyników dla szczególnych klas grafów.

Twierdzenie 5.10. *Dla grafu G spełniającego* $\Delta(G) \leq \frac{1}{3}(n(G)-1)$ *zachodzi* $\chi_g^\infty(G) = \lfloor \frac{1}{2}n(G) \rfloor + 1$.

Pozwoliło to na otrzymanie pełnych wyników dla grafów podkubicznych.

Twierdzenie 5.11. *Dla grafów G z $\Delta(G) \leq 3$ zachodzi*

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

Pełne wyniki otrzymano również dla grafów k-dzielnych pełnych:

Twierdzenie 5.12. *Niech l będzie liczbą nieparzystych liczb w zbiorze $\{r_1, r_2, \ldots, r_k\}$. Wówczas*

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

Najważniejszym wynikiem pracy [G2] jest pokazanie, że istnieje strategia dla Bogdana dowodząca, że dla wszystkich grafów zachodzi $\chi_g^{\infty}(G) \leq n(G) - \alpha'(\bar{G})$ oraz strategia dla Alicji, że dla wszystkich grafów o nieparzystej liczbie wierzchołków zachodzi $\chi_g^{\infty}(G) \geq n(G) - \alpha'(\bar{G})$.

5.7. Rekonstrukcja hipergrafów

Jedną z operacji rozważanych w teoretycznych badaniach nad hipergrafami jest tzw. 2-podział hipergrafu. Operacja ta polega na zastąpieniu każdej hiperkrawędzi w grafie przez klikę i dodatkowo na usunięciu powstałych zdublowanych krawędzi. Analizy hipergrafów poprzez powstałe w wyniku takich operacji grafy znalazły zastosowania m.in. w biologii obliczeniowej [42], teorii języków [52] i optymalizacji kompilatorów [88].

Ponieważ jednak to podejście oznacza utratę informacji o krotnościach krawędzi, można zmodyfikować ten problem pozostawiając wielokrotne krawędzie lub, równoważnie, naturalne wagi na krawędziach, i zdefiniować następujący problem rekonstrukcji hipergrafów:

Definicja 5.10 (Problem decyzyjny rekonstrukcji hipergrafów)**.** *Czy dla danego grafu G oraz funkcji wag $w\colon E(G) \to \mathbb{N}_+$ istnieje hipergraf H taki, że (G, w) jest jego ważonym 2-podziałem?*

Dla optymalizacyjnej wersji tego problemu tj. gdy szukamy odpowiedniego hipergrafu z minimalną liczbą hiperkrawędzi dowiedziono, że należy on do klasy FPT ze względu na parametryzację wielkością wyjścia [41]. Jeśli natomiast mamy problem optymalizacyjny, ale bez podanej funkcji wag, to wykazano, że problem jest NP-trudny dla grafów planarnych [79], niezawierających K_4 [81] i split grafów [101], ale istnieją algorytmy FPT dla grafów planarnych i grafów bez K_4, gdy parametrem jest rozmiar wyjścia [47].

W pracy [H1] pokazano, że problem decyzyjny jest również NP-zupełny dla grafów pełnych nawet, gdy wagi krawędzi należą do zbioru $\{1, 2\}$. Dowód polegał na odpowiedniej redukcji ze znanego NP-trudnego problemu krawędziowego 3-kolorowania spójnych grafów kubicznych [58].

W ramach wyników konstruktywno-algorytmicznych podano natomiast, że możliwe jest rozwiązanie problemu i znalezienie odpowiedniego hipergrafu w czasie wielomianowym, dla częściowych 2-drzew, dla grafów 2-zdegenerowanych i dla grafów o $\Delta(G) \leq 4$.

6. Osiągnięcia dydaktyczne i organizacyjne

6.1. Dydaktyka

6.1.1. Kursy przygotowane przez habilitanta

- Algorytmy tekstowe (wykład, ćwiczenia) – na Uniwersytecie Jagiellońskim.
- Język programowania: C++ (wykład, laboratorium) – na Uniwersytecie Jagiellońskim.
- Język programowania: Python (wykład, laboratorium) – na Uniwersytecie Jagiellońskim.
- Programowanie mobilne (laboratorium) – na Uniwersytecie Jagiellońskim.
- Języki programowania (wykład, laboratorium) – na Politechnice Gdańskiej.

6.1.2. Inne zajęcia dydaktyczne

- Programowanie współbieżne (laboratorium) – na Uniwersytecie Jagiellońskim.
- Systemy rozproszone (laboratorium) – na Uniwersytecie Jagiellońskim.
- Podstawy programowania (laboratorium) – na Politechnice Gdańskiej.

- Algorytmy i struktury danych (laboratorium) – na Politechnice Gdańskiej.
- Algorytmy optymalizacji dyskretnej (laboratorium) – na Politechnice Gdańskiej.
- Podstawy analizy algorytmów (ćwiczenia) – na Politechnice Gdańskiej.
- Elementy bioinformatyki (laboratorium) – na Politechnice Gdańskiej.
- Bioinformatyka (laboratorium) – na Politechnice Gdańskiej.
- Modelowanie i symulacja systemów (laboratorium) – na Politechnice Gdańskiej.

6.2. Praca promotorska

6.2.1. Promowane prace magisterskie na Uniwersytecie Jagiellońskim

1. Adrian Siwiec, *Rozpoznawanie i kolorowanie grafów doskonałych*, 2020.
2. Paweł Palenica, *Algorytmy kompresji grafów losowych*, 2020.
3. Wojciech Grabis, *Implementacja wybranych modeli makroekonomicznych DSGE*, 2022.
4. Michał Stobierski, *Szybkie obliczanie maksymalnych przepływów metodami kombinatorycznymi*, 2022.

6.2.2. Promowane prace licencjackie na Uniwersytecie Jagiellońskim

1. Mateusz Górski, *Uogólnienie liczby Turána*, 2020.
2. Marcin Serwin, *Przegląd algorytmów operujących na kografach*, 2020.
3. Krzysztof Michalik, *O λ-szkieletowym kolorowaniu klik z drzewem w szkielecie w czasie liniowym*, 2021.
4. Mikołaj Twaróg, *Wielomianowe schematy aproksymacyjne dla problemów NP-trudnych na grafach planarnych*, 2021.
5. Mateusz Pach, *Protokoły rozproszonego konsensusu*, 2022.
6. Inka Sokołowska, *Szeregowanie zadań z blokowym grafem ograniczeń*, 2022.

7. Uzyskane granty

Obecnie jestem kierownikiem grantu „Charakteryzacja treści informacyjnej struktur grafowych” (nr 2020/39/D/ST6/00419) przyznanego w ramach konkursu SONATA-16 przez Narodowe Centrum Nauki.

Literatura

[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.