Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki

Paweł Prałat

Grafy proteuszowe

Praca doktorska
napisana pod kierunkiem
prof. dr hab. Tomasza Łuczaka

Poznań 2004

Pragnę podziękować mojemu promotorowi,
prof. dr hab. Tomaszowi Łuczakowi,
za inspirację, cierpliwość i okazaną pomoc.

Spis treści

Proteusz: *Światowych chwytów jeszcześ nie zaniechał*
Tales: *Postać swą zmieniać jest wciąż twą uciechą*
Johann Wolfgang von Goethe *"Faust"*

Wstęp

W ciągu ostatnich kilku lat intensywnie badane są istniejące w świecie systemy zależności, takie jak sieć wzajemnych powiązań ludzkich [18, 27], aktorów filmowych [26], współpracujących ze sobą naukowców [16] czy graczy futbolowych [25]. Znajomość własności omawianych sieci jest dla nas istotna, dzięki niej poznajemy naturę rozprzestrzeniania się różnych chorób, przekazywania wiadomości wśród społeczeństwa czy propagacji danych w różnych sieciach radiowych np. w sieci telefonii komórkowej. Na szczególne wyeksponowanie zasługuje sieć internetowa i odpowiadający jej graf internetowy [2, 13, 20], któremu ostatnio poświęca się wiele uwagi. Tematyki tej dotyczy wiele artykułów, istnieje nawet niezależne czasopismo. Zainteresowanie to spowodowane jest nie tylko dlatego, że to wdzięczny i ciekawy obiekt badań, ale dzięki wiedzy o tym jakie własności posiada sieć internetowa możemy efektywniej projektować algorytmy działające w tej sieci, np. algorytm wyszukiwania stron internetowych (ang. *search engine*). Naturalne więc jest, że duże firmy takie jak *Microsoft*, *AltaVista* czy *IBM* włączają się do badań i powodują przyspieszenie tempa prac.

Niezależnie od badań nad strukturą sieci internetowej, naukowcy wprowadzają i badają modele grafów losowych posiadających podobne własności jak graf internetowy [7, 8, 10, 19]. W niniejszej rozprawie proponuję nowy model, graf proteuszowy. Zasadniczą własnością odróżniającą graf proteuszowy od innych pojawiających się w literaturze modeli grafu internetowego jest fakt, że jego krawędzie, odpowiadające połączeniom w sieci, mogą pojawiać się, ale także zanikać. W przeciwieństwie do niemal wszystkich modeli grafu internetowego, graf proteuszowy może być niespójny; podkreślmy, że graf internetowy, wbrew powszechnemu mniemaniu, również składa się z wielu różnych składowych [13]. Inną istotną cechą grafu proteuszowego jest to, że przy właściwym doborze parametrów, model ten pokrywa się ze standardowym modelem grafu losowego $G(n,p)$, czy, po pewnych modyfikacjach, modelem sieci typu *peer-to-peer* – jest on zatem uogólnieniem znanych w literaturze grafów losowych. Wprowadzony model jest interesujący nie tylko jako model sieci internetowej, ale wydaje się atrakcyjny również z teoretycznego punktu widzenia. Posiada on bowiem bogatą strukturę zależności oraz, w przeciwieństwie do innych modeli grafów losowych, definiujący go proces proteuszowy jest łańcuchem Markowa, który znajduje się w stanie stacjonarnym.

Pierwszy rozdział niniejszej pracy zawiera dość swobodnie potraktowane omówienie rzeczywistych sieci. Okazuje się, iż posiadają one pewne wspólne własności, które odróżniają je od innych tego rodzaju struktur. Szczególnie dużo uwagi poświęcam grafowi internetowemu i jego najbardziej charakterystycznym cechom jak rozkładowi stopni, spójności czy średnicy. W końcowej części rozdziału wprowadzam formalną definicję procesu proteuszowego $\mathfrak{P}_n(d,\eta)$ oraz grafu proteuszowego $\mathcal{P}_n(d,\eta)$.

W drugim rozdziale badam podstawowe własności grafu proteuszowego. Pokazuję m.in., że $\mathcal{P}_n(d,\eta)$ (w przypadku, gdy $\eta \in (0,1)$) posiada podobną strukturę do grafu o wierzchołkach ze zbioru $[n]$, w którym dwa dowolne wierzchołki i,j, $1 \leq i < j \leq n$ są połączone krawędzią z prawdopodobieństwem $(1-\eta)\frac{d}{n}(\frac{j}{i})^{\eta}$, niezależnym dla każdej pary wierzchołków. Wykazuję ponadto, że rozkład stopni w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ jest rozkładem potęgowym (potęga zależy od parametru η). Własność ta pozwala traktować $\mathcal{P}_n(d,\eta)$ jako model sieci internetowej.

Pokazuję również, że prawdopodobieństwo, iż ustalony wierzchołek jest wierzchołkiem izolowanym (podobnie jest w przypadku oczekiwanego stopnia) zależy od miejsca w grafie, w którym się on znajduje. W "najgorszym" położeniu znajdują się wierzchołki w środku grafu proteuszowego. Dla nich prawdopodobieństwo bycia wierzchołkiem izolowanym jest największe. Możemy więc powiedzieć, że część wierzchołków w grafie proteuszowym przeżywa "kryzys wieku średniego".

W trzecim rozdziale zajmuję się spójnością grafu proteuszowego w przypadku, gdy $\eta \in (0,1)$ oraz $\eta = 0$. Okazuje się, że struktura zależności w omawianym grafie wpływa znacząco na próg spójności. W rozdziale tym badaniom poddano rozmiar największej składowej oraz jej średnicę. Korzystając m.in. z teorii procesów gałązkowych pokazuję, że średnica największej składowej w grafie proteuszowym wynosi $\Theta(\log n)$.

W końcowej części rozdziału znajduję rozkład asymptotyczny "czasu powrotu" do stanu spójności grafu proteuszowego. Jest to przykład wielkości związanej z procesem proteuszowym, dla której trudno byłoby znaleźć odpowiednik w innych modelach struktur losowych. Rozpatruję bowiem zachowanie grafu proteuszowego ponad progiem dla pewnej określonej własności (w tym przypadku spójności) i zastanawiam się jak szybko graf proteuszowy, który w trakcie procesu proteuszowego utracił "typową" dla siebie własność, ponownie własność tę odzyska.

W ostatnim rozdziale opisano wyniki przeprowadzonych symulacji. Ze względu na możliwość modelowania sieci internetowej, skupiłem się na dwóch szczególnych wartościach parametru η ($\eta_{\text{out}} = 0{,}59$ i $\eta_{\text{in}} = 0{,}91$) oraz w przypadku, gdy średni stopień w grafie proteuszowym nie jest zbyt duży ($d = 10$).

1 Definicje

1.1 Mały wielki świat

Chcąc odpocząć od codzienności wybieramy się w dalekie podróże do odległych zakątków kraju czy świata. Mamy nadzieję, że będziemy się tam czuli anonimowo i beztrosko spędzimy wolny czas. Ogromne zdziwienie nas ogarnia, gdy spotykamy tam kolegów z pracy, rodzinę czy znajomych. Myślę, że każdemu z nas choć raz przytrafiła się taka sytuacja. Najczęściej padają wtedy słowa: "Jaki ten świat jest mały!".

Istotnie, jak szacuje *The United Nations Department of Economic and Social Affairs*, w dniu 12 września 1999 roku liczba mieszkańców naszej planety przekroczyła sześć miliardów, niemniej jednak jeżeli spojrzymy na sieć wzajemnych powiązań ludzkich (ang. *social network*) odległość między osobami znajdującymi się w tej sieci jest zaskakująco mała [18, 27].

Pierwszą osobą badającą sieć wzajemnych powiązań ludzkich był Stanley Milgram z Harvard University [23], który w latach sześćdziesiątych przeprowadził prosty, aczkolwiek bardzo ciekawy, eksperyment. Sporządził on listy i zaadresował je do znajomych maklerów giełdowych mieszkających na terenie Bostonu (stan Massachusetts). Listy te wysłał do losowo wybranych osób w stanie Nebraska z prośbą o ich dostarczenie. Listów nie należało jednak wysyłać bezpośrednio do adresatów, można je było przekazywać wyłącznie osobom, które dobrze znamy. Osoby biorące udział w doświadczeniu, nie znające adresatów musiały podjąć decyzję do kogo przesłać list, aby "zbliżyć" się do celu. Zazwyczaj wybierały one rodzinę bądź znajomych mieszkających w stanie Massachusetts (albo w samym Bostonie) lub kogoś z branży finansowej, mając nadzieję, iż osoba ta zna adresata. Okazało się, że znacząca liczba listów przygotowanych przez Milgrama dotarła pod wskazany adres. Co więcej, średnio wystarczyło tylko sześć przesłań listu by list z Nebraski dotarł do adresata w Bostonie. Milgram przypuszczał, że podobnego wyniku można by się spodziewać, gdyby chciano

skontaktować ze sobą dwie wybrane osoby przebywające w dowolnym miejscu na świecie. Średnia odległość w tym przypadku będzie zapewne nieco większa niż sześć, niemniej liczba ta zakorzeniła się na tyle mocno, że do dzisiaj termin *"six degrees of separation"* [28], oznacza, że dowolne dwie osoby z danej grupy są "połączone ze sobą" względnie krótkim łańcuchem, w którym sąsiadujące osoby są w pewnym sensie "blisko siebie".

Własność tę ma również wiele innych, w miarę naturalnie zdefiniowanych sieci. W sieci aktorów, w której dwie osoby sąsiadują ze sobą, gdy zagrały w tym samym filmie [26], każda para aktorów (hollywoodzkich!) jest połączona łańcuchem o długości nie większej niż osiem. Podobne badania przeprowadzono dla sieci współpracujących ze sobą naukowców [16] oraz graczy futbolowych [25]. Kolejną, bardzo ważną i często badaną, tego rodzaju siecią jest sieć internetowa, której poświęcono osobny rozdział 1.2.

Omawiane sieci posiadają stosunkowo mały średni stopień, który, jak się wydaje, nie rośnie wraz z rozmiarem sieci. Prostym modelem spełniającym ten wymóg, jest duża składowa grafu losowego $G(n,p)$, dla $np = d > 1$, której średnica jest rzędu $\log n$ (a zatem jest mała nawet dla dużych grafów).

We wspomnianych sieciach zauważa się również tendencję do grupowania się elementów w nich występujących. Przyjaciele naszych przyjaciół są często i naszymi przyjaciółmi, naukowcy z którymi współpracujemy niejednokrotnie współpracują również między sobą. Podobnie jest w sieci aktorów czy graczy futbolowych. Własność tę mierzymy definiując współczynnik skupienia (ang. *clustering coefficient*).

Definicja 1. *Niech G będzie dowolnym grafem. Dla dowolnego wierzchołka $v \in V(G)$ przez Γ_v oznaczmy graf jego sąsiadów*

$$\begin{aligned} V(\Gamma_v) &= \{x \in V(G) : \{v, x\} \in E(G)\}\ , \\ E(\Gamma_v) &= \{\{x, y\} \in E(G) : x, y \in V(\Gamma_v)\}\ . \end{aligned}$$

Współczynnik skupienia C^v wierzchołka $v \in V(G)$ definiujemy jako

$$C^v = \frac{|E(\Gamma_v)|}{\binom{deg(v)}{2}},$$

a przez C^G będziemy oznaczać współczynnik skupienia dla grafu G będący średnim współczynnikiem skupienia w tym grafie.

$$C^G = \frac{\sum_{v \in V(G)} C^v}{|V(G)|}$$

Zauważmy, że dla dowolnego grafu G współczynnik skupienia $C^G \in [0, 1]$, przy czym dla grafu pustego mamy $C^G = 0$, gdy G jest grafem pełnym, wtedy $C^G = 1$. Dla grafu losowego $G(n, p)$, w którym krawędzie pojawiają się niezależnie od pozostałych z prawdopodobieństwem $p = d/n$, wartość oczekiwana C^G wynosi $p = d/n$ i dąży do zera gdy $n \to \infty$.

Wartości współczynników skupienia dla różnych sieci obliczone przez Watts'a i Strogatz'a [29] przedstawiają się następująco:

rodzaj sieci	n	d	C	C_{los}
sieć aktorów	225 226	3,65	0,79	0,00027
układ nerwowy dżdżownicy	282	2,65	0,28	0,05
sieć elektryczna w USA	4 941	18,7	0,08	0,0005

W powyższej tabeli n oznacza liczbę elementów znajdujących się w sieci, d średni stopień sieci, a C to współczynnik skupienia. W ostatniej kolumnie umieszczono, dla porównania, współczynnik skupienia C_{los}, który wystąpiłby w grafie losowym mającym identyczną liczbę wierzchołków oraz średni stopień. Powyższe wyniki sugerują, że graf losowy (pomimo, że jego średnica jest względnie mała) nie jest idealnym modelem dla rzeczywistych sieci, których współczynnik skupienia jest znacząco większy niż $O(n^{-1})$.

Obecnie dużo uwagi poświęca się badaniom istniejących sieci (grafów) typu ‘‘small world”. Chociaż trudno jest podać ich formalną definicję, większość grafów określanych tą nazwą posiada trzy następujące własności:

- stały (tzn. niezależny od liczby wierzchołków n) średni stopień grafu,
- średnia odległość pomiędzy wierzchołkami jest rzędu co najwyżej $\log n$

$$\bar{\delta}_G = O(\log n) \ ,$$

- współczynnik skupienia jest znacząco większy niż w grafie losowym

$$C^G \gg n^{-1} \ .$$

1.2 Graf internetowy i jego własności

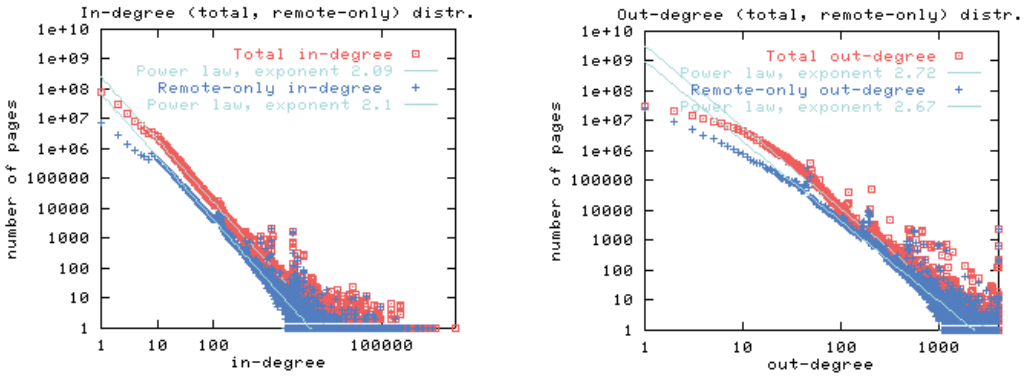
Szczególnym rodzajem sieci typu “small world” jest sieć internetowa, lub inaczej *graf internetowy*. Jest to graf skierowany $\mathcal{G}$, którego wierzchołki $V(\mathcal{G})$ odpowiadają stronom internetowym, natomiast łuki $A(\mathcal{G})$ odpowiadają odsyłaczom pomiędzy tymi stronami. Łuk $(x, y) \in A(\mathcal{G})$ wtedy i tylko wtedy, gdy na stronie odpowiadającej wierzchołkowi $x \in V(\mathcal{G})$ znajduje się odsyłacz do strony odpowiadającej wierzchołkowi $y \in V(\mathcal{G})$.

Graf internetowy jest jednym z najintensywniej badanych obiektów współczesnej informatyki. Każdego miesiąca pojawiają się dziesiątki artykułów na ten temat, istnieje nawet niezależne czasopismo poświęcone własnościom grafu internetowego i różnym jego modelom. Poniższy opis grafu internetowego oparty został na artykule [13], którego autorzy badali strukturę sieci internetowej z 1999 roku (składającą się podówczas z około 200 mln stron WWW oraz 1,5 mld odsyłaczy).

Rozkład stopni

Najważniejszą własnością grafu internetowego jest rozkład stopni. To właśnie on wyróżnia go spośród innych naturalnych przykładów grafów. Okazuje się, że rozkład ten jest rozkładem potęgowym. Powyższa własność zachodzi zarówno dla stopni wejściowych jak i wyjściowych (wykładniki potęgi różnią się dla obu tych przypadków).

$$P(deg^{IN}(v) = k) \sim k^{-2,1} \text{ dla każdego } v \in V(\mathcal{G})$$
$$P(deg^{OUT}(v) = k) \sim k^{-2,72} \text{ dla każdego } v \in V(\mathcal{G})$$



Niemal identyczne wyniki otrzymano analizując dane pochodzące z maja 1999 roku jak i z września tegoż samego roku. Co więcej, podobne wykładniki dla rozkładu stopni wierzchołków otrzymał wcześniej Kumar *et. al.* [20] badając pięciokrotnie mniejszą sieć pochodzącą z 1997 roku. Można zatem przypuszczać, iż pomimo szybkiego rozwoju sieci internetowej rozkład stopni pozostaje niezmienny i w pewien sposób powiązany z mechanizmem tworzenia omawianej sieci, choć związek ten do dzisiaj pozostaje dość niejasny.

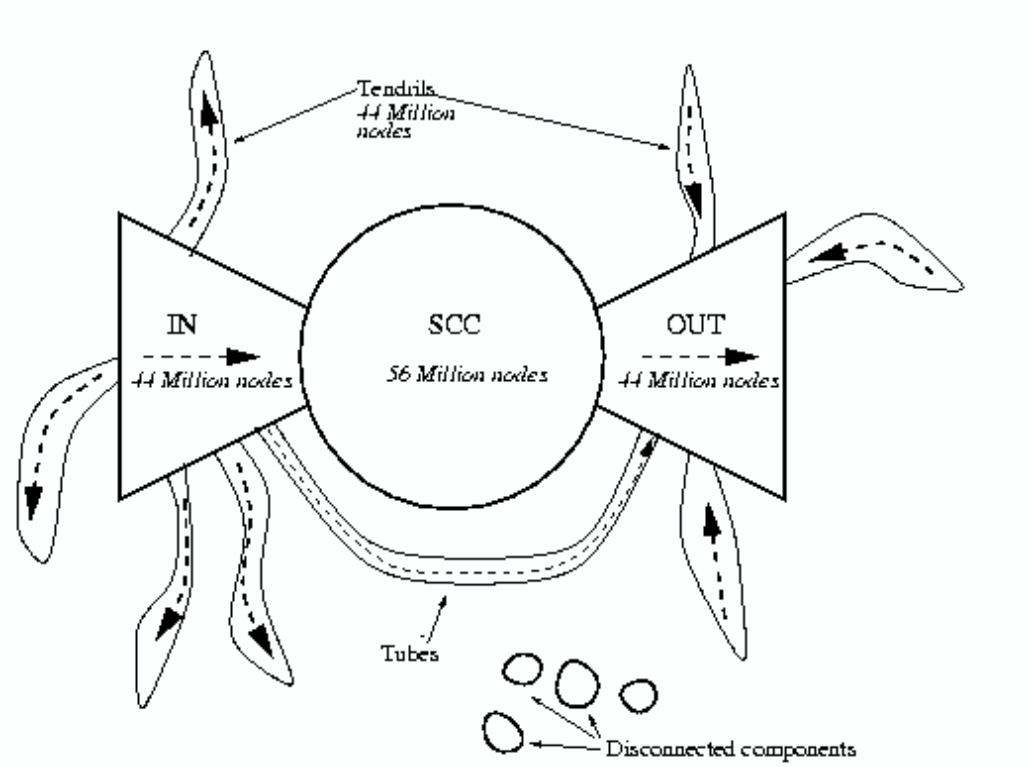
Słaba spójna składowa

Największa słaba spójna składowa w grafie internetowym składa się z 186 mln wierzchołków, co stanowi 91% wszystkich wierzchołków. Mogłoby się wydawać, iż duży rozmiar największej składowej jest wynikiem istnienia wierzchołków o dużym stopniu wejściowym. Tak jednak nie jest. Nawet jeżeli usuniemy wszystkie wierzchołki o stopniu wejściowym większym lub równym 5, graf nadal będzie zawierał słabą spójną składową zawierającą 59 mln wierzchołków. Własność ta jest niezwykle ważna dla projektowania i analizy algorytmów działających na tego typu strukturach. Poniższa tabela pokazuje rozmiar największej słabej spójnej składowej (w mln) po usunięciu wierzchołków o stopniu większym lub równym k.

k	1000	100	10	5	4	3
rozmiar składowej (w mln)	177	167	105	59	41	15
rozmiar składowej (w %)	87	82	52	29	20	7

Silnie spójna składowa

Największa silnie spójna składowa w grafie internetowym składa się z 56 mln wierzchołków, co stanowi 28% wszystkich wierzchołków. Jeżeli zestawimy tę wartość z rozmiarem słabej spójnej składowej możemy zadać pytanie: gdzie jest reszta wierzchołków? Dokładniejsza analiza grafu internetowego ukazuje nam jego charakterystyczną strukturę, określaną potocznie w literaturze jako "muszka" (ang. *"bow–tie"*).



Wierzchołki w grafie internetowym możemy podzielić na pięć klas:

- **SCC** – wierzchołki znajdujące się w największej silnie spójnej składowej,
- **OUT** – wierzchołki nie będące w SCC, do których można dojść z dowolnego wierzchołka znajdującego się w SCC,
- **IN** – wierzchołki nie będące w SCC, z których można dojść do dowolnego wierzchołka znajdującego się w SCC (a zatem i do dowolnego wierzchołka znajdującego się w OUT),

- **TENDRILS** – wierzchołki będące w największej słabej spójnej składowej, ale nie będące w żadnym ze zbiorów SCC, OUT, IN; mogą to być przykładowo wierzchołki do których można dojść z wierzchołka znajdującego się w IN, bądź wierzchołki z których można dojść do wierzchołka znajdującego się w OUT, bądź wierzchołki, które leżą na ścieżce od wierzchołka w IN do wierzchołka w OUT (**TUBES**),

- **DISCONNECTED** – wierzchołki poza największą słabą spójną składową.

Oto rozmiary poszczególnych klas:

klasa	rozmiar	rozmiar (w %)
SCC	56 463 993	27,7
IN	43 343 168	21,3
OUT	43 166 185	21,2
TENDRILS	43 797 944	21,5
DISCONNECTED	16 777 756	8,2
W sumie	203 549 046	100,0

Średnica i średnia odległość

Ze względu na wielkość grafu internetowego i złożoność problemu nie obliczono dokładnej wartości jego średnicy. Wielokrotne przeszukiwanie grafu wszerz (startując z losowo wybranego wierzchołka), doprowadziło jednak do kilku oszacowań.

Średnica SCC wynosi co najmniej 28, natomiast maksymalna długość najkrótszej skończonej ścieżki łączącej dowolne dwa wierzchołki wynosi co najmniej 503 lecz nie więcej niż 905.

Średnia odległość pomiędzy losową parą wierzchołków wynosi 16,12. Natomiast w wersji nieskierowanej (strzałki zamieniamy na krawędzie) średnia odległość wynosi 6,83. Analiza struktury tego grafu skłoniła Alberta *et. al.* [2], do przypuszczenia, że średnia odległość wierzchołków w grafie internetowym rośnie z grubsza jak $\bar{\delta}_{\mathcal{G}} = 0,35+2,06\log_{10} n$, gdzie n jest liczbą wierzchołków (stron WWW) tego grafu. Dla $n = 2\cdot 10^8$ (rozmiar rzeczywistej sieci w 1999 roku) $\bar{\delta}_{\mathcal{G}} \approx 17,45$.

1.3 Modele grafu internetowego

Niezależnie od badań nad strukturą sieci internetowej, w ostatnich latach opublikowano ponad sto artykułów poświęconych modelom grafów losowych mających własności podobne do grafu internetowego. W większości tych modeli, interesujący nas graf jest stanem pewnego łańcucha Markowa $\{G_t\}_{t=0}^{\infty} = \{(V_t, E_t)\}_{t=0}^{\infty}$, gdzie w t-tym kroku do grafu G_t dodajemy jeden wierzchołek łącząc go krawędziami z pozostałymi wierzchołkami zgodnie z pewną, przyjętą wcześniej przez autorów modelu, regułą [8, 19]. Niektóre z modeli dopuszczają również generowanie krawędzi pomiędzy już istniejącymi wierzchołkami w grafie [7, 10]. Zwróćmy uwagę, iż modele zdefiniowane w powyższy sposób zawsze generują graf spójny. Nie odzwierciedlają zatem zbyt wiernie właściwości sieci internetowej, która, jak pokazują badania, posiada wiele spójnych składowych.

Wprowadzony i badany w niniejszej pracy graf proteuszowy dopuszcza istnienie wielu składowych. Jego zasadniczą cechą, jest to, że w procesie proteuszowym, w trakcie którego generujemy graf proteuszowy, dopuszczamy zarówno dodawanie jak i usuwanie krawędzi. Prócz tego, w odróżnieniu od innych modeli grafu internetowego, podobnie jak w standardowym modelu grafu losowego $G(n,p)$, w czasie procesu proteuszowego liczba wierzchołków pozostaje stała; co więcej, proces ten pozostaje w stanie stacjonarnym co pozwala badać takie jego własności jak czas powrotu (ang. *recovery time*), czyli czas, w którym graf ponownie będzie posiadał typową własność, którą utracił w trakcie procesu.

1.4 Definicja grafu proteuszowego

W niniejszym rozdziale wprowadzimy pojęcie grafu i procesu proteuszowego. Zanim jednak przedstawimy formalną definicję, chcielibyśmy omówić krótko ideę, która jest prosta i intuicyjna. Rozważmy dowolny graf G składający się z n wierzchołków. W każdym kroku procesu wybieramy losowo jeden wierzchołek i usuwamy go z grafu G wraz z incydentnymi z nim krawędziami (odpowiada to sytuacji, w której losowa strona WWW zostaje usunięta z sieci internetowej). Następnie dodajemy nowy wierzchołek oraz łączymy go z pozostałymi wierzchołkami w grafie zgodnie z ustalonym rozkładem $\mathbf{X} = \mathbf{X}_{n-1}$ (w internecie pojawia się nowa strona, a na niej odsyłacze do innych stron). Istotny jest fakt, iż dopuszczamy w tym miejscu, aby rozkład $\mathbf{X}_{n-1}$ zależał od "wieku" wierzchołków pozostających w grafie (wydaje się to naturalne, że na nowopowstałych stronach częściej umieszczane są odsyłacze do starszych, znanych już dobrze stron internetowych).

Przejdźmy teraz do formalnej definicji. Niech $\mathbf{X}_{n-1} = (X_1, \ldots, X_{n-1})$ będzie $(n–1)$-wymiarową, nieujemną, całkowitoliczbową zmienną losową, G będzie dowolnym grafem o zbiorze wierzchołków $V(G) = [n] = \{1, 2, \ldots, n\}$, a σ będzie dowolną permutacją zbioru $[n]$. Rozważmy następujący jednorodny łańcuch Markowa $\{(\tilde{G}_k, \tilde{\sigma}_k, A_k)\}_{k=0}^{\infty}$, w którym stanami są trójki $(\tilde{G}_k, \tilde{\sigma}_k, A_k)$, gdzie $\tilde{G}_k$ jest grafem o wierzchołkach ze zbioru $[n]$, $\tilde{\sigma}_k : [n] \to [n]$ jest permutacją zbioru $[n]$, oraz $A_k \subseteq [n]$. Stanem początkowym procesu jest $(\tilde{G}_0, \tilde{\sigma}_0, A_0) = (G, \sigma, \emptyset)$. W k-tym kroku procesu wybieramy losowo wierzchołek $i \in [n]$, a $\tilde{\sigma}_k$ otrzymujemy przez przeniesienie wierzchołka i na koniec permutacji $\tilde{\sigma}_{k-1}$, tzn. dla $k > 0$

$$\tilde{\sigma}_k(j) = \begin{cases} \tilde{\sigma}_{k-1}(j) & \text{dla } \tilde{\sigma}_{k-1}(j) < \tilde{\sigma}_{k-1}(i) \\ \tilde{\sigma}_{k-1}(j) - 1 & \text{dla } \tilde{\sigma}_{k-1}(j) > \tilde{\sigma}_{k-1}(i) \\ n & \text{dla } j = i\,. \end{cases}$$

Następnie tworzymy graf $\tilde{G}_k$ z grafu $\tilde{G}_{k-1}$ poprzez usunięcie wszystkich krawędzi incydentnych z wierzchołkiem i oraz wygenerowanie losowo nowych krawędzi zgodnie z rozkładem $\mathbf{X}_{n-1}$. Mówiąc ściślej, jeśli przez $d^i(\tilde{\sigma}_k^{-1}(\ell))$, $\ell = 1, 2, \ldots, n-1$, oznaczymy liczbę krawędzi łączących wierzchołek i z wierzchołkiem $\tilde{\sigma}_k^{-1}(\ell)$, to wektor losowy

$$\left(d^i(\tilde{\sigma}_k^{-1}(1)), d^i(\tilde{\sigma}_k^{-1}(2)), \ldots, d^i(\tilde{\sigma}_k^{-1}(n-1))\right)$$

odpowiadający nowo wybranym krawędziom ma mieć rozkład $\mathbf{X}_{n-1}$. Przyjmujemy również $A_k = A_{k-1} \cup \{i\}$.

Innymi słowy, aby otrzymać $\tilde{G}_k$ usuwamy z grafu $\tilde{G}_{k-1}$ losowy wierzchołek (wraz z jego krawędziami), przenumerowujemy odpowiednio pozostałe wierzchołki, dodajemy nowy wierzchołek n oraz łączymy go z innymi wierzchołkami zgodnie z rozkładem $\mathbf{X}_{n-1}$. Przez A_k oznaczamy zbiór wszystkich wierzchołków, które zostały wybrane do tej pory.

Zdefiniujmy

$$L = \min\left\{k \colon A_k = [n]\right\},$$

tak by po L krokach, każdy z wierzchołków grafu G_0 został wybrany (i przenumerowany) przynajmniej raz. **Proces proteuszowy** $\mathfrak{P}(\mathbf{X}_{n-1})$ jest zdefiniowany jako łańcuch Markowa $\{(G_i, \sigma_i)\}_{i=0}^{\infty}$, w którym stanami są pary (G_i, σ_i), gdzie $G_i = \tilde{G}_{i+L}$ oraz $\sigma_i = \tilde{\sigma}_{i+L}$. Zauważmy, że zdefiniowany łańcuch $\mathfrak{P}(\mathbf{X}_{n-1}) = \{(G_i, \sigma_i)\}_{i=0}^{\infty}$ jest łańcuchem nieprzywiedlnym (wszystkie stany są istotne i komunikujące się) oraz ergodycznym (wszystkie stany są powracające, niezerowe i nieokresowe). Łańcuch ten znajduje się w stanie stacjonarnym, czyli rozkład zdeterminowany przez G_i na zbiorze wszystkich grafów o wierzchołkach ze zbioru $[n] = \{1, 2, \ldots, n\}$ jest identyczny dla wszystkich $i \geq 0$. Zauważmy również, że rozkład ten nie zależy od wyboru początkowego grafu G oraz permutacji σ. **Graf proteuszowy** $\mathcal{P}(\mathbf{X}_{n-1})$ jest grafem losowym zdefiniowanym poprzez ten rozkład. W celu uproszczenia rozważań, w dalszej części pracy założymy, że "wiek" wierzchołków w grafie proteuszowym odpowiada ich etykietom. Wierzchołek "najstarszy" (wierzchołek, który najdłużej nie był wybierany)

to wierzchołek posiadający numer 1, zaś "najmłodszy" (wierzchołek, który został wybrany w bieżącym kroku) to wierzchołek n. Zatem graf proteuszowy $\mathcal{P}(\mathbf{X}_{n-1})$ będziemy utożsamiać z $G_{\tilde{L}}$, gdzie

$$\tilde{L} = \min\left\{i : \sigma_i \text{ jest identycznością}\right\}.$$

Rzecz jasna, struktura grafu proteuszowego $\mathcal{P}(\mathbf{X}_{n-1})$ zależy od rozkładu zmiennej losowej $\mathbf{X}_{n-1}$. Jeżeli, na przykład, $\mathbf{X}_{n-1}$ posiada rozkład dwumianowy $B(n-1,p)$, to $\mathcal{P}(\mathbf{X}_{n-1})$ pokrywa się ze standardowym modelem grafu losowego $G(n,p)$. Innym prostym rozkładem jest rozkład, w którym wybrany w trakcie procesu wierzchołek łączy się z d losowo wybranymi wierzchołkami. W tym przypadku, otrzymany model jest podobny do wprowadzonego przez Pandurangana *et. al.* [24] modelu sieci typu *peer-to-peer* [9], w którym nowo dodany wierzchołek łączy się ze stałą, ustaloną liczbą wierzchołków ze zbioru możliwych kandydatów. Aby otrzymać sieć typu P2P należy, jak to zrobiono we wspomnianej pracy, wprowadzić dodatkowe krawędzie, aby uczynić graf proteuszowy spójnym. Wprowadzony model jest zatem uogólnieniem wcześniejszych modeli, pozwalającym jak się wkrótce przekonamy, przy właściwym doborze zmiennej losowej $\mathbf{X}_{n-1}$, modelować również rzeczywistą sieć internetową.

W niniejszej rozprawie rozważamy wyłącznie specjalny rodzaj grafu proteuszowego o n wierzchołkach, gdzie w każdym kroku procesu "nowy" wierzchołek wybiera niezależnie d sąsiadów spośród pozostałych wierzchołków, a prawdopodobieństwo wyboru ustalonego wierzchołka jest proporcjonalne do jego "wieku" podniesionego do potęgi η. Mówiąc dokładniej, dla $1 \leq s \leq n-1$ niech

$$\delta_s = (0,0,...,0,1,0,...,0) \quad \text{(jedynka na s-tym miejscu)},$$

$n, d \in \mathbb{N}$, $\eta \geq 0$, $\mathbf{X}_{n-1}^{\eta}$ oznacza nieujemną całkowitoliczbową zmienną losową zdefiniowaną poprzez następujący rozkład

$$\mathbb{P}(\mathbf{X}_{n-1}^{\eta} = \delta_s) = s^{-\eta} / \sum_{i=1}^{n-1} i^{-\eta}.$$

Przez $\mathbf{X}_{n-1}^{\eta}(i)$, $i = 1, 2, \ldots, d$ oznaczmy niezależne kopie $\mathbf{X}_{n-1}^{\eta}$, a

$$\mathbf{X}_{n-1}^{\eta,d} = \sum_{i=1}^{d} \mathbf{X}_{n-1}^{\eta}(i)\,.$$

W pracy będziemy badać własności **grafu proteuszowego** $\mathcal{P}(\mathbf{X}_{n-1}^{\eta,d})$ oznaczanego przez $\mathcal{P}_n(d,\eta)$ i **procesu proteuszowego** $\mathfrak{P}(\mathbf{X}_{n-1}^{\eta,d})$, który będziemy zapisywali jako $\mathfrak{P}_n(d,\eta)$.

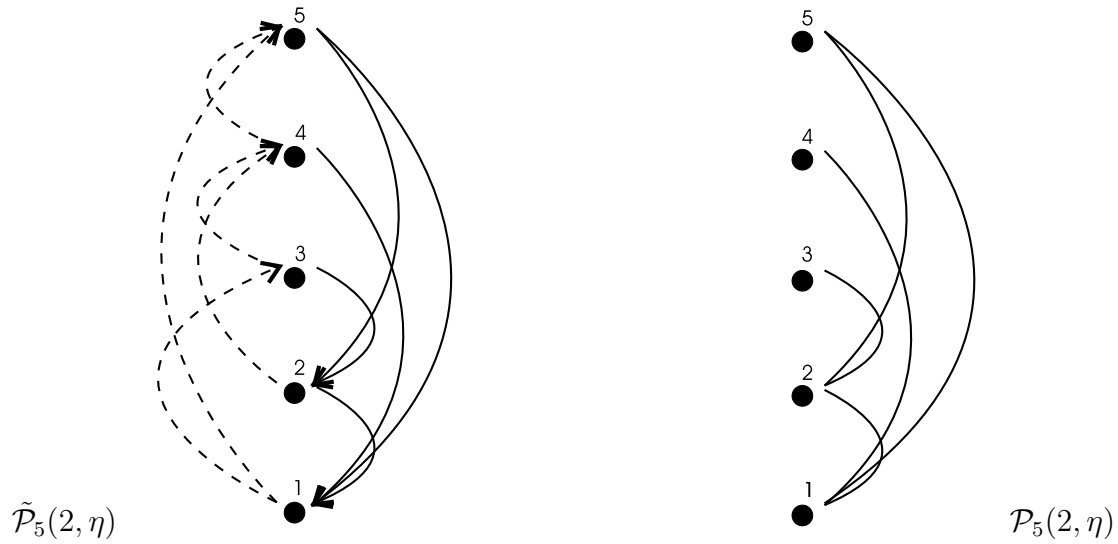
2 Stopnie wierzchołków grafu proteuszowego

2.1 Prawdopodobieństwo istnienia krawędzi

W niniejszym rozdziale obliczymy prawdopodobieństwo $p^{(d,\eta,n)}(j,i)$ występowania krawędzi pomiędzy wierzchołkami i oraz j $(i < j)$ w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ pokazując, że krawędzie w grafie proteuszowym pojawiają się "niemal niezależnie" (Twierdzenie 5).

Zgodnie z definicją, wybrany w trakcie procesu proteuszowego wierzchołek za każdym razem usuwa wszystkie krawędzie z nim połączone. Zatem jedynie ostatni wybór ustalonego wierzchołka wpływa na ostateczny kształt grafu proteuszowego. Wierzchołek, w momencie, kiedy po raz ostatni został wybrany, przesuwamy na koniec permutacji i łączymy z d losowo wybranymi wierzchołkami. Jednak nie wszystkie utworzone w ten sposób krawędzie "przetrwają" do zakończenia procesu. W grafie proteuszowym znajdą się tylko te z krawędzi, które są połączone z wierzchołkami, które do zakończenia procesu nie zostaną ponownie wybrane. Ze względu na czytelniejszą notację zatrzymujemy proces proteuszowy w momencie, gdy wierzchołki posortowane są według czasów ostatniego wyboru (numer wierzchołka pokrywa się z jego "wiekiem"; wierzchołek o numerze 1 to wierzchołek "najstarszy", natomiast wierzchołek o numerze n jest "najmłodszy").

Niech $\tilde{\mathcal{P}}_n(d,\eta)$ będzie grafem skierowanym o n wierzchołkach pojawiającym się w trakcie procesu proteuszowego. Wierzchołek j w tym grafie jest połączony z wierzchołkiem i wtedy i tylko wtedy, gdy w momencie kiedy wierzchołek j był po raz ostatni wybrany (w trakcie procesu proteuszowego) połączył się z wierzchołkiem i. Mając graf $\tilde{\mathcal{P}}_n(d,\eta)$ możemy łatwo utworzyć graf proteuszowy $\mathcal{P}_n(d,\eta)$, umieszczając krawędź $\{j,i\}$ w grafie $\mathcal{P}_n(d,\eta)$ wtedy, gdy łuk (j,i) pojawia się w grafie $\tilde{\mathcal{P}}_n(d,\eta)$ oraz $j > i$. Zasadę tę ilustruje rysunek, na którym umieściliśmy jedną z realizacji grafu losowego $\tilde{\mathcal{P}}_5(2,\eta)$, oraz otrzymany na jej podstawie graf proteuszowy $\mathcal{P}_5(2,\eta)$. Łuki, które "nie przetrwały" do zakończenia procesu proteuszowego zaznaczono linią przerywaną.



Przykładowo wierzchołek 3, w momencie kiedy był po raz ostatni wybrany, połączył się z wierzchołkami 2 oraz 4. Niestety krawędź pomiędzy wierzchołkami 3 i 4 nie będzie występowała w grafie proteuszowym $\mathcal{P}_5(2,\eta)$, ponieważ w kolejnych krokach procesu wybrano wierzchołek 4 i usunięto wszystkie krawędzie z nim połączone.

Z powyższych rozważań wynika, że prawdopodobieństwo istnienia krawędzi pomiędzy dowolnym wierzchołkiem j a wierzchołkiem i $(j > i)$ zależy od położenia (w permutacji) wierzchołka i w momencie, gdy po raz ostatni został wybrany wierzchołek j. Położenie to jest zmienną losową, której rozkład opisuje poniższe twierdzenie.

Twierdzenie 2. *Niech $1 \leq i < j \leq n$. Rozważmy proces proteuszowy $\{(G_i, \sigma_i)\}_{i=0}^{\infty}$ i niech $(\hat{G}^j, \hat{\sigma}^j)$ oznacza stan, w którym po raz ostatni (do momentu wygenerowania grafu proteuszowego) wybrano wierzchołek j, $1 \leq j \leq n$. Niech $U^n(j,i) = \hat{\sigma}^j(i)$, dla $1 \leq i < j$.*

Wtedy dla każdego k, $i \leq k \leq n - j + i$,

$$\mathbb{P}(U^n(j,i) = k) = \frac{\binom{k-1}{i-1}\binom{n-k-1}{j-i-1}}{\binom{n-1}{j-1}},$$

oraz

$$\mathbb{E}(U^n(j,i)) = \frac{in}{j} .$$

Co więcej, zmienna losowa $U^n(j,i)$ *jest silnie skoncentrowana wokół swojej wartości oczekiwanej. Dla* $\log^3 n < i < j \leq n$ *zachodzi bowiem następująca nierówność*

$$\mathbb{P}(|U^n(j,i) - \mathbb{E}(U^n(j,i))| \geq \frac{1}{\sqrt{\log n}}\mathbb{E}(U^n(j,i))) = o(\exp(-\log^{7/4} n)) .$$

Zanim udowodnimy powyższe twierdzenie przedstawimy pomocniczy lemat, mówiący o rozkładzie występującej w twierdzeniu permutacji $\hat{\sigma}^j$.

Lemat 3. *Oznaczmy przez* $\bar{\sigma}_n^j$ *losową permutację zbioru* $[n]$ *otrzymaną z jednostajnej losowej permutacji przez nałożenie warunku, że elementy* $1, \ldots, j$ *występują w tej permutacji we właściwym porządku oraz, że* $\bar{\sigma}_n^j(j) = n$.

$$\bar{\sigma}_n^j(1) < \bar{\sigma}_n^j(2) < \cdots < \bar{\sigma}_n^j(j) = n$$

Wtedy permutacja $\hat{\sigma}^j$ *może być identyfikowana z losową permutacją* $\bar{\sigma}_n^j$.

Dowód. Niech $\tilde{a} = \{a_i\}_{i=-\infty}^0$ oznacza losowy ciąg liczb całkowitych ze zbioru $[n]$, gdzie dla dowolnego $i \leq 0$ oraz $1 \leq r \leq n$,

$$\mathbb{P}(a_i = r) = 1/n .$$

Dla dowolnego $i \in [n]$ oraz $t \leq 0$ zdefiniujmy $T(t,i)$ w następujący sposób

$$T(t,i) = \max\{j \leq t : a_j = i\}$$

(jeżeli i nie występuje w nieskończonym ciągu $\{a_i\}_{i=-\infty}^t$, co zachodzi z prawdopodobieństwem równym 0, $T(t,i) = -\infty$). Niech teraz $\hat{a} = \{a_i\}_{i=-\infty}^0$ będzie losowym ciągiem otrzymanym z $\tilde{a}$ poprzez nałożenie warunku

$$-\infty < T(0,1) < T(0,2) < \cdots < T(0,n) = 0 .$$

Zauważmy, że ciąg $\hat{a}$ można interpretować jako ciąg wierzchołków wybieranych w czasie trwania procesu proteuszowego, w którym a_0 to

numer ostatniego wybranego wierzchołka (w tym kroku wygenerowano graf proteuszowy), a_{-i} to numer wierzchołka wybranego i kroków wcześniej, a $T(t,i)$ oznacza czas ostatniego wyboru wierzchołka i do momentu t.

Z definicji T wynika, że $T(T(0,j),i) - T(0,j)$ jest czasem, który upłynął od ostatniego wyboru wierzchołka i do momentu, w którym po raz ostatni wybrano wierzchołek j. A więc $\hat{\sigma}^j(i) = k$ wtedy i tylko wtedy, gdy k-ty najmniejszy element w ciągu

$$T(T(0,j),1), T(T(0,j),2), \ldots, T(T(0,j),n),$$

wynosi $T(T(0,j),i)$. Wprost z definicji wynika, że $T(T(0,j),j) = T(0,j)$ jest największym elementem w ciągu, co oznacza, że $\hat{\sigma}^j(j) = n$. Co więcej, dla dowolnego $i < j$ mamy $T(0,i) < T(0,j)$, czyli $T(T(0,j),i) = T(0,i)$. Ponieważ dla $1 \leq k < l \leq j$ zachodzi $T(0,k) < T(0,l)$ otrzymujemy, że

$$T(T(0,j),1) < T(T(0,j),2) < \cdots < T(T(0,j),j) \ .$$

Ostatecznie

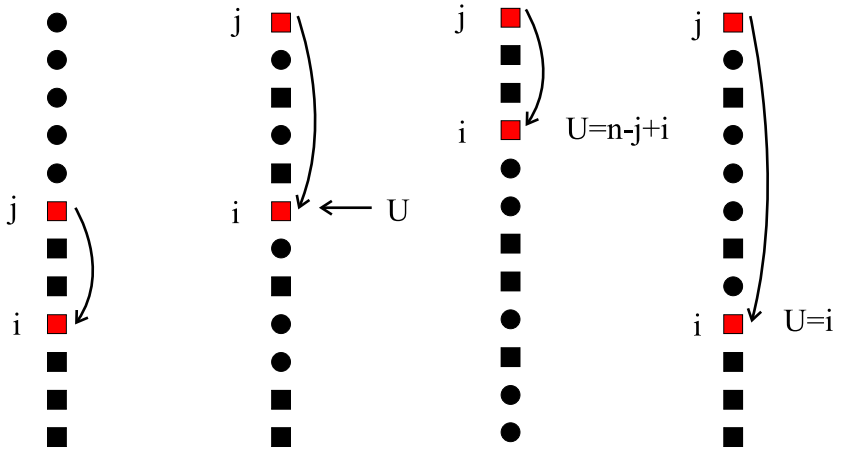
$$\hat{\sigma}^j(1) < \hat{\sigma}^j(2) < \cdots < \hat{\sigma}^j(j) = n \ ,$$

a więc permutacja $\hat{\sigma}^j$ może być identyfikowana z losową permutacją $\bar{\sigma}_n^j$. □

Powyższy lemat będzie pomocny w dowodzie Twierdzenia 2.

Dowód Twierdzenia 2. Niech $1 \leq i < j \leq n$, a $U^n(j,i)$ będzie zmienną losową zdefiniowaną w założeniu twierdzenia.

W dowodzie Lematu 3 pokazaliśmy, że permutację $\hat{\sigma}^j$ możemy utożsamiać z losową permutacją $\bar{\sigma}_n^j$, co między innymi oznacza, że wierzchołki $\{1,2,\ldots,j-1\}$ w permutacji $\hat{\sigma}^j$ występują w tej samej kolejności (nie muszą znajdować się na tych samych miejscach). Oznacza to, że zmienna losowa $U^n(j,i)$ może przyjmować wartości wyłącznie z przedziału $[i, n-j+i]$. Na rysunku poniżej przedstawiono kilka przykładowych sytuacji, obrazujących zachowanie się zmiennej losowej $U^n(j,i)$.



Wierzchołki $\{1, 2, \ldots, j\}$ są zaznaczone jako kwadraty. Wierzchołki i oraz j wyróżnione zostały innym kolorem. Prawdopodobieństwo istnienia krawędzi pomiędzy wierzchołkiem j a wierzchołkiem i (pierwszy rysunek od lewej) zależy od położenia w permutacji $\hat{\sigma}^j$ wierzchołka i (drugi rysunek), a więc od rozkładu zmiennej $U^n(j,i)$. Zmienna losowa osiąga swoje maksimum dla $U^n(j,i) = n - j + i$ (trzeci rysunek) oraz minimum w przypadku, gdy $U^n(j,i) = i$ (czwarty rysunek).

Policzmy teraz prawdopodobieństwo, że zmienna losowa $U^n(j,i)$ jest równa k. Istnieje $\binom{n-1}{j-1}(n-j)!$ sposobów ustawienia n elementów tak by wierzchołek j znalazł się na końcu permutacji, a $j-1$ ustalonych elementów (czarne kwadraty) wystąpiło w ustalonym porządku. Natomiast na $\binom{k-1}{i-1}\binom{n-k-1}{j-i-1}(n-j)!$ sposobów można ustawić te wierzchołki, tak by na k-tej pozycji znajdował się wierzchołek i. Zatem szukane prawdopodobieństwo wynosi

$$\mathbb{P}(U^n(j,i) = k) = \frac{\binom{k-1}{i-1}\binom{n-k-1}{j-i-1}}{\binom{n-1}{j-1}} .$$

Dokonując prostych przekształceń otrzymujemy

$$\begin{aligned}\mathbb{E}(U^n(j,i)) &= \sum_{k=i}^{n-j+i} k \cdot \mathbb{P}(U^n(j,i)=k) \\ &= \sum_{k=i}^{n-j+i} k \frac{\binom{k-1}{i-1}\binom{n-k-1}{j-i-1}}{\binom{n-1}{j-1}} \\ &= \frac{in}{j} \sum_{k=i}^{n-j+i} \frac{\binom{k}{i}\binom{n-k-1}{j-i-1}}{\binom{n}{j}} .\end{aligned}$$

Rozumując podobnie jak poprzednio, można pokazać, że ustawiając losowo n elementów, gdzie j z nich (wyróżnionych) znajduje się w ustalonym porządku, prawdopodobieństwo, że $i+1$-szy z wyróżnionych elementów znajduje się na $k+1$-szym miejscu wynosi $\binom{k}{i}\binom{n-k-1}{j-i-1}/\binom{n}{j}$, gdzie k może przyjmować wartości z przedziału $[i, n-j+i]$. Zatem suma w powyższym wzorze, jako suma prawdopodobieństw po wszystkich możliwych stanach, jest równa jeden i ostatecznie $\mathbb{E}(U^n(j,i)) = in/j$.

Aby pokazać, że zmienna losowa $U^n(j,i)$ jest silnie skoncentrowana wokół swojej wartości oczekiwanej, skorzystamy z analogicznej, dobrze znanej własności, która zachodzi dla zmiennej losowej o rozkładzie hipergeometrycznym [4, 5, 17].

Spośród n kul, z czego k pomalowano na czarno, wybieramy losowo j kul. Niech $H^n(k,j)$ będzie zmienną losową oznaczającą liczbę wylosowanych czarnych kul. Dla dowolnego $i \in N$

$$\mathbb{P}\Big(H^n(k,j)=i\Big) = \frac{\binom{k}{i}\binom{n-k}{j-i}}{\binom{n}{j}} .$$

Można pokazać, że

$$\mathbb{E}\Big(H^n(k,j)\Big) = \frac{kj}{n}$$

oraz, że dla dowolnego $\varepsilon \in (0,1)$ (zobacz [17] Twierdzenie 2.10),

$$\mathbb{P}\Big(\Big|H^n(k,j) - \mathbb{E}\big(H^n(k,j)\big)\Big| \geq \varepsilon \mathbb{E}\big(H^n(k,j)\big)\Big) \leq 2\exp\Big[-\frac{\varepsilon^2}{3}\mathbb{E}\big(H^n(k,j)\big)\Big] .$$

Niech $0 \leq \varepsilon \leq \frac{1}{4}$ oraz niech

$$\begin{aligned} k_\varepsilon^- &= \mathbb{E}\big(U^n(j,i)\big)(1-\varepsilon) = \frac{in}{j}(1-\varepsilon), \\ k_\varepsilon^+ &= \mathbb{E}\big(U^n(j,i)\big)(1+\varepsilon) = \frac{in}{j}(1+\varepsilon)\,. \end{aligned}$$

Pokażemy najpierw, że następujące prawdopodobieństwa $\mathbb{P}\big(U^n(j,i) = k_\varepsilon^-\big)$ oraz $\mathbb{P}\big(U^n(j,i) = k_\varepsilon^+\big)$, a tym samym $\mathbb{P}\big(U^n(j,i) \leq k_\varepsilon^-\big)$ oraz $\mathbb{P}\big(U^n(j,i) \geq k_\varepsilon^+\big)$, są małe. Istotnie

$$\begin{aligned} \mathbb{P}\big(U^n(j,i) = k_\varepsilon^-\big) &= \frac{\binom{k_\varepsilon^- - 1}{i-1}\binom{n-k_\varepsilon^- - 1}{j-i-1}}{\binom{n-1}{j-1}} = \frac{i}{k_\varepsilon^-} \cdot \frac{j-i}{n-k_\varepsilon^-} \cdot \frac{n}{j} \cdot \frac{\binom{k_\varepsilon^-}{i}\binom{n-k_\varepsilon^-}{j-i}}{\binom{n}{j}} \\ &= \frac{i}{\frac{i\cdot n}{j}(1-\varepsilon)} \cdot \frac{j-i}{n-\frac{i\cdot n}{j}(1-\varepsilon)} \cdot \frac{n}{j} \cdot \mathbb{P}\big(H^n(k_\varepsilon^-, j) = i\big) \\ &\leq \frac{1}{1-\varepsilon} \cdot \frac{j-i}{n-\frac{i\cdot n}{j}} \cdot \mathbb{P}\Big(H^n(k_\varepsilon^-, j) = \frac{k_\varepsilon^- j}{n(1-\varepsilon)}\Big) \\ &= \frac{1}{1-\varepsilon} \cdot \frac{j}{n} \cdot \mathbb{P}\Big(H^n(k_\varepsilon^-, j) = \frac{\mathbb{E}(H^n(k_\varepsilon^-, j))}{1-\varepsilon}\Big) \\ &\leq (1+2\varepsilon) \cdot \frac{j}{n} \cdot \mathbb{P}\Big(H^n(k_\varepsilon^-, j) \geq \frac{\mathbb{E}(H^n(k_\varepsilon^-, j))}{1-\varepsilon}\Big) \\ &\leq \frac{3}{2} \cdot \frac{j}{n} \cdot \mathbb{P}(H^n(k_\varepsilon^-, j) \geq (1+\varepsilon)\mathbb{E}(H^n(k_\varepsilon^-, j))) \\ &\leq \frac{3}{2} \cdot \frac{j}{n} \cdot 2\exp\Big(-\frac{\varepsilon^2}{3}\mathbb{E}(H^n(k_\varepsilon^-, j))\Big) \\ &= 3 \cdot \frac{j}{n} \cdot \exp\Big(-\frac{\varepsilon^2}{3}\frac{k_\varepsilon^- j}{n}\Big) \\ &\leq 3 \cdot \frac{j}{n} \cdot \exp\Big(-\frac{\varepsilon^2}{4} i\Big)\,, \end{aligned}$$

co oznacza, że

$$\begin{aligned} \mathbb{P}\big(U^n(j,i) \leq k_\varepsilon^-\big) &\leq k_\varepsilon^- \cdot \mathbb{P}\big(U^n(j,i) = k_\varepsilon^-\big) \\ &\leq 3i\,\exp\Big(-\frac{\varepsilon^2}{4} i\Big)\,. \end{aligned}$$

Podobnie można pokazać, że

$$\begin{aligned}
\mathbb{P}\big(U^n(j,i)=k_\varepsilon^+\big) &= \frac{\binom{k_\varepsilon^+-1}{i-1}\binom{n-k_\varepsilon^+-1}{j-i-1}}{\binom{n-1}{j-1}} = \frac{i}{k_\varepsilon^+}\cdot\frac{j-i}{n-k_\varepsilon^+}\cdot\frac{n}{j}\cdot\frac{\binom{k_\varepsilon^+}{i}\binom{n-k_\varepsilon^+}{j-i}}{\binom{n}{j}} \\
&= \frac{i}{\frac{i\cdot n}{j}(1+\varepsilon)}\cdot\frac{j-i}{n-\frac{i\cdot n}{j}(1+\varepsilon)}\cdot\frac{n}{j}\cdot\mathbb{P}\big(H^n(k_\varepsilon^+,j)=i\big) \\
&\le \frac{j-i}{n-\frac{i\cdot n}{j}(1+\varepsilon)}\cdot\mathbb{P}\Big(H^n(k_\varepsilon^+,j)=\frac{k_\varepsilon^+ j}{n(1+\varepsilon)}\Big) \\
&= \frac{j-i}{j-i(1+\varepsilon)}\cdot\frac{j}{n}\cdot\mathbb{P}\Big(H^n(k_\varepsilon^+,j)=\frac{\mathbb{E}(H^n(k_\varepsilon^+,j))}{1+\varepsilon}\Big) \\
&\le \Big(1+\frac{\varepsilon i}{j-i(1+\varepsilon)}\Big)\cdot\frac{j}{n}\cdot\mathbb{P}\Big(H^n(k_\varepsilon^+,j)\le\frac{\mathbb{E}(H^n(k_\varepsilon^+,j))}{1+\varepsilon}\Big) \\
&\le \Big(1+\frac{\varepsilon i}{j-i}\Big)\cdot\frac{j}{n}\cdot\mathbb{P}(H^n(k_\varepsilon^+,j)\le(1-\frac{\varepsilon}{2})\mathbb{E}(H^n(k_\varepsilon^+,j))) \\
&\le \Big(1+\frac{\varepsilon i}{j-i}\Big)\cdot\frac{j}{n}\cdot 2\exp\Big(-\frac{\varepsilon^2}{12}\mathbb{E}(H^n(k_\varepsilon^+,j))\Big) \\
&= 2\Big(1+\frac{\varepsilon i}{j-i}\Big)\cdot\frac{j}{n}\,\exp\Big(-\frac{\varepsilon^2}{12}\frac{k_\varepsilon^+ j}{n}\Big) \\
&\le 2\Big(1+\frac{\varepsilon i}{j-i}\Big)\cdot\frac{j}{n}\,\exp\Big(-\frac{\varepsilon^2}{12}i\Big)\,,
\end{aligned}$$

oraz

$$\begin{aligned}
\mathbb{P}\big(U^n(j,i)\ge k_\varepsilon^+\big) &\le (n-k_\varepsilon^+)\cdot\mathbb{P}\big(U^n(j,i)=k_\varepsilon^+\big) \\
&\le (n-\frac{i\cdot n}{j})\cdot 2\Big(1+\frac{\varepsilon i}{j-i}\Big)\cdot\frac{j}{n}\,\exp\Big(-\frac{\varepsilon^2}{12}i\Big) \\
&= 2\big(j-(1-\varepsilon)i)\big)\,\exp\Big(-\frac{\varepsilon^2}{12}i\Big) \\
&\le 2j\,\exp\Big(-\frac{\varepsilon^2}{12}i\Big)\,.
\end{aligned}$$

Zatem dla $\log^3 n < i \leq j \leq n$ oraz dla $\varepsilon = \log^{-1/2} n$

$$\begin{aligned}
\mathbb{P}\Big(\Big|U^n(j,i) - \mathbb{E}\big(U^n(j,i)\big)\Big| \geq \varepsilon \mathbb{E}\big(U^n(j,i)\big)\Big) \\
&= \mathbb{P}\big(U^n(j,i) \leq k_\varepsilon^-\big) + \mathbb{P}\big(U^n(j,i) \geq k_\varepsilon^+\big) \\
&\leq 3i \, \exp\Big(-\frac{\varepsilon^2}{4} i\Big) + 2j \, \exp\Big(-\frac{\varepsilon^2}{12} i\Big) \\
&\leq 5n \, \exp\Big(-\frac{\varepsilon^2}{12} i\Big) \\
&= o(\exp(-\log^{7/4} n)) \, .
\end{aligned}$$

□

Przedstawimy i udowodnimy teraz główne twierdzenie z tego podrozdziału, mówiące o prawdopodobieństwie istnienia (bądź nieistnienia) krawędzi w grafie proteuszowym $\mathcal{P}_n(d, \eta)$.

Nasze rozważania rozpoczniemy od następującej obserwacji. Przypuśćmy, że do m urn wrzucono d kul (jedna po drugiej), za każdym razem losując urny niezależnie z prawdopodobieństwami równymi odpowiednio $\rho_1, \ldots, \rho_m$, gdzie $\sum_{i=1}^m \rho_i = 1$. Niech teraz $S_1, S_2 \subseteq [m]$ będą rozłącznymi zbiorami urn, gdzie $|S_1| \leq d$, i niech $p(S_1, S_2)$ oznacza prawdopodobieństwo, że żadna z kul nie znajdzie się w urnie ze zbioru S_2, a każda urna ze zbioru S_1 zawierać będzie przynajmniej jedną kulę. Wtedy zachodzi następujący fakt.

Lemat 4. *Stosując powyższe oznaczenia otrzymujemy*

$$p(S_1, S_2) \geq (1 - \sum_{j \in S_1 \cup S_2} \rho_j)^{d - |S_1|} d(d-1) \ldots (d - |S_1| + 1) \prod_{i \in S_1} \rho_i \, ,$$

oraz

$$p(S_1, S_2) \leq (1 - \sum_{j \in S_2} \rho_j)^{d - |S_1|} d(d-1) \ldots (d - |S_1| + 1) \prod_{i \in S_1} \rho_i \, .$$

Dowód. Istotnie, wystarczy zauważyć, że wyraz po prawej stronie pierwszego oszacowania to prawdopodobieństwo, że każda z urn ze zbioru S_1 została wybrana dokładnie jeden raz, a urny z S_2 nie zostały wybrane ani razu. Wyrażenie po prawej stronie drugiej nierówności to oszacowanie z góry (niektóre zdarzenia mogą być policzone wielokrotnie) na prawdopodobieństwo zdarzenia, że każda z urn ze zbioru S_1 została wybrana przynajmniej raz, a żadna z urn z S_2 nie została wybrana. □

Wprowadźmy następujące oznaczenia. Niech $0 < \eta < 1$, $d \in \mathbb{N}$, oraz

$$E_1, E_2 \subseteq \big\{\{i,j\} : \log^3 n < i < j \le n\big\}, \quad E_1 \cap E_2 = \emptyset\,.$$

Ponadto, dla każdego $i, j \in [n]$, $r = 1, 2$, niech

$$V_r(j) = \big\{i < j : \{i,j\} \in E_r\big\}\,,$$

$$w(j,i) = (1-\eta)\frac{1}{n}\Big(\frac{j}{i}\Big)^{\eta} = \big(1 + O\big(n^{\eta-1}\big)\big)\frac{(in/j)^{-\eta}}{\sum_{s=1}^{n} s^{-\eta}}\,,$$

$$w_r(j) = \sum_{i \in V_r(j)} w(j,i)\,.$$

Twierdzenie 5. *Niech $\eta \in (0,1)$, $d \in \mathbb{N}$, E_1, E_2, $V_1(j)$, $w(j,i)$, $w_1(j)$, $w_2(j)$ będą zdefiniowane jak powyżej, oraz niech $|V_1(j)| \le d$ dla każdego $j \in [n]$.*

Przez $P_n(E_1, E_2, d, \eta)$ oznaczmy prawdopodobieństwo, że wszystkie pary ze zbioru E_1 są krawędziami w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ oraz żadna para ze zbioru E_2 nie jest krawędzią w $\mathcal{P}_n(d,\eta)$. Wtedy prawdziwe są następujące oszacowania:

$$\begin{aligned} P_n(E_1, E_2, d, \eta) &\le o(\exp(-\log^{3/2} n)) \\ &+ \prod_{j=1}^{n}(1 - (1 + O(\log^{-1/2} n))w_2(j))^{d-|V_1(j)|} \\ &\quad \times d(d-1)\dots(d-|V_1(j)|+1) \prod_{i \in V_1(j)} (1 + O(\log^{-1/2} n))w(j,i) \end{aligned}$$

oraz

$$P_n(E_1, E_2, d, \eta) \geq o(\exp(-\log^{3/2} n)) + \prod_{j=1}^{n} (1 - (1 + O(\log^{-1/2} n))(w_1(j) + w_2(j)))^{d-|V_1(j)|} \times d(d-1)\dots(d-|V_1(j)|+1) \prod_{i \in V_1(j)} (1 + O(\log^{-1/2} n)) w(j,i)\,.$$

Dowód. Policzmy najpierw ile wynosi suma wag W wszystkich wierzchołków w grafie proteuszowym $\mathcal{P}_n(d, \eta)$.

$$\begin{aligned} W &= \sum_{i=1}^{n} i^{-\eta} = O(1) + \int_0^n \lceil x \rceil^{-\eta}\, dx = \\ &= O(1) + \frac{n^{1-\eta}}{1-\eta} = (1 + O(n^{\eta-1}))\frac{n^{1-\eta}}{1-\eta}\,. \end{aligned} \tag{1}$$

Prawdopodobieństwo występowania (nie występowania) krawędzi pomiędzy dowolnym wierzchołkiem j a wierzchołkiem i zależy od zmiennej losowej $U^n(j,i)$, określającej położenie wierzchołka i w momencie, gdy po raz ostatni wybrano wierzchołek j. Z Twierdzenia 2 wynika, że dla $\varepsilon = \log^{-1/2} n$

$$\mathbb{P}\Big(\frac{in}{j}(1-\varepsilon) \leq U^n(j,i) \leq \frac{in}{j}(1+\varepsilon)\Big) = 1 - o(\exp(-\log^{7/4} n)).$$

Oznacza to, że z prawdopodobieństwem $1 - o(\exp(-\log^{7/4} n))$ położenie wierzchołka i w momencie, gdy po raz ostatni wybrano wierzchołek j wynosi

$$\frac{in}{j}\Big(1 + O(\log^{-1/2} n)\Big)\,,$$

a co za tym idzie jego waga wynosi

$$\begin{aligned} \big(U^n(j,i)\big)^{-\eta}/W &= \Big(\frac{in}{j}\big(1 + O(\log^{-1/2} n)\big)\Big)^{-\eta}(1 + O(n^{\eta-1}))(1-\eta)n^{\eta-1} \\ &= \big(1 + O(\log^{-1/2} n)\big)\frac{1-\eta}{n}\Big(\frac{j}{i}\Big)^{\eta} \\ &= \big(1 + O(\log^{-1/2} n)\big)w(j,i)\,. \end{aligned}$$

Przypomnijmy, że wierzchołek j, w momencie kiedy został po raz ostatni wybrany w trakcie procesu proteuszowego, wygenerował losowo d krawędzi (niezależnie) incydentnych z j. Przy czym, prawdopodobieństwo wyboru ustalonego wierzchołka jest wprost proporcjonalne do jego wagi. Z powyższych obserwacji oraz z Lematu 4 otrzymujemy następujące oszacowanie z góry na prawdopodobieństwo zdarzenia, że wierzchołek j, w grafie proteuszowym $\mathcal{P}_n(d,\eta)$, jest połączony z wierzchołkami ze zbioru $V_1(j)$ oraz nie jest połączony z wierzchołkami z $V_2(j)$

$$\begin{aligned}
&o(\exp(-\log^{7/4} n))(|V_1(j)|+|V_2(j)|)\\
&\quad +(1-(1+O(\log^{-1/2} n))w_2(j))^{d-|V_1(j)|}\\
&\qquad \times d(d-1)\dots(d-|V_1(j)|+1)\prod_{i\in V_1(j)}(1+O(\log^{-1/2} n))w(j,i)\,.
\end{aligned}$$

Analogicznie można uzyskać następujące oszacowanie z dołu

$$\begin{aligned}
&o(\exp(-\log^{7/4} n))(|V_1(j)|+|V_2(j)|)\\
&\quad +(1-(1+O(\log^{-1/2} n))(w_1(j)+w_2(j)))^{d-|V_1(j)|}\\
&\qquad \times d(d-1)\dots(d-|V_1(j)|+1)\prod_{i\in V_1(j)}(1+O(\log^{-1/2} n))w(j,i)\,.
\end{aligned}$$

Aby zakończyć dowód twierdzenia wystarczy teraz zauważyć, że generowanie krawędzi, w dowolnym kroku procesu proteuszowego, zależy jedynie od wag wierzchołków (od położenia wierzchołków w permutacji); nie zależy w szczególności od rozkładu istniejących już w grafie krawędzi. □

Analogiczne twierdzenie do Twierdzenia 5 zachodzi w przypadku, gdy $\eta = 0$, jednak w tym przypadku waga wierzchołka nie zależy od położenia wierzchołka w grafie (tj. $w(j,i) = 1/n$ dla każdego $1 \le i < j \le n$), co znacznie upraszcza tezę twierdzenia.

Twierdzenie 6. *Niech $d \in \mathbb{N}$, E_1, E_2, $V_1(j)$, $V_2(j)$ będą zdefiniowane jak powyżej, oraz niech $|V_1(j)| \leq d$ dla każdego $j \in [n]$.*

Przez $P_n(E_1, E_2, d, 0)$ oznaczmy prawdopodobieństwo, że wszystkie pary ze zbioru E_1 są krawędziami w grafie proteuszowym $\mathcal{P}_n(d, 0)$ oraz żadna para ze zbioru E_2 nie jest krawędzią w $\mathcal{P}_n(d, 0)$. Wtedy prawdziwe są następujące oszacowania:

$$\begin{aligned} P_n(E_1, E_2, d, 0) \;&\leq\; \prod_{j=1}^{n} (1 - |V_2(j)|/n)^{d-|V_1(j)|} \\ &\qquad \times d(d-1)\dots(d-|V_1(j)|+1)n^{-|V_1(j)|} \end{aligned}$$

oraz

$$\begin{aligned} P_n(E_1, E_2, d, 0) \;&\geq\; \prod_{j=1}^{n} (1 - |V_1(j)|/n - |V_2(j)|/n)^{d-|V_1(j)|} \\ &\qquad \times d(d-1)\dots(d-|V_1(j)|+1)n^{-|V_1(j)|}. \end{aligned}$$

Na koniec zauważmy, że Twierdzenia 2 i 5 sugerują, że własności grafu proteuszowego $\mathcal{P}_n(d, \eta)$, są podobne do własności grafu o wierzchołkach ze zbioru $[n]$, w którym dwa dowolne wierzchołki i, j, $1 \leq i < j \leq n$ są połączone krawędzią z prawdopodobieństwem

$$p(j, i) = dw(j, i) = (1 - \eta)\frac{d}{n}\Big(\frac{j}{i}\Big)^{\eta}, \tag{2}$$

niezależnie dla każdej pary wierzchołków. Istotnie, jeśli $|V_1(j)| = o(d)$ dla $j \in [n]$, to na mocy Twierdzenia 5, prawdopodobieństwo, że wszystkie pary ze zbioru E_1 są krawędziami w grafie proteuszowym $\mathcal{P}_n(d, \eta)$ oraz żadna para ze zbioru E_2 nie jest krawędzią w $\mathcal{P}_n(d, \eta)$ wynosi

$$\begin{aligned} P_n(E_1, E_2, d, \eta) \;&\approx\; \prod_{j=1}^{n} \Big(1 - \sum_{i \in V_2(j)} w(j, i)\Big)^{d} d^{|V_1(j)|} \prod_{i \in V_1(j)} w(j, i) \\ &=\; (1 + o(1)) \exp\Big(-d \sum_{\{i,j\} \in E_2} w(j, i)\Big) \prod_{\{i,j\} \in E_1} dw(j, i) \\ &=\; (1 + o(1)) \exp\Big(-\sum_{\{i,j\} \in E_2} p(j, i)\Big) \prod_{\{i,j\} \in E_1} p(j, i)\,. \end{aligned}$$

Natomiast w przypadku grafu z niezależnymi krawędziami, prawdopodobieństwo analogicznego zdarzenia wynosi

$$\prod_{\{i,j\}\in E_2}\Big(1-p(j,i)\Big)\prod_{\{i,j\}\in E_1}p(j,i)$$
$$=(1+o(1))\exp\Big(-\sum_{\{i,j\}\in E_2}p(j,i)\Big)\prod_{\{i,j\}\in E_1}p(j,i)\,.$$

W podobny sposób można argumentować, że własności grafu proteuszowego $\mathcal{P}_n(d,0)$ są zbieżne do własności grafu, w którym dwa dowolne wierzchołki są połączone krawędzią z prawdopodobnieństwem d/n.

2.2 Rozkład stopni

W niniejszym rozdziale pokażemy, że rozkład stopni w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ jest rozkładem potęgowym, w którym potęga zależy od parametru η, co pozwala mieć nadzieję, że będzie on przydatnym modelem sieci internetowej. Rozważania nasze zacznijmy jednak od znacznie prostszego zadania: oszacowania wartości oczekiwanej stopnia ustalonego wierzchołka.

Twierdzenie 7. *Niech $d=o(n^{(1-\eta)/2})$, $\eta\in[0,1)$. Oczekiwana liczba krawędzi w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ wynosi $\big(1+o(1)\big)dn/2$. Oczekiwany stopień wierzchołka $i=i(n)$, $\log^4 n\le i\le n$ wynosi*

$$\mathbb{E}\deg(i)=\big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\cdot\left(\left(\frac{n}{i}\right)^{\eta}+\frac{2\eta}{1-\eta}\frac{i}{n}\right).$$

Dowód. Przez $\deg^{>}(i)$ będziemy oznaczać liczbę krawędzi pomiędzy wierzchołkiem i a dowolnym wierzchołkiem $j>i$. Analogicznie $\deg^{<}(i)$ oznacza liczbę krawędzi pomiędzy wierzchołkiem i a dowolnym wierzchołkiem $j<i$. Ostatecznie $\deg(i)=\deg^{>}(i)+\deg^{<}(i)$ oznacza stopień wierzchołka i.

Policzmy najpierw oczekiwaną wartość $\deg^<(i)$. Zauważmy, że waga każdego wierzchołka jest nie większa niż $1/W$, gdzie W jest dane równaniem (1). Zatem w dowolnym momencie procesu proteuszowego, dowolny zbiór składający się z $\log^3 n$ wierzchołków posiada całkowitą wagę nie większą niż $\log^3 n/W = O(n^{\eta-1}\log^3 n)$. Oznacza to, że oczekiwana liczba krawędzi pomiędzy wierzchołkiem i a dowolnym z pierwszych $\log^3 n$ wierzchołków wynosi

$$d \cdot O(n^{\eta-1}\log^3 n) = o(n^{-(1-\eta)/2}\log^3 n) = o(1)\ .$$

Co więcej, z Twierdzenia 2 wynika, że z prawdopodobieństwem $1 - o(\exp(-\log^{3/2} n))$, całkowita waga wszystkich wierzchołków j, $\log^3 n \le j < i$, wynosi

$$\begin{aligned}
\big(1+o(1)\big)\sum_{j=\log^3 n}^{i} w(j,i) &= \big(1+o(1)\big)(1-\eta)\frac{1}{n}i^{\eta}\sum_{j=\log^3 n}^{i} j^{-\eta} \\
&= \big(1+o(1)\big)(1-\eta)\frac{1}{n}i^{\eta}n\int_{\log^3 n/n}^{i/n}\lceil nx\rceil^{-\eta}dx \\
&= \big(1+o(1)\big)(1-\eta)\Big(\frac{i}{n}\Big)^{\eta}\int_{0}^{i/n} x^{-\eta}dx \\
&= \big(1+o(1)\big)(1-\eta)\Big(\frac{i}{n}\Big)^{\eta}\frac{(i/n)^{1-\eta}}{1-\eta} \\
&= \big(1+o(1)\big)i/n\,,
\end{aligned}$$

a zatem

$$\mathbb{E}\deg^<(i) = (1+o(1))di/n + o(1)\ .$$

Prawdziwość tej równości można pokazać również w następujący sposób. W momencie, gdy po raz ostatni wierzchołek i został wybrany (w trakcie procesu proteuszowego) połączył się z d innymi wierzchołkami. Od tego momentu, aż do zakończenia procesu generowania grafu proteuszowego, wybrano (być może niektóre wierzchołki kilka razy) dokładnie $n-i$ wierzchołków. Tym samym oczekiwana wartość usuniętych krawędzi wynosi $d(n-i)/n$, a więc oczekiwana wartość krawędzi, które "przetrwały" do końca procesu wynosi di/n.

Z drugiej strony, korzystając z Twierdzenia 5 i wzoru (2) mamy

$$\begin{aligned}
\mathbb{E}\deg^{>}(i) &= \big(1+o(1)\big)\sum_{j=i+1}^{n} p(j,i) \\
&= \big(1+o(1)\big)\frac{(1-\eta)d}{ni^{\eta}}\sum_{j=i+1}^{n} j^{\eta} \\
&= \big(1+o(1)\big)\frac{(1-\eta)d}{ni^{\eta}} n\int_{i/n}^{1} \lceil nx\rceil^{\eta} dx \\
&= \big(1+o(1)\big)d(1-\eta)\Big(\frac{n}{i}\Big)^{\eta}\int_{i/n}^{1} x^{\eta}dx \\
&= \big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\Big(\frac{n}{i}\Big)^{\eta}\Big(1-\Big(\frac{i}{n}\Big)^{1+\eta}\Big) \\
&= \big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\Big(\Big(\frac{n}{i}\Big)^{\eta}-\frac{i}{n}\Big)\,.
\end{aligned}$$

Ostatecznie więc

$$\mathbb{E}\deg(i) = \mathbb{E}\deg^{<}(i) + \mathbb{E}\deg^{>}(i) = \big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\Big(\Big(\frac{n}{i}\Big)^{\eta} + \frac{2\eta}{1-\eta}\frac{i}{n}\Big)\,.$$

Na koniec zauważmy, że oczekiwana liczba krawędzi w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ wynosi

$$\sum_{i=1}^{n}\mathbb{E}\deg^{<}(i) = \big(1+o(1)\big)\sum_{i=1}^{n} di/n = \big(1+o(1)\big)dn/2\,.$$

□

Oznaczmy przez $Z_k = Z_k(n;d;\eta)$ liczbę wierzchołków stopnia k w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ oraz przez $Z_{\geq k} = \sum_{i\geq k} Z_i$ liczbę wierzchołków o stopniu większym lub równym k. Pokażemy, że z dużym prawdopodobieństwem stopień wierzchołka jest bliski swojej wartości oczekiwanej i korzystając z tej własności oszacujemy $Z_{\geq k}$ dla k znacznie większych od d.

Twierdzenie 8. *Niech $d = O(log^2 n)$, $\eta \in (0,1)$, $k = k(n) \geq \log^4 n$. Wtedy*

$$Z_{\geq k} = (1+o(1))n\Big(\frac{1-\eta}{1+\eta}\cdot\frac{d}{k}\Big)^{1/\eta} + O(\log^3 n)\,.$$

Dowód. Z Twierdzenia 7 wynika, że

$$\begin{aligned}\mathbb{E}\,\deg(i) &= \big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\Big[\Big(\frac{n}{i}\Big)^{\eta} + \frac{2\eta}{1-\eta}\frac{i}{n}\Big] \\ &= \big(1+o(1)\big)d\frac{1-\eta}{1+\eta}\Big(\frac{n}{i}\Big)^{\eta} + O(d) \\ &= \mathbb{E}\,\deg^{>}(i) + O(d)\,.\end{aligned}$$

Oznacza to, że dla stosunkowo małego

$$i_0 = i_0(n) = n\Big(\frac{1-\eta}{1+\eta}\cdot\frac{d}{k_0}\Big)^{1/\eta} = o(n)\,,$$

oczekiwany stopień wierzchołka i_0 wynosi z grubsza k_0. Pokażemy, że dla dowolnego $\varepsilon > 0$ prawie na pewno wszystkie wierzchołki i takie, że

$$i \geq (1+\varepsilon)i_0 = (1+\varepsilon)n\Big(\frac{1-\eta}{1+\eta}\cdot\frac{d}{k_0}\Big)^{1/\eta} \geq \log^3 n\,,$$

mają mniej niż k_0 sąsiadów oraz wszystkie wierzchołki i takie, że

$$i \leq (1-\varepsilon)i_0 = (1-\varepsilon)n\Big(\frac{1-\eta}{1+\eta}\cdot\frac{d}{k_0}\Big)^{1/\eta}\,,$$

mają więcej niż k_0 sąsiadów.

Niech $0 < \varepsilon < 1$ oraz $i \geq (1+\varepsilon)i_0$. Niech $\delta > 0$ będzie dowolną stałą, dla której zachodzą następujące nierówności: $\delta \leq \varepsilon^4/6$ i $\delta < (1-\varepsilon^2)(1+\varepsilon)^{\eta} - 1$ ($\frac{1+\delta}{(1+\varepsilon)^{\eta}} < 1-\varepsilon^2$). Zauważmy, że dla funkcji $f(\varepsilon) = (1-\varepsilon^2)(1+\varepsilon)^{\eta} - 1$ mamy $f'(0) = \eta > 0$, a zatem dla odpowiednio małego parametru ε wartość funkcji f w tym punkcie jest większa od zera ($f(\varepsilon) > 0$) i stała $\delta > 0$ o powyższych własnościach istnieje.

Pokażemy teraz, że prawie na pewno $\deg^>(i) < k_0 - d = k_0(1+o(1))$ $(\deg(i) < k_0)$.

$$\begin{aligned}
&\mathbb{P}(\deg^>(i) \geq k_0(1+o(1))) \\
&\quad = o(\exp(-\log^{3/2} n)) + \sum_{k \geq k_0(1+o(1))} \mathbb{P}(\deg^>(i) = k) \\
&\quad = o(\exp(-\log^{3/2} n)) + \sum_{k \geq k_0(1+o(1))} \sum_{x_1,\ldots,x_k > i} \\
&\qquad \prod_{j=1}^{k} (1+o(1))p(x_j,i) \prod_{j>i, j \notin \{x_1,\ldots,x_k\}} (1-(1+o(1))p(j,i)) \\
&\quad \leq \sum_{k \geq k_0(1+o(1))} \sum_{x_1,\ldots,x_k > i} \prod_{j=1}^{k} (1+\delta)p(x_j,i) \prod_{j>i, j \notin \{x_1,\ldots,x_k\}} (1-(1+\delta)p(j,i)) \\
&\qquad \times \prod_{j>i} \frac{1-(1+o(1))p(j,i)}{1-(1+\delta)p(j,i)} \,.
\end{aligned}$$

Oszacujmy najpierw ostatni iloczyn

$$\begin{aligned}
\prod_{j>i} \frac{1-(1+o(1))p(j,i)}{1-(1+\delta)p(j,i)} &= \prod_{j>i} \Big(1 + \frac{(\delta+o(1))p(j,i)}{1-(1+\delta)p(j,i)}\Big) \\
&= \prod_{j>i} \Big(1 + (1+o(1))\delta p(j,i)\Big) \\
&\leq \exp\Big((1+o(1))\delta \sum_{j>i} p(j,i)\Big) \,. \qquad (3)
\end{aligned}$$

Wprowadźmy pomocniczą zmienną losową $X_i = \sum_{j>i} X_i(j)$, gdzie $X_i(j)$, $i < j \leq n$ jest rodziną niezależnych zmiennych losowych zdefiniowanych w następujący sposób

$$\begin{aligned}
\mathbb{P}(X_i(j) = 1) &= (1+\delta)p(j,i)\,, \\
\mathbb{P}(X_i(j) = 0) &= 1-(1+\delta)p(j,i)\,.
\end{aligned}$$

Wtedy

$$\mathbb{P}(\deg^>(i) \geq k_0(1+o(1))) \leq \mathbb{P}(X_i \geq k_0(1+o(1))) \exp\Big((1+o(1))\delta \sum_{j>i} p(j,i)\Big)\,.$$

Aby pokazać, że powyższe prawdopodobieństwo dąży do zera skorzystamy ze znanej własności sumy niezależnych zmiennych losowych o rozkładzie Bernouliego (patrz na przykład Twierdzenie 2.8 [17]) mówiącej, że dla dowolnego $0 < \varepsilon < 3/2$

$$\mathbb{P}\Big(|X_i - \mathbb{E}X_i| \geq \varepsilon \mathbb{E}X_i\Big) \leq 2\exp\Big(-\frac{\varepsilon^2}{3}\mathbb{E}X_i\Big)\,. \tag{4}$$

Oszacujmy teraz wartość oczekiwaną zmiennej $X_{(1+\varepsilon)i_0}$

$$\begin{aligned}
\mathbb{E}X_{(1+\varepsilon)i_0} &= \sum_{j>(1+\varepsilon)i_0} (1+\delta)p(j,(1+\varepsilon)i_0) \\
&= (1+\delta)\frac{1-\eta}{n} d\Big((1+\varepsilon)i_0\Big)^{-\eta} \sum_{j>(1+\varepsilon)i_0} j^{\eta} \\
&= (1+o(1))(1+\delta)(1+\varepsilon)^{-\eta}(1+\eta)k_0 \int_0^1 x^{\eta}dx \\
&= (1+o(1))\frac{1+\delta}{(1+\varepsilon)^{\eta}}k_0 \leq (1-\varepsilon^2)k_0\,.
\end{aligned}$$

Zatem

$$\begin{aligned}
&\mathbb{P}(\deg^{>}(i) \geq k_0(1+o(1))) \\
&\qquad \leq \mathbb{P}(\deg^{>}((1+\varepsilon)i_0) \geq k_0(1+o(1))) \\
&\qquad \leq \mathbb{P}(X_{(1+\varepsilon)i_0} \geq k_0(1+o(1)))\exp\Big((1+o(1))\delta \sum_{j>(1+\varepsilon)i_0} p(j,(1+\varepsilon)i_0)\Big) \\
&\qquad \leq \mathbb{P}(|X_{(1+\varepsilon)i_0} - \mathbb{E}X_{(1+\varepsilon)i_0}| \geq \varepsilon^2 k_0)\exp\Big(\frac{1+o(1)}{1+\delta}\delta\mathbb{E}X_{(1+\varepsilon)i_0}\Big) \\
&\qquad \leq \mathbb{P}(|X_{(1+\varepsilon)i_0} - \mathbb{E}X_{(1+\varepsilon)i_0}| \geq \varepsilon^2 \mathbb{E}X_{(1+\varepsilon)i_0})\exp\Big(\delta\mathbb{E}X_{(1+\varepsilon)i_0}\Big)
\end{aligned}$$

i ostatecznie, na mocy (4),

$$\begin{aligned}
\mathbb{P}(\deg^{>}(i) \geq k_0(1+o(1))) &\leq 2\exp\Big(-\frac{\varepsilon^4}{3}\mathbb{E}X_{(1+\varepsilon)i_0}\Big)\exp\Big(\delta\mathbb{E}X_{(1+\varepsilon)i_0}\Big) \\
&\leq 2\exp\Big(-\frac{\varepsilon^4}{6}\mathbb{E}X_{(1+\varepsilon)i_0}\Big) \\
&\leq 2\exp\Big(-\Omega(\log^2 n)\Big) = o(n^{-1})\,.
\end{aligned}$$

Oznacza to, że prawdopodobieństwo, że stopień wierzchołka i jest większy lub równy k_0 dąży bardzo szybko do zera (jest mniejsze niż $1/n$). Zatem prawie na pewno wszystkie wierzchołki i takie, że $i \geq (1+\varepsilon)i_0$ mają mniej niż k_0 sąsiadów.

Analogicznie można pokazać, że prawie na pewno wszystkie wierzchołki i takie, że $\log^3 n < i \leq (1-\varepsilon)i_0$ mają więcej niż k_0 sąsiadów. (zwróćmy uwagę, że Twierdzenie 5 nie pozwala oszacować prawdopodobieństwa istnienia krawędzi incydentnych z wierzchołkiem $i < \log^3 n$.) Istotnie, niech teraz $i \leq (1-\varepsilon)i_0$. Zakładamy tym razem, że zachodzi następująca nierówność $(1+\varepsilon^2)(1-\varepsilon)^\eta - 1 \leq 0$ ($\frac{1}{(1-\varepsilon)^\eta} \geq 1+\varepsilon^2$). Zauważmy, że dla funkcji $g(\varepsilon) = (1+\varepsilon^2)(1-\varepsilon)^\eta - 1$ mamy $g'(0) = -\eta < 0$, a zatem dla odpowiednio małego parametru ε wartość funkcji f w tym punkcie jest mniejsza od zera i powyższa nierówność zachodzi. Niech $0 < \delta \leq \varepsilon^4/24$ będzie dowolną stałą. Pokażemy teraz, że prawie na pewno $\deg^>(i) > k_0$, a tym samym $\deg(i) > k_0$. Korzystając z (3) oraz wprowadzonej wcześniej pomocniczej zmiennej losowej X_i otrzymujemy

$$\begin{aligned}
\mathbb{P}(\deg^>(i) \leq k_0) &= o(\exp(-\log^{3/2} n)) + \sum_{k \leq k_0} \mathbb{P}(\deg^>(i) = k) \\
&= o(\exp(-\log^{3/2} n)) + \sum_{k \leq k_0} \sum_{x_1,\ldots,x_k > i} \\
&\quad \prod_{j=1}^{k} (1+o(1))p(x_j,i) \prod_{j>i, j \notin \{x_1,\ldots,x_k\}} (1-(1+o(1))p(j,i)) \\
&\leq \sum_{k \leq k_0} \sum_{x_1,\ldots,x_k > i} \prod_{j=1}^{k} (1+\delta)p(x_j,i) \prod_{j>i, j \notin \{x_1,\ldots,x_k\}} (1-(1+\delta)p(j,i)) \\
&\qquad \times \prod_{j>i} \frac{1-(1+o(1))p(j,i)}{1-(1+\delta)p(j,i)} \\
&\leq \mathbb{P}(X_i \leq k_0) \exp\left((1+o(1))\delta \sum_{j>i} p(j,i)\right).
\end{aligned}$$

Oszacujmy teraz wartość oczekiwaną zmiennej $X_{(1-\varepsilon)i_0}$.

$$\begin{aligned}
\mathbb{E}X_{(1-\varepsilon)i_0} &= \sum_{j>(1-\varepsilon)i_0} (1+\delta)p(j,(1-\varepsilon)i_0) \\
&= (1+\delta)\frac{1-\eta}{n} d\Big((1-\varepsilon)i_0\Big)^{-\eta} \sum_{j>(1-\varepsilon)i_0} j^{\eta} \\
&= (1+o(1))(1+\delta)(1-\varepsilon)^{-\eta}(1+\eta)k_0 \int_0^1 x^{\eta}dx \\
&= (1+o(1))\frac{1+\delta}{(1-\varepsilon)^{\eta}}k_0 \geq \frac{k_0}{(1-\varepsilon)^{\eta}} \geq (1+\varepsilon^2)k_0\,.
\end{aligned}$$

Zatem

$$\begin{aligned}
\mathbb{P}(\deg^{>}(i) \leq k_0) &\leq \mathbb{P}(\deg^{>}((1-\varepsilon)i_0) \leq k_0) \\
&\leq \mathbb{P}(X_{(1-\varepsilon)i_0} \leq k_0) \exp\Big((1+o(1))\delta \sum_{j>(1-\varepsilon)i_0} p(j,(1-\varepsilon)i_0)\Big) \\
&\leq \mathbb{P}(|X_{(1-\varepsilon)i_0} - \mathbb{E}X_{(1-\varepsilon)i_0}| \geq \varepsilon^2 k_0) \exp\Big(\frac{1+o(1)}{1+\delta}\delta\mathbb{E}X_{(1-\varepsilon)i_0}\Big) \\
&\leq \mathbb{P}(|X_{(1-\varepsilon)i_0} - \mathbb{E}X_{(1-\varepsilon)i_0}| \geq \frac{\varepsilon^2}{1+\varepsilon}\mathbb{E}X_{(1-\varepsilon)i_0}) \exp\Big(\delta\mathbb{E}X_{(1-\varepsilon)i_0}\Big) \\
&\leq \mathbb{P}(|X_{(1-\varepsilon)i_0} - \mathbb{E}X_{(1-\varepsilon)i_0}| \geq \frac{\varepsilon^2}{2}\mathbb{E}X_{(1-\varepsilon)i_0}) \exp\Big(\delta\mathbb{E}X_{(1-\varepsilon)i_0}\Big)
\end{aligned}$$

i ostatecznie, na mocy (4),

$$\begin{aligned}
\mathbb{P}(\deg^{>}(i) \leq k_0) &\leq 2\exp\Big(-\frac{\varepsilon^4}{12}\mathbb{E}X_{(1-\varepsilon)i_0}\Big)\exp\Big(\delta\mathbb{E}X_{(1-\varepsilon)i_0}\Big) \\
&\leq 2\exp\Big(-\frac{\varepsilon^4}{24}\mathbb{E}X_{(1-\varepsilon)i_0}\Big) \\
&\leq 2\exp\Big(-\Omega(\log^2 n)\Big) = o(n^{-1})\,.
\end{aligned}$$

Oznacza to, że prawdopodobieństwo, że stopień wierzchołka i jest mniejszy lub równy k_0 dąży bardzo szybko do zera (jest mniejsze niż $1/n$). Zatem prawie na pewno wszystkie wierzchołki i takie, że $i \leq (1-\varepsilon)i_0$ mają więcej niż k_0 sąsiadów. □

Przypomnijmy, że w grafie internetowym liczba wierzchołków o stopniu wejściowym wynoszącym k jest proporcjonalna do $k^{-2,1}$, natomiast liczba wierzchołków o stopniu wyjściowym k maleje jak $k^{-2,7}$. Oznacza to, że liczba wierzchołków o stopniu wejściowym nie mniejszym niż k wynosi

$$\sum_{l \geq k} O(l^{-2,1}) = O(k^{-1,1}) \ ,$$

wyjściowym natomiast

$$\sum_{l \geq k} O(l^{-2,7}) = O(k^{-1,7}) \ .$$

Dla nieskierowanej wersji grafu internetowego frakcja wierzchołków o stopniu nie mniejszym niż k wynosi

$$\sum_{l \geq k} \big(O(l^{-2,1}) + O(l^{-2,7}) \big) = O(k^{-1,1}) \ ,$$

zatem, ze względu na możliwość modelowania sieci internetowej, szczególnie interesujący jest graf proteuszowy dla którego parametr η wynosi $\eta = \eta_{\text{in}} \sim 1/1,1 \sim 0{,}91$, a także, w pewnym stopniu $\eta = \eta_{\text{out}} \sim 1/1{,}7 = 0{,}59$.

Niniejszy rozdział zakończymy dwoma prostymi wnioskami z Twierdzenia 7.

Wniosek 9. *Niech $d \in N$, $\eta \in (0,1)$ i*

$$c_0(\eta) = \Big(\frac{1-\eta}{2}\Big)^{\frac{1}{\eta+1}} .$$

Wtedy w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ najmniejszy oczekiwany stopień ma wierzchołek o numerze $(1+o(1))c_0(\eta)n$.

Dowód. Zgodnie z Twierdzeniem 7 oczekiwany stopień wierzchołka $k = \lceil cn \rceil$, $c \in (0,1)$ wynosi

$$\begin{aligned} \mathbb{E}\deg(k) &= (1+o(1))d\frac{1-\eta}{1+\eta}(c^{-\eta} + \frac{2\eta}{1-\eta}c) \\ &= (1+o(1))d\frac{1-\eta}{1+\eta}f(c) , \end{aligned}$$

gdzie funkcja $f : (0,1) \longrightarrow \mathbb{R}$ określona jest wzorem

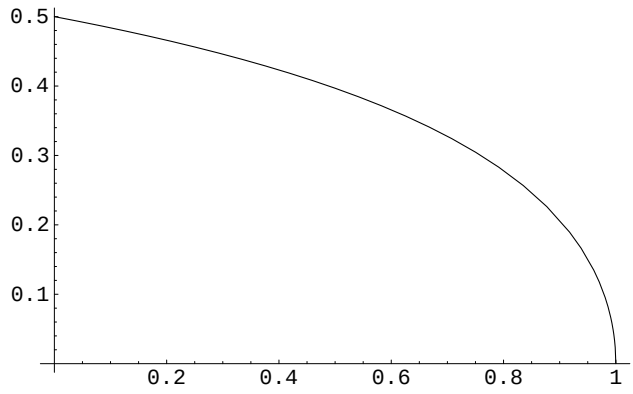
$$f(c) = c^{-\eta} + \frac{2\eta}{1-\eta}c .$$

Aby obliczyć minimum $c_0(\eta)$ funkcji f policzmy jej pierwszą i drugą pochodną.

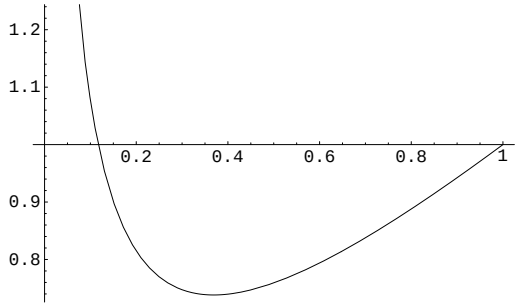
$$\begin{aligned} f'(c) &= -\eta c^{-1-\eta} + \frac{2\eta}{1-\eta} \\ f''(c) &= \eta(1+\eta)c^{-2-\eta} \end{aligned}$$

Dla każdego $c \in (0,1)$ $f''(c) > 0$ oraz $f'(c) = 0$ wtedy i tylko wtedy, gdy $c = (\frac{1-\eta}{2})^{\frac{1}{\eta+1}}$. Zatem najmniejszy oczekiwany stopień ma wierzchołek o etykiecie $(1+o(1))c_0(\eta)n$, gdzie $c_0(\eta) = (\frac{1-\eta}{2})^{\frac{1}{\eta+1}}$. □

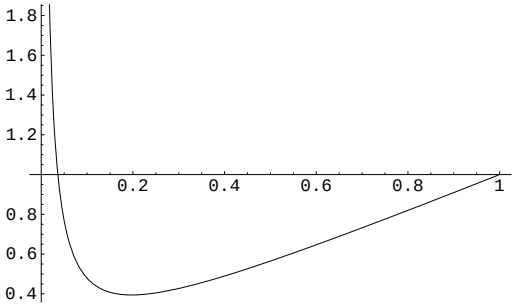
Z powyższego twierdzenia wynika, że gdy $\eta \to 0$, wierzchołkiem posiadającym najmniejszy oczekiwany stopień jest wierzchołek o numerze $(1 + o(1))n/2$. Na poniższym wykresie przedstawiono jak zmienia się $c_0(\eta)$ (oś OY) w zależności od parametru η (oś OX).



W interesujących nas przypadkach, gdy $\eta_{\text{out}} = 0{,}59$ oraz $\eta_{\text{in}} = 0{,}91$ (ze względu na możliwość modelowania sieci internetowej), szukane minimum znajduje się odpowiednio w punktach $c_0(\eta_{\text{out}}) \approx 0{,}369$ oraz $c_0(\eta_{\text{in}}) \approx 0{,}197$. Na poniższych wykresach przedstawiono oczekiwany stopień w zależności od położenia wierzchołka w grafie proteuszowym (dla $d = 1$).



$\eta_{\text{out}} = 0{,}59$
$c_0(\eta_{\text{out}}) \approx 0{,}369$



$\eta_{\text{in}} = 0{,}91$
$c_0(\eta_{\text{in}}) \approx 0{,}197$

Wiedząc, który wierzchołek posiada najmniejszy oczekiwany stopień możemy łatwo oszacować z dołu oczekiwany stopień dowolnego wierzchołka.

Wniosek 10. *Niech $d \in \mathbb{N}$, $\eta \in (0,1)$. Dla dowolnego wierzchołka $i \in [n]$ w grafie proteuszowym $\mathcal{P}_n(d,\eta)$*

$$\mathbb{E}\deg(i) \geq (1+o(1))2d\Big(\frac{1-\eta}{2}\Big)^{\frac{1}{1+\eta}} .$$

Dowód. Bezpośrednio z Twierdzenia 7 oraz Wniosku 9 wynika, że minimalny oczekiwany stopień wynosi

$$\begin{aligned}
\mathbb{E}\deg\big(c_0(\eta)n\big) &= (1+o(1))d\frac{1-\eta}{1+\eta}\Big(c_0(\eta)^{-\eta}+\frac{2\eta}{1-\eta}c_0(\eta)\Big)\\
&= (1+o(1))d\frac{1-\eta}{1+\eta}\Big[\Big(\frac{1-\eta}{2}\Big)^{\frac{-\eta}{1+\eta}}+\frac{2\eta}{1-\eta}\Big(\frac{1-\eta}{2}\Big)^{\frac{1}{1+\eta}}\Big]\\
&= (1+o(1))d\frac{1-\eta}{1+\eta}\Big[\Big(\frac{1-\eta}{2}\Big)^{\frac{-\eta}{1+\eta}}+\eta\Big(\frac{1-\eta}{2}\Big)^{\frac{-\eta}{1+\eta}}\Big]\\
&= (1+o(1))d(1-\eta)\Big(\frac{1-\eta}{2}\Big)^{\frac{-\eta}{1+\eta}}\\
&= (1+o(1))2d\Big(\frac{1-\eta}{2}\Big)^{\frac{1}{1+\eta}} .
\end{aligned}$$

□

W szczególności więc, dla dowolnego wierzchołka $i \in [n]$

$$\begin{aligned}
&\mathbb{E}\deg(i) \geq 0{,}73\cdot d \qquad \text{dla} \qquad \eta_{\text{out}} = 0{,}59 ,\\
&\mathbb{E}\deg(i) \geq 0{,}39\cdot d \qquad \text{dla} \qquad \eta_{\text{in}} = 0{,}91 .
\end{aligned}$$

2.3 "Kryzys wieku średniego"

Rozdział ten jest poświęcony rozważaniom na temat prawdopodobieństwa bycia wierzchołkiem izolowanym. Wierzchołki o stosunkowo małych numerach (wierzchołki, które dawno nie były "odświeżane") posiadają duże wagi, a co za tym idzie są często wybierane przez inne wierzchołki. Wierzchołki o dużych numerach (wierzchołki, które stosunkowo niedawno były "odświeżane") w momencie wyboru połączyły się z d losowo wybranymi wierzchołkami i z dużym prawdopodobieństwem przynajmniej niektóre z utworzonych w ten sposób krawędzi przetrwają do końca procesu. W najgorszym położeniu znajdują się wierzchołki w środku grafu proteuszowego. Dla nich prawdopodobieństwo bycia wierzchołkiem izolowanym jest największe. Możemy więc powiedzieć, że część wierzchołków w grafie proteuszowym przeżywa "kryzys wieku średniego". Poniższe twierdzenia precyzują powyższą obserwację.

Twierdzenie 11. *Niech $d \in N$, $\eta \in (0,1)$ oraz $k = \lceil xn \rceil \in [n]$ będzie dowolnym wierzchołkiem w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ ($x \in (0,1)$).*
Niech $x_0 = x_0(\eta) \in (0,1)$ będzie punktem, w którym funkcja $g : (0,1) \longrightarrow \mathbb{R}$, określona wzorem

$$g(x) = \frac{1-\eta}{1+\eta}(x^{-\eta} - x) - \log(1-x) \; ,$$

posiada minimum, tj. $x_0(\eta)$ jest rozwiązaniem równania

$$(1-\eta)\eta x^{-1-\eta} + 1 - \eta = \frac{1+\eta}{1-x} \; . \tag{5}$$

Prawdopodobieństwo, że wierzchołek $k = \lceil xn \rceil$ jest wierzchołkiem izolowanym wynosi

$$\begin{aligned} p_{izol}^{(d,\eta,n)}(\lceil xn \rceil) &= (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big)x\Big]^d \\ &\quad \times \exp\Big[- \big(1+o(1)\big)\frac{1-\eta}{1+\eta} d(x^{-\eta} - x)\Big] \; . \end{aligned}$$

Co więcej, prawdopodobieństwo, że ustalony wierzchołek jest wierzchołkiem izolowanym jest największe dla wierzchołka o numerze

$$k_0(\eta) = (1+o(1))x_0(\eta)n \; .$$

Dowód. Niech $x \in (0,1)$. Wierzchołek $k = \lceil xn \rceil$ jest wierzchołkiem izolowanym wtedy i tylko wtedy, gdy w momencie ostatniego jego wyboru (w czasie trwania procesu proteuszowego) wybrał on wyłącznie wierzchołki o numerach większych niż k i żaden wierzchołek o numerze większym niż k nie wybrał tego wierzchołka. Zgodnie z Twierdzeniem 5 pierwsze zdarzenie zachodzi z prawdopodobieństwem

$$\begin{aligned}
&o(\exp(-\log^{3/2} n)) + \Big[1 - \big(1 + O(\log^{-1/2} n)\big) \sum_{i=1}^{k-1} \frac{(1-\eta)}{n} \Big(\frac{k}{i}\Big)^{\eta}\Big]^d \\
&\quad = (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big)(1-\eta) x^{\eta} \int_0^x t^{-\eta} dt\Big]^d \\
&\quad = (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big) x\Big]^d ,
\end{aligned}$$

drugie zaś z prawdopodobieństwem

$$\begin{aligned}
&o(\exp(-\log^{3/2} n)) + \prod_{i=k+1}^{n} \Big[1 - \big(1+o(1)\big)(1-\eta)\frac{d}{n}\Big(\frac{i}{k}\Big)^{\eta}\Big] \\
&\quad = (1+o(1)) \prod_{i=k+1}^{n} \exp\Big[- \big(1+o(1)\big)(1-\eta)\frac{d}{n}\Big(\frac{i}{k}\Big)^{\eta}\Big] \\
&\quad = (1+o(1)) \exp\Big[- \big(1+o(1)\big)(1-\eta)\frac{d}{n} \sum_{i=k+1}^{n} \Big(\frac{i}{k}\Big)^{\eta}\Big] \\
&\quad = (1+o(1)) \exp\Big[- \big(1+o(1)\big)(1-\eta) d x^{-\eta} \int_x^1 t^{\eta} dt\Big] \\
&\quad = (1+o(1)) \exp\Big[- \big(1+o(1)\big)\frac{1-\eta}{1+\eta} d x^{-\eta}(1 - x^{1+\eta})\Big] \\
&\quad = (1+o(1)) \exp\Big[- \big(1+o(1)\big)\frac{1-\eta}{1+\eta} d(x^{-\eta} - x)\Big] .
\end{aligned}$$

Ostatecznie więc

$$\begin{aligned}
p_{izol}^{(d,\eta,n)}(\lceil xn \rceil) \;&=\; (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big) x\Big]^d \\
&\quad \times \exp\Big[- \big(1+o(1)\big)\frac{1-\eta}{1+\eta} d(x^{-\eta} - x)\Big] .
\end{aligned}$$

Pokażemy teraz, że obliczone prawdopodobieństwo przyjmuje wartość maksymalną, gdy $x = (1 + o(1))x_0(\eta)$. Istotnie, niech $g(x)$ będzie funkcją zdefiniowaną wzorem

$$g(x) = \frac{1-\eta}{1+\eta}(x^{-\eta} - x) - \log(1-x)\,.$$

Wtedy

$$\begin{aligned} p_{izol}^{(d,\eta,n)}(k) &= (1+o(1))(1+O(\log^{-1/2} n))^d (1-x)^d \\ &\quad \times \exp\Big[-\big(1+o(1)\big)\frac{1-\eta}{1+\eta} d(x^{-\eta} - x)\Big] \\ &= (1+o(1))(1+O(\log^{-1/2} n))^d \Big[-\big(1+o(1)\big) d g(x)\Big]\,. \end{aligned}$$

Oznacza to, żeby zmaksymalizować funkcję $p_{izol}^{(d,\eta,n)}$ należy znaleźć minimum funkcji g. Policzmy zatem pierwszą oraz drugą pochodną funkcji g

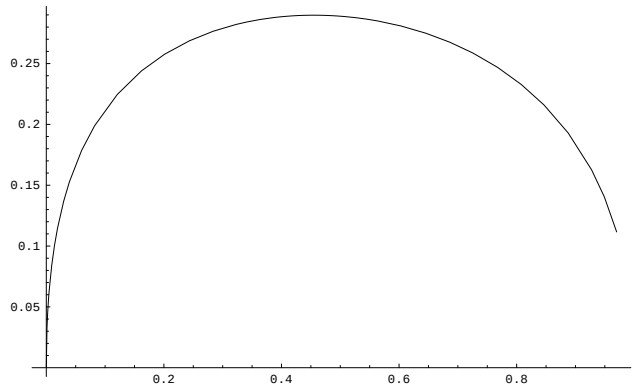
$$\begin{aligned} g'(x) &= -\frac{1-\eta}{1+\eta}\eta x^{-1-\eta} - \frac{1-\eta}{1+\eta} + \frac{1}{1-x}\,, \\ g''(x) &= (1-\eta)\eta x^{-2-\eta} + \frac{1}{(1-x)^2}\,. \end{aligned}$$

Zauważmy, że $g''(x) > 0$ dla każdego $x \in (0,1)$ oraz

$$\begin{aligned} \lim_{x\to 0^+} g'(x) &= -\infty\,, \\ \lim_{x\to 1^-} g'(x) &= +\infty\,. \end{aligned}$$

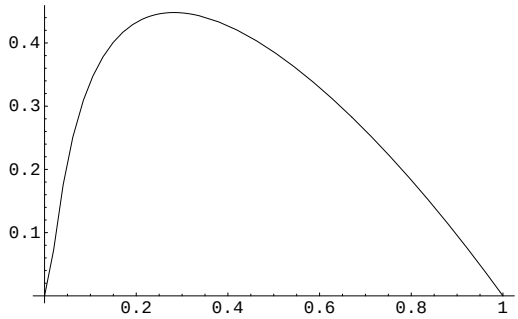
Oznacza to, że istnieje dokładne jedno rozwiązanie $x_0(\eta) \in (0,1)$ równania $g'(x) = 0$, oraz, że funkcja g posiada minimum w punkcie $x_0(\eta)$ (a tym samym funkcja $p_{izol}^{(d,\eta,n)}$ posiada maksimum w punkcie $(1+o(1))x_0(\eta)n$). □

Zależność $x_0(\eta)$ od parametru η przedstawia poniższy wykres.

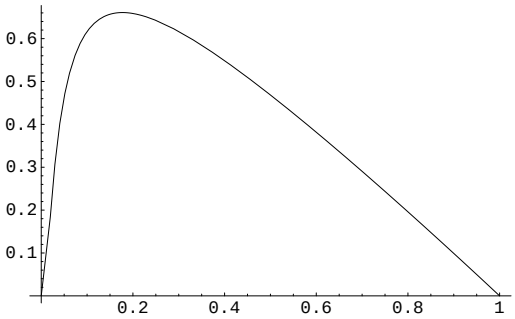


Na wykresie przedstawiono jak zmienia się $x_0(\eta)$, czyli wartość dla której prawdopodobieństwo, że wierzchołek $x_0 n$ jest wierzchołkiem izolowanym jest największe (oś OY) w zależności od parametru η (oś OX). Podkreślmy, że wykres funkcji $x_0(\eta)$ nie jest symetryczny względem prostej $\eta = 1/2$ (choć rysunek mógłby to sugerować).

W najbardziej interesujących nas przypadkach, gdy $\eta_{\text{out}} = 0{,}59$ oraz $\eta_{\text{in}} = 0{,}91$ (ze względu na możliwość modelowania sieci internetowej), szukane maksimum znajduje się odpowiednio w punktach $x_0(\eta_{\text{out}}) \approx 0{,}282$ oraz $x_0(\eta_{\text{in}}) \approx 0{,}177$. Na poniższych wykresach przedstawiono prawdopodobieństwa bycia wierzchołkiem izolowanym w zależności od położenia wierzchołka w grafie proteuszowym (dla $d = 1$).



$\eta = 0{,}59$

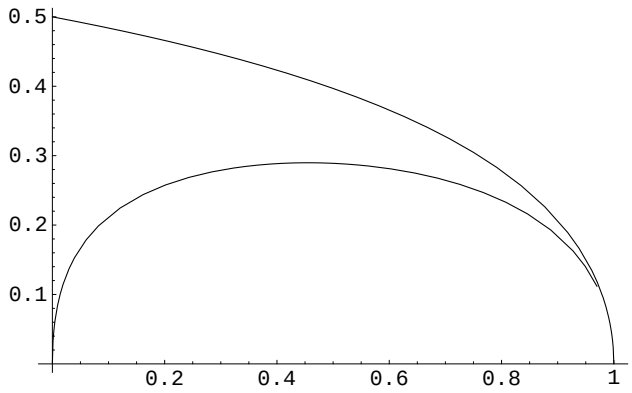


$\eta = 0{,}91$

Istnienie wierzchołków, które "przeżywają kryzys wieku średniego" można również pokazać badając oczekiwany stopień wierzchołków w zależności od położenia w grafie proteuszowym. W Twierdzeniu 9 pokazaliśmy, że najmniejszy oczekiwany stopień, w grafie proteuszowym $\mathcal{P}_n(d, \eta)$, posiada wierzchołek o numerze $(1 + o(1))c_0(\eta)n \in [n]$, gdzie

$$c_0(\eta) = \left(\frac{1-\eta}{2}\right)^{\frac{1}{\eta+1}} .$$

Zwróćmy na końcu uwagę na dość interesujący fakt, a mianowicie wierzchołki o najmniejszym oczekiwanym stopniu znajdują się w innym miejscu niż wierzchołki, dla których prawdopodobieństwo bycia wierzchołkiem izolowanym jest największe. Na poniższym wykresie przedstawiono zależność od parametru η zarówno funkcji $x_0(\eta)$ jak i $c_0(\eta)$.



Wierzchołki, których oczekiwany stopień jest najmniejszy posiadają większe numery niż wierzchołki, dla których prawdopodobieństwo bycia wierzchołkiem izolowanym jest największe (dla dowolnego $\eta \in (0, 1)$ $c_0(\eta) > x_0(\eta)$).

2.4 Wierzchołki izolowane

Policzymy teraz ile wynosi oczekiwana liczba wierzchołków izolowanych w grafie proteuszowym $\mathcal{P}_n(d,\eta)$. Liczba ta zależy od d oraz η. Dla różnych wartości parametru η (w przypadku, gdy $\eta \in (0,1)$, $\eta = 0$ oraz $\eta = 1$) oczekiwana liczba wierzchołków istotnie się różni. Dlatego te trzy przypadki rozpatrujemy w osobnych twierdzeniach.

Twierdzenie 12. *Niech $d \in \mathbb{N}$, $\eta \in (0,1)$. Niech $x_0 = x_0(\eta) \in (0,1)$ będzie punktem, w którym funkcja $g : (0,1) \longrightarrow \mathbb{R}$, określona wzorem*

$$g(x) = \frac{1-\eta}{1+\eta}(x^{-\eta} - x) - \log(1-x) \ ,$$

posiada minimum, tj. $x_0(\eta)$ jest rozwiązaniem równania

$$(1-\eta)\eta x^{-1-\eta} + 1 - \eta = \frac{1+\eta}{1-x} \ .$$

Oczekiwana liczba wierzchołków izolowanych w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ wynosi

$$\mathbb{E}Y_{izol}^{(d,\eta,n)} = n\sqrt{2\pi}\,\frac{\exp\big[-(1+o(1))dg(x_0)\big]}{\sqrt{(1+o(1))g''(x_0)d}}\big(1+O(d^{-1/2})\big)\big(1+O(\log^{-1/2} n)\big)^d \ .$$

Dowód. Zgodnie z Twierdzeniem 11 prawdopodobieństwo, że wierzchołek $k = \lceil xn \rceil \in [n]$ jest wierzchołkiem izolowanym, wynosi

$$\begin{aligned} p_{izol}^{(d,\eta,n)}(\lceil x \cdot n \rceil) \ &= \ (1+o(1))\Big[1 - \big(1+O(\log^{-1/2} n)\big)x\Big]^d \\ &\qquad \times \exp\Big[-\big(1+o(1)\big)\frac{1-\eta}{1+\eta}d(x^{-\eta}-x)\Big] \ . \end{aligned}$$

Niech Y_i, $1 \le i \le n$ będzie rodziną zmiennych losowych zdefiniowanych w następujący sposób

$$Y_{i,j} = \begin{cases} 1 & \text{gdy wierzchołek } i \text{ jest wierzchołkiem izolowanym} \\ 0 & \text{w przeciwnym razie .} \end{cases}$$

Wtedy oczekiwana liczba wierzchołków izolowanych wynosi

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,\eta,n)} &= \sum_{i=1}^{n}\mathbb{E}Y_i \\
&= (1+o(1))n\int_0^1 \Big[1-\big(1+O(\log^{-1/2}n)\big)x\Big]^d \\
&\quad \times \exp\Big[-\big(1+o(1)\big)\frac{1-\eta}{1+\eta}d(x^{-\eta}-x)\Big]dx \\
&= \big(1+O(\log^{-1/2}n)\big)^d n\int_0^1 \exp\Big[-\big(1+o(1)\big)dg(x)\Big]dx\ .
\end{aligned}$$

W dowodzie Twierdzenia 11 pokazano, że funkcja g osiąga minimum w punkcie $x_0 = x_0(\eta) \in (0,1)$, gdzie $x_0(\eta)$ jest pierwiastkiem równania

$$(1-\eta)\eta x^{-1-\eta} + 1 - \eta = \frac{1+\eta}{1-x}\ .$$

Rozwijając funkcję g w szereg Taylora (w pobliżu punktu x_0) otrzymujemy, że dla dowolnego $x \in (0,1)$

$$g(x) = g(x_0) + \frac{g''(x_0)}{2}(x-x_0)^2 + O((x-x_0)^3)\ ,$$

co oznacza, że

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,\eta,n)} &= \big(1+O(\log^{-1/2}n)\big)^d n\int_0^1 \exp\Big[-\big(1+o(1)\big)dg(x_0) \\
&\qquad -\big(1+o(1)\big)d\frac{g''(x_0)}{2}(x-x_0)^2 + O\big(d(x-x_0)^3\big)\Big]dx \\
&= \big(1+O(\log^{-1/2}n)\big)^d n\exp\big[-\big(1+o(1)\big)dg(x_0)\big] \\
&\qquad \int_0^1 \exp\Big[-\big(1+o(1)\big)d\frac{g''(x_0)}{2}(x-x_0)^2 + O\big(d(x-x_0)^3\big)\Big]dx\ .
\end{aligned}$$

Podstawiając najpierw za $x = y + x_0$, a później za

$$y = z\Big[\big(1+o(1)\big)g''(x_0)d\Big]^{-1/2},$$

otrzymujemy

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,\eta,n)} &= \big(1+O(\log^{-1/2} n)\big)^d n \exp\big[-\big(1+o(1)\big)dg(x_0)\big] \\
&\quad \times \int_{-x_0}^{1-x_0} \exp\Big[-\big(1+o(1)\big)d\frac{g''(x_0)}{2}y^2 + O\big(dy^3\big)\Big]dy \\
&= \big(1+O(\log^{-1/2} n)\big)^d n \frac{\exp\big[-(1+o(1))dg(x_0)\big]}{\sqrt{(1+o(1))g''(x_0)d}} \\
&\quad \times \int_{-x_0\sqrt{(1+o(1))g''(x_0)d}}^{(1-x_0)\sqrt{(1+o(1))g''(x_0)d}} \exp\Big[-\frac{z^2}{2} + O\big(\frac{z^3}{\sqrt{d}}\big)\Big]dz\,.
\end{aligned}$$

Niech $\varepsilon > 0$ będzie dowolną liczbą taką, że

$$d^{\varepsilon} < \min(x_0, 1-x_0)\sqrt{(1+o(1))g''(x_0)d}\,.$$

Wtedy

$$\mathbb{E}Y_{izol}^{(d,\eta,n)} = \big(1+O(\log^{-1/2} n)\big)^d n \frac{\exp\big[-(1+o(1))dg(x_0)\big]}{\sqrt{(1+o(1))g''(x_0)d}}(\gamma_1+\gamma_2+\gamma_3)\,,$$

gdzie

$$\begin{aligned}
\gamma_1 &= \int_{-x_0\sqrt{(1+o(1))g''(x_0)d}}^{-d^{\varepsilon}} \exp\Big[-\frac{z^2}{2} + O\big(\frac{z^3}{\sqrt{d}}\big)\Big]dz\,, \\
\gamma_2 &= \int_{-d^{\varepsilon}}^{d^{\varepsilon}} \exp\Big[-\frac{z^2}{2} + O\big(\frac{z^3}{\sqrt{d}}\big)\Big]dz\,, \\
\gamma_3 &= \int_{d^{\varepsilon}}^{(1-x_0)\sqrt{(1+o(1))g''(x_0)d}} \exp\Big[-\frac{z^2}{2} + O\big(\frac{z^3}{\sqrt{d}}\big)\Big]dz\,.
\end{aligned}$$

O zachowaniu $\mathbb{E}Y_{izol}^{(d,\eta,n)}$ decyduje wyraz γ_2, który możemy oszacować w następujący sposób

$$\begin{aligned}\gamma_2 &= \int_{-d^{\varepsilon}}^{d^{\varepsilon}} \exp\Big[-\frac{z^2}{2}+O\big(\frac{(d^{\varepsilon})^3}{\sqrt{d}}\big)\Big]dz \\ &= \big(1+O(d^{-1/2+3\varepsilon})\big)\int_{-d^{\varepsilon}}^{d^{\varepsilon}} \exp\Big[-\frac{z^2}{2}\Big]dz \\ &= \big(1+O(d^{-1/2+3\varepsilon})\big)\int_{-\infty}^{\infty} \exp\Big[-\frac{z^2}{2}\Big]dz \\ &= \sqrt{2\pi}\big(1+O(d^{-1/2+3\varepsilon})\big)\,.\end{aligned}$$

Oszacujmy teraz γ_1

$$\begin{aligned}\gamma_1 &\leq \int_{-x_0\sqrt{(1+o(1))g''(x_0)d}}^{-d^{\varepsilon}} \exp\Big[-\frac{d^{2\varepsilon}}{2}+O\big(\frac{d^{3\varepsilon}}{\sqrt{d}}\big)\Big]dz \\ &\leq O(d^{1/2})\,\exp\Big[-\frac{d^{2\varepsilon}}{2}+O\big(\frac{d^{3\varepsilon}}{\sqrt{d}}\big)\Big] \\ &= o(d^{-9})\,.\end{aligned}$$

Analogicznie można pokazać, że $\gamma_3 = o(d^{-9})$, a zatem na mocy dowolności wyboru ε otrzymujemy

$$\mathbb{E}Y_{izol}^{(d,\eta,n)} = n\sqrt{2\pi}\,\frac{\exp\big[-(1+o(1))dg(x_0)\big]}{\sqrt{(1+o(1))g''(x_0)d}}\big(1+O(d^{-1/2})\big)\big(1+O(\log^{-1/2} n)\big)^d\,.$$

□

Twierdzenie 13. *Niech $d \in \mathbb{N}$. Oczekiwana liczba wierzchołków izolowanych, w grafie proteuszowym $\mathcal{P}_n(d,0)$, wynosi*

$$\mathbb{E}Y_{izol}^{(d,0,n)} = ne^{-d}\sqrt{\frac{\pi}{2d}}\big(1+O(d^{-1/2})\big)\,.$$

Dowód. Zauważmy, że w tym szczególnym przypadku grafu proteuszowego (w przypadku, gdy $\eta = 0$) waga każdego wierzchołka wynosi $1/n$. Oznacza to, że prawdopodobieństwo istnienia w $\mathcal{P}_n(d,0)$ dowolnej krawędzi $\{i,j\}$ wynosi

$$p^{(d,0,n)}(j,i) = (1+o(1))d\cdot p^{(1,0,n)}(j,i) = (1+o(1))\frac{d}{n}\,.$$

Policzmy teraz prawdopodobieństwo, że ustalony wierzchołek jest wierzchołkiem izolowanym. Niech $x \in (0,1)$. Wierzchołek $k = \lceil xn \rceil \in [n]$ jest wierzchołkiem izolowanym wtedy i tylko wtedy, gdy w momencie ostatniego jego wyboru (w czasie trwania procesu proteuszowego) wybrał on wyłącznie wierzchołki o numerach większych niż k. Co więcej, żaden wierzchołek o numerze większym niż k nie wybrał tego wierzchołka. Pierwsze zdarzenie zachodzi z prawdopodobieństwem

$$\Big[1 - \sum_{i=1}^{k-1} \frac{1}{n}\Big]^d = (1 - x + O(1/n))^d = (1 + o(1))(1-x)^d \,,$$

drugie zaś z prawdopodobieństwem

$$\begin{aligned} \prod_{i=k+1}^{n} \Big(1 - \frac{d}{n}\Big) &= (1+o(1)) \prod_{i=k+1}^{n} \exp\Big(-\frac{d}{n}\Big) \\ &= (1+o(1)) \exp\Big(-\sum_{i=k+1}^{n} \frac{d}{n}\Big) \\ &= (1+o(1)) \exp\Big[-d(1-x)\Big] \,. \end{aligned}$$

Ostatecznie więc

$$p_{izol}^{(d,0,n)}(k) = (1+o(1))(1-x)^d \exp\Big[-d(1-x)\Big] \,.$$

Niech Y_i, $1 \le i \le n$ będzie rodziną zmiennych losowych zdefiniowanych w następujący sposób

$$Y_{i,j} = \begin{cases} 1 & \text{gdy wierzchołek } i \text{ jest wierzchołkiem izolowanym} \\ 0 & \text{w przeciwnym razie .} \end{cases}$$

Wtedy oczekiwana liczba wierzchołków izolowanych wynosi

$$\begin{aligned} \mathbb{E}Y_{izol}^{(d,0,n)} &= \sum_{i=1}^{n} \mathbb{E}Y_i \\ &= (1+o(1))n \int_0^1 (1-x)^d \exp\Big[-d(1-x)\Big]dx \\ &= (1+o(1))ne^{-d} \int_0^1 \Big[(1-x)e^x\Big]^d dx \,. \end{aligned}$$

Wprowadźmy pomocniczą funkcję $f : (0,1) \longrightarrow \mathbb{R}$, określoną następującym wzorem

$$f(x) = (1-x)e^x \ .$$

Rozwijając funkcję f w szereg Taylora (w pobliżu punktu $x_0 = 0$) otrzymujemy, że dla dowolnego $x \in (0,1)$

$$\begin{aligned} f(x) &= f(x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + O((x-x_0)^3) \\ &= 1 - \frac{1}{2}x^2 + O(x^3) \ , \end{aligned}$$

a zatem

$$\begin{aligned} \mathbb{E}Y_{izol}^{(d,0,n)} &= (1+o(1))ne^{-d}\int_0^1 \Big[1 - \frac{1}{2}x^2 + O(x^3)\Big]^d dx \\ &= (1+o(1))ne^{-d}\int_0^1 \exp\Big[-\frac{d}{2}x^2 + O(dx^3)\Big]dx \ . \end{aligned}$$

Podstawiając za $x = y/\sqrt{d}$ otrzymujemy

$$\mathbb{E}Y_{izol}^{(d,0,n)} = (1+o(1))\frac{ne^{-d}}{\sqrt{d}}\int_0^{\sqrt{d}} \exp\Big[-\frac{y^2}{2} + O(\frac{y^3}{\sqrt{d}})\Big]dy \ .$$

Niech $0 < \varepsilon < 1/2$ będzie dowolną liczbą, wtedy

$$\mathbb{E}Y_{izol}^{(d,0,n)} = (1+o(1))\frac{ne^{-d}}{\sqrt{d}}(\gamma_1 + \gamma_2) \ ,$$

gdzie

$$\begin{aligned} \gamma_1 &= \int_0^{d^\varepsilon} \exp\Big[-\frac{y^2}{2} + O(\frac{y^3}{\sqrt{d}})\Big]dy \ , \\ \gamma_2 &= \int_{d^\varepsilon}^{\sqrt{d}} \exp\Big[-\frac{y^2}{2} + O(\frac{y^3}{\sqrt{d}})\Big]dy \ . \end{aligned}$$

Obliczmy najpierw główny wyraz γ_1

$$\begin{aligned} \gamma_1 &= \int_0^{d^\varepsilon} \exp\Big[-\frac{y^2}{2} + O(\frac{(d^\varepsilon)^3}{\sqrt{d}})\Big]dy \\ &= \big(1 + O(d^{-1/2+3\varepsilon})\big)\int_0^{d^\varepsilon} \exp\Big[-\frac{y^2}{2}\Big]dy \\ &= \big(1 + O(d^{-1/2+3\varepsilon})\big)\int_0^{\infty} \exp\Big[-\frac{y^2}{2}\Big]dy \\ &= \big(1 + O(d^{-1/2+3\varepsilon})\big)\frac{\sqrt{2\pi}}{2} \ . \end{aligned}$$

Dla γ_2 otrzymujemy

$$\begin{aligned}\gamma_2 &\leq \int_{d^\varepsilon}^{\sqrt{d}} \exp\Big[-\frac{d^{2\varepsilon}}{2}+O(\frac{d^{3\varepsilon}}{\sqrt{d}})\Big]dy \\ &\leq \sqrt{d}\exp\Big[-\frac{d^{2\varepsilon}}{2}+O(\frac{d^{3\varepsilon}}{\sqrt{d}})\Big]dy \\ &= o(d^{-9})\,.\end{aligned}$$

Na mocy dowolności wyboru ε otrzymujemy

$$\begin{aligned}\mathbb{E}Y_{izol}^{(d,0,n)} &= \frac{ne^{-d}}{\sqrt{d}}\big(1+O(d^{-1/2})\big)\frac{\sqrt{2\pi}}{2} \\ &= ne^{-d}\sqrt{\frac{\pi}{2d}}\big(1+O(d^{-1/2})\big)\,.\end{aligned}$$

□

Twierdzenie 14. *Niech $d \in \mathbb{N}$. Oczekiwana liczba wierzchołków, w grafie proteuszowym $\mathcal{P}_n(d,1)$, wynosi*

$$\begin{aligned}\mathbb{E}Y_{izol}^{(d,1,n)} &= n2^{-1/4}\sqrt{\pi}d^{-1/2}\log^{-1/4}n\,\exp\Big(\frac{-(1+o(1))d\sqrt{2}}{\sqrt{\log n}}\Big) \\ &\qquad \big(1+O(\log^{-1/2}n)\big)^d\big(1+O(d^{-1/2}\log^{1/4}n)\big)\,.\end{aligned}$$

Dowód. Policzmy najpierw ile wynosi suma wag wszystkich wierzchołków w grafie proteuszowym $\mathcal{P}_n(d,1)$. Przybliżając sumę przez całkę otrzymamy

$$\begin{aligned}W=\sum_{i=1}^{n}\frac{1}{i} &\geq \int_1^{n+1}\frac{dx}{x}=\log(n+1)\,, \\ W &\leq 1+\int_1^n\frac{dx}{x}=1+\log n\,.\end{aligned}$$

Oznacza to, że

$$W=\big(1+O(\log^{-1}n)\big)\log n\,.$$

Prawdopodobieństwo występowania (bądź nie) krawędzi pomiędzy dowolnym wierzchołkiem j a wierzchołkiem i zależy od zmiennej losowej $U^n(j,i)$

(położenia wierzchołka i w momencie, gdy po raz ostatni wybrano wierzchołek j). Z Twierdzenia 2 wynika, że dla $\varepsilon = \log^{-1/2} n$

$$\mathbb{P}\Big(\frac{in}{j}(1-\varepsilon) \le U^n(j,i) \le \frac{in}{j}(1+\varepsilon)\Big) = 1 - o(\exp(-\log^{3/2} n)).$$

Oznacza to, że z prawdopodobieństwem $1 - o(\exp(-\log^{3/2} n))$ położenie wierzchołka i w momencie, gdy po raz ostatni wybrano wierzchołek j wynosi

$$\frac{in}{j}\Big(1 + O(\frac{1}{\sqrt{\log n}})\Big)\ .$$

Z tej obserwacji wynika, że prawdopodobieństwo, że w grafie proteuszowym istnieje dowolna krawędź $\{i,j\}$ $(j > i)$ wynosi

$$\begin{aligned} p^{(d,1,n)}(j,i) &= d\Big(\frac{in}{j}\big(1 + O(\log^{-1/2} n)\big)\Big)^{-1}/W \\ &= \big(1 + O(\log^{-1/2} n)\big)\frac{d}{n\log n}\frac{j}{i}\ . \end{aligned}$$

Policzmy teraz prawdopodobieństwo, że ustalony wierzchołek jest wierzchołkiem izolowanym. Niech $x \in (0,1)$. Wierzchołek $k = \lceil xn \rceil \in [n]$ jest wierzchołkiem izolowanym wtedy i tylko wtedy, gdy w momencie ostatniego jego wyboru (w czasie trwania procesu proteuszowego) wybrał on wyłącznie wierzchołki o numerach większych niż k. Co więcej, żaden wierzchołek o numerze większym niż k nie wybrał tego wierzchołka. Pierwsze zdarzenie zachodzi z prawdopodobieństwem

$$\begin{aligned} o(\exp(-\log^{3/2} n)) &+ \Big[1 - \sum_{i=1}^{k-1} \frac{\big(1 + O(\log^{-1/2} n)\big)}{n\log n}\frac{k}{i}\Big]^d \\ &= (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big)\frac{x}{\log n}\sum_{i=1}^{k-1}\frac{1}{i}\Big]^d \\ &= (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big)\frac{x}{\log n}\log(xn)\Big]^d \\ &= (1+o(1))\left[1 - \big(1 + O(\log^{-1/2} n)\big)x\Big(1 + \frac{\log x}{\log n}\Big)\right]^d \\ &= (1+o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big)x\Big]^d\ , \end{aligned}$$

drugie zaś z prawdopodobieństwem

$$\begin{aligned} o(\exp(-\log^{3/2} n)) &+ \prod_{i=k+1}^{n} \Big[1 - \big(1+o(1)\big)\frac{d}{n\log n}\frac{i}{k}\Big] \\ &= (1+o(1)) \prod_{i=k+1}^{n} \exp\Big[-\big(1+o(1)\big)\frac{d}{n\log n}\frac{i}{k}\Big] \\ &= (1+o(1))\exp\Big[-\big(1+o(1)\big)\frac{d}{xn^2\log n}\sum_{i=k+1}^{n} i\Big] \\ &= (1+o(1))\exp\Big[-\big(1+o(1)\big)\frac{d}{xn^2\log n}\frac{k+1+n}{2}(n-k)\Big] \\ &= (1+o(1))\exp\Big[-\big(1+o(1)\big)\frac{d(1+x)(1-x)}{2x\log n}\Big]. \end{aligned}$$

Ostatecznie więc

$$\begin{aligned} p_{izol}^{(d,1,n)}(k) &= (1+o(1))\Big[1-\big(1+O(\log^{-1/2} n)\big)x\Big]^d \\ &\qquad \exp\Big[-\big(1+o(1)\big)\frac{d(1+x)(1-x)}{2x\log n}\Big]. \end{aligned}$$

Niech Y_i, $1 \le i \le n$ będzie rodziną zmiennych losowych zdefiniowanych w następujący sposób

$$Y_{i,j} = \begin{cases} 1 & \text{gdy wierzchołek } i \text{ jest wierzchołkiem izolowanym} \\ 0 & \text{w przeciwnym razie .} \end{cases}$$

Wtedy oczekiwana liczba wierzchołków izolowanych wynosi

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,1,n)} &= \sum_{i=1}^{n} \mathbb{E}Y_i \\
&= (1+o(1))n \int_0^1 \Big[1-\big(1+O(\log^{-1/2} n)\big)x\Big]^d \\
&\qquad \exp\Big[-\big(1+o(1)\big)\frac{d(1-x^2)}{2x\log n}\Big]dx \\
&= \big(1+O(\log^{-1/2} n)\big)^d n \int_0^1 \exp\left[(1+o(1))d\Big(-x-\frac{1-x^2}{2x\log n}\Big)\right]dx \\
&= \big(1+O(\log^{-1/2} n)\big)^d n \\
&\qquad \int_0^1 \exp\left[(1+o(1))d\Big(-x\Big(1-\frac{1}{2\log n}\Big)-\frac{1}{2x\log n}\Big)\right]dx \\
&= \big(1+O(\log^{-1/2} n)\big)^d n \int_0^1 \exp\left[(1+o(1))d\Big(-x-\frac{1}{2x\log n}\Big)\right]dx\,.
\end{aligned}$$

Podstawiając za $x = y/\sqrt{\log n}$ otrzymujemy

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,1,n)} &= \big(1+O(\log^{-1/2} n)\big)^d \frac{n}{\sqrt{\log n}} \\
&\qquad \int_0^{\sqrt{\log n}} \exp\left[(1+o(1))\Big(-y-\frac{1}{2y}\Big)\frac{d}{\sqrt{\log n}}\right]dy\,.
\end{aligned}$$

Wprowadźmy pomocniczą funkcję $h : (0,1) \longrightarrow \mathbb{R}$, określoną następującym wzorem

$$h(y) = -y - \frac{1}{2y}\,.$$

Rozwijając funkcję h w szereg Taylora (w pobliżu punktu $y_0 = 1/\sqrt{2}$), dla dowolnego $y \in (0,1)$ otrzymujemy

$$\begin{aligned}
h(y) &= h(y_0) + \frac{h''(y_0)}{2}(y-y_0)^2 + O((y-y_0)^3) \\
&= -\sqrt{2}\left[1+\Big(y-\frac{1}{\sqrt{2}}\Big)^2 + O\Big(\Big(y-\frac{1}{\sqrt{2}}\Big)^3\Big)\right].
\end{aligned}$$

Wykorzystując powyższy wzór, a następnie wykonując podstawienie za $y = z + 1/\sqrt{2}$ oraz za $z = t2^{-3/4}\log^{1/4} n d^{-1/2}$ otrzymujemy

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,1,n)} &= \big(1+O(\log^{-1/2} n)\big)^d \frac{n}{\sqrt{\log n}} \\
&\quad \int_0^{\sqrt{\log n}} \exp\left[(1+o(1))\Big(1+\Big(y-\frac{1}{\sqrt{2}}\Big)^2 + O\Big(\Big(y-\frac{1}{\sqrt{2}}\Big)^3\Big)\Big)\frac{-d\sqrt{2}}{\sqrt{\log n}}\right] dy \\
&= \big(1+O(\log^{-1/2} n)\big)^d n \frac{\exp(\frac{-(1+o(1))d\sqrt{2}}{\sqrt{\log n}})}{\sqrt{\log n}} \\
&\quad \int_0^{\sqrt{\log n}} \exp\left[(1+o(1))\Big(\Big(y-\frac{1}{\sqrt{2}}\Big)^2 + O\Big(\Big(y-\frac{1}{\sqrt{2}}\Big)^3\Big)\Big)\frac{-d\sqrt{2}}{\sqrt{\log n}}\right] dy \\
&= \big(1+O(\log^{-1/2} n)\big)^d n \frac{\exp(\frac{-(1+o(1))d\sqrt{2}}{\sqrt{\log n}})}{\sqrt{\log n}} \\
&\quad \int_{-1/\sqrt{2}}^{\sqrt{\log n}-1/\sqrt{2}} \exp\left[(1+o(1))\big(z^2+O(z^3)\big)\frac{-d\sqrt{2}}{\sqrt{\log n}}\right] dz \\
&= \big(1+O(\log^{-1/2} n)\big)^d n 2^{-3/4} d^{-1/2}\log^{-1/4} n \exp\Big(\frac{-(1+o(1))d\sqrt{2}}{\sqrt{\log n}}\Big) \\
&\quad \int_{-O(d^{1/2}\log^{-1/4} n)}^{O(d^{1/2}\log^{1/4} n)} \exp\Big[-\frac{t^2}{2}+O(t^3 d^{-1/2}\log^{1/4} n)\Big] dt\,.
\end{aligned}$$

Niech $\varepsilon > 0$ będzie dowolną liczbą taką, że $d^\varepsilon = O(d^{1/2}\log^{-1/4} n)$. Wtedy

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,1,n)} &= \big(1+O(\log^{-1/2} n)\big)^d n 2^{-3/4} d^{-1/2}\log^{-1/4} n \\
&\qquad \exp\Big(\frac{-(1+o(1))d\sqrt{2}}{\sqrt{\log n}}\Big)(\gamma_1+\gamma_2+\gamma_3)\,,
\end{aligned}$$

gdzie

$$\begin{aligned}
\gamma_1 &= \int_{-O(d^{1/2}\log^{-1/4} n)}^{-d^\varepsilon} \exp\Big[-\frac{t^2}{2}+O(t^3 d^{-1/2}\log^{1/4} n)\Big] dt\,, \\
\gamma_2 &= \int_{-d^\varepsilon}^{d^\varepsilon} \exp\Big[-\frac{t^2}{2}+O(t^3 d^{-1/2}\log^{1/4} n)\Big] dt\,, \\
\gamma_3 &= \int_{d^\varepsilon}^{O(d^{1/2}\log^{1/4} n)} \exp\Big[-\frac{t^2}{2}+O(t^3 d^{-1/2}\log^{1/4} n)\Big] dt\,.
\end{aligned}$$

Obliczmy najpierw główny wyraz γ_2

$$\begin{aligned}
\gamma_2 &= \int_{-d^{\varepsilon}}^{d^{\varepsilon}} \exp\Big[-\frac{t^2}{2}+O(d^{3\varepsilon}d^{-1/2}\log^{1/4}n)\Big]dt \\
&= \big(1+O(d^{-1/2+3\varepsilon}\log^{1/4}n)\big)\int_{-d^{\varepsilon}}^{d^{\varepsilon}} \exp\Big[-\frac{t^2}{2}\Big]dt \\
&= \big(1+O(d^{-1/2+3\varepsilon}\log^{1/4}n)\big)\int_{-\infty}^{\infty} \exp\Big[-\frac{t^2}{2}\Big]dt \\
&= \sqrt{2\pi}\big(1+O(d^{-1/2+3\varepsilon}\log^{1/4}n)\big)
\end{aligned}$$

Oszacujmy teraz γ_3.

$$\begin{aligned}
\gamma_3 &\leq \int_{d^{\varepsilon}}^{O(d^{1/2}\log^{1/4}n)} \exp\Big[-\frac{d^{2\varepsilon}}{2}+O(d^{3\varepsilon}d^{-1/2}\log^{1/4}n)\Big]dt \\
&\leq O(d^{1/2}\log^{1/4}n)\exp\Big[-\frac{d^{2\varepsilon}}{2}+O(d^{3\varepsilon}d^{-1/2}\log^{1/4}n)\Big] \\
&= o(d^{-9})
\end{aligned}$$

Analogicznie możemy pokazać, że $\gamma_1 = o(d^{-9})$, a zatem na mocy dowolności wyboru ε otrzymujemy

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,1,n)} &= n2^{-1/4}\sqrt{\pi}d^{-1/2}\log^{-1/4}n\exp\Big(\frac{-(1+o(1))d\sqrt{2}}{\sqrt{\log n}}\Big) \\
&\qquad \big(1+O(\log^{-1/2}n)\big)^d\big(1+O(d^{-1/2}\log^{1/4}n)\big)\ .
\end{aligned}$$

□

3 Składowe spójności grafu proteuszowego

W rzeczywistych sieciach (w szczególności w sieciach internetowych) istotne jest, aby dowolna para wierzchołków była połączona ścieżką i, jeśli to możliwe, by długość tej ścieżki była jak najmniejsza. Naturalne więc wydają się pytania o próg spójności czy o średnicę największej składowej. Spójność oraz stosunkowo niewielka średnica dają nadzieję na szybką zbieżność algorytmów działających na tego typu strukturach. W niniejszym rozdziale zajmiemy się tymi zagadnieniami.

3.1 Funkcja progowa dla spójności

Niech $\rho_n(d,\eta)$ oznacza prawdopodobieństwo, że graf proteuszowy $\mathcal{P}_n(d,\eta)$ jest spójny. Nasze rozważania rozpoczniemy od najprostszego przypadku, gdy $\eta = 0$. Oznacza to, że w każdym kroku procesu proteuszowego $\mathfrak{P}_n(d,0)$ wszystkie wierzchołki posiadają identyczną wagę, bez względu na ich położenie w permutacji (tj. bez względu na ich "wiek"). W tym przypadku prawdopodobieństwo, że dwa dowolne wierzchołki są połączone krawędzią wynosi

$$p(i,j) = p'(n) = 1 - (1-1/n)^d = d/n + O(d^2/n^2)\ .$$

Mogłoby się wydawać zatem, iż próg spójności w tym grafie będzie identyczny jak w standardowym modelu $G(n,p')$. Okazuje się jednak, że struktura zależności grafu proteuszowego sprawia, że funkcje progowe dla spójności w przypadku grafów $\mathcal{P}_n(d,\eta)$ i $G(n,p')$ różnią się wyrazami drugiego rzędu (przypomnijmy, że dla $G(n,p)$ funkcja ta wynosi $p(n) = (\log n + O(1))/n$, patrz Twierdzenie 7.3 [6]).

Twierdzenie 15. *Niech* $d = d(n) = \log n - \frac{1}{2}\log\log n + a(n)$. *Wtedy*

$$\lim_{n\to\infty} \rho_n(d,0) = \begin{cases} 1 & \text{gdy} \quad a(n) \to \infty \\ \exp\big(-\sqrt{\pi/2}e^{-a}\big) & \text{gdy} \quad a(n) \to a \\ 0 & \text{gdy} \quad a(n) \to -\infty\,. \end{cases}$$

Dowód. Zgodnie z Twierdzeniem 13 oczekiwana liczba $Y_{izol}^{(d,0,n)}$ wierzchołków izolowanych, w grafie proteuszowym $\mathcal{P}_n(d,0)$, wynosi

$$\begin{aligned}
\mathbb{E}Y_{izol}^{(d,0,n)} &= \big(1+O(d^{-1/2})\big)ne^{-d}\sqrt{\frac{\pi}{2d}} \\
&= \big(1+O(\log^{-1/2} n)\big)ne^{-\log n+\frac{1}{2}\log\log n-a(n)} \\
&\qquad \times\sqrt{\frac{\pi}{2(\log n-\frac{1}{2}\log\log n+a(n))}} \\
&= \big(1+o(1)\big)e^{-a(n)}\sqrt{\frac{\pi}{2}}\,.
\end{aligned}$$

Co więcej, można pokazać, że dla dowolnego $r \geq 2$, r-ty moment silniowy zmiennej losowej $Y_{izol}^{(d,0,n)}$, dąży do $e^{-r\cdot a(n)}(\pi/2)^{r/2}$. Istotnie, podobnie jak w dowodzie Twierdzenia 13, z Twierdzenia 5 wynika, że prawdopodobieństwo, że ustalona r-ka wierzchołków o numerach $k_1,\ldots,k_r$, gdzie $k_i=\lceil x_i n\rceil \in [n]$ oraz $0\leq x_i\leq 1$ dla $i=1,2,\ldots,r$, jest izolowana wynosi

$$(1+o(1))\prod_{i=1}^{r} p_{izol}^{(d,0,n)}(k_i)\,,$$

gdzie

$$p_{izol}^{(d,0,n)}(k) = (1-x)^d\,\exp\Big[-d(1-x)\Big]\,.$$

Stąd

$$\begin{aligned}
\mathbb{E}_r Y_{izol}^{(d,0,n)} &= Y_{izol}^{(d,0,n)}(Y_{izol}^{(d,0,n)}-1)\ldots(Y_{izol}^{(d,0,n)}-r+1) \\
&= (1+o(1))n^r\int_0^1\ldots\int_0^1 p_{izol}^{(d,0,n)}(x_1 n)\ldots p_{izol}^{(d,0,n)}(x_r n)\,dx_r\ldots dx_1 \\
&= (1+o(1))n^r\Big(e^{-d}\sqrt{\frac{\pi}{2d}}\big(1+O(d^{-1/2})\big)\Big)^r \\
&= (1+o(1))e^{-r\cdot a(n)}(\pi/2)^{r/2}\,.
\end{aligned}$$

Oznacza to, że zmienna losowa $Y_{izol}^{(d,0,n)}$ dąży do zmiennej losowej o rozkładzie Poissona z parametrem $e^{-a(n)}\sqrt{\pi/2}$, w szczególności zaś, że prawdopodobieństwo, że graf proteuszowy $\mathcal{P}_n(d,0)$ nie zawiera wierzchołków izolowanych dąży do $\exp\big(-e^{-a(n)}\sqrt{\pi/2}\big)$, przy $n\to\infty$.

Pokażemy teraz, że dla dowolnego $d = d(n) > 0,99 \log n$, graf proteuszowy $\mathcal{P}_n(d,0)$ składa się z dużej składowej oraz, być może, z pewnej liczby wierzchołków izolowanych. W grafie proteuszowym $\mathcal{P}_n(d,0)$ nie ma składowych o rozmiarze k, $2 \le k \le 2n/3$. Istotnie, na $\binom{n}{k}$ sposobów można wybrać wierzchołki wchodzące w skład takiej składowej. Na ustalonych k wierzchołkach można utworzyć k^{k-2} różnych drzew rozpinających tę składową (wzór Cayley'a). Zauważmy, że co najwyżej $2k/\sqrt{d}$ wierzchołków drzewa ma stopień większy niż $\sqrt{d}$. Oznacza to, że prawdopodobieństwo, że żaden wierzchołek drzewa nie jest połączony z wierzchołkami na zewnątrz drzewa można oszacować z góry przez (patrz uwaga po Twierdzeniu 6)

$$(1-d/n)^{(k-2k/\sqrt{d})(n-k)} = (1-d/n)^{(1+o(1))k(n-k)},$$

a prawdopodobieństwo istnienia $k-1$ krawędzi tego drzewa przez $(d/n)^{k-1}$. Zatem prawdopodobieństwo, że $\mathcal{P}_n(d,0)$ zawiera składową o rozmiarze k, $2 \le k \le 2n/3$, można oszacować z góry przez

$$\begin{aligned}
& \sum_{k=2}^{2n/3} \binom{n}{k} k^{k-2} (1-d/n)^{(1+o(1))k(n-k)} (d/n)^{k-1} \\
\le\ & \sum_{k=2}^{2n/3} \Big(\frac{en}{k}\Big)^k k^{k-2} \exp\big(-(1+o(1))dk(n-k)/n\big)(d/n)^{k-1} \\
\le\ & \sum_{k=2}^{2n/3} nk^{-2} e^k d^{k-1} \exp\big(-(1+o(1))dk(n-k)/n\big) \\
\le\ & \sum_{k=2}^{2n/3} ne^k \log^k n \exp\big(-0,98k(n-k)\log n/n\big) \\
\le\ & ne^2 \log^2 n \cdot n^{-1,94} + ne^3 \log^3 n \cdot n^{-2,91} + n \sum_{k=4}^{2n/3} \Big(e \log n \cdot n^{-0,32}\Big)^k \\
\le\ & o(n^{-0,25}) + ne^4 \log^4 n \cdot n^{-4\cdot 0,32} \sum_{k=0}^{\infty} \Big(e \log n \cdot n^{-0,32}\Big)^k \\
=\ & o(n^{-0,25}) + o(n^{-0,25})(1+o(1)) = o(n^{-0,25}) \,.
\end{aligned}$$

Oznacza to, że prawie na pewno graf proteuszowy $\mathcal{P}_n(d,0)$ składa się wyłącznie z dużej składowej oraz (być może) wierzchołków izolowanych, co kończy dowód. □

Znalezienie funkcji progowej dla spójności $\mathcal{P}_n(d,\eta)$, gdy $\eta \in (0,1)$ jest zadaniem nieco trudniejszym. Poniższe twierdzenie szacuje tę funkcję z dokładnością do czynnika $1+o(1)$.

Twierdzenie 16. *Niech $\eta \in (0,1)$, $d = d(n) = a\log n$, gdzie $a > 0$ jest stałą. Wtedy*

$$\lim_{n\to\infty} \rho_n(d,\eta) = \begin{cases} 1 & \text{gdy} \quad a > 1/g(x_0(\eta)) \\ 0 & \text{gdy} \quad a < 1/g(x_0(\eta))\,, \end{cases}$$

gdzie $x_0 = x_0(\eta) \in (0,1)$ będzie punktem, w którym funkcja $g : (0,1) \longrightarrow \mathbb{R}$, określona wzorem

$$g(x) = \frac{1-\eta}{1+\eta}(x^{-\eta} - x) - \log(1-x)\,,$$

posiada minimum, tj. $x_0(\eta)$ jest rozwiązaniem równania

$$(1-\eta)\eta x^{-1-\eta} + 1 - \eta = \frac{1+\eta}{1-x}\,.$$

Dowód. Dowód tego twierdzenia jest oparty na rozumowaniu podobnym do tego, którego użyliśmy do pokazania Twierdzenia 15. Zgodnie z Twierdzeniem 12 oczekiwana liczba $Y_{izol}^{(d,\eta,n)}$ wierzchołków izolowanych, w grafie proteuszowym $\mathcal{P}_n(d,\eta)$, wynosi

$$\begin{aligned} \mathbb{E}Y_{izol}^{(d,\eta,n)} &= n\sqrt{2\pi}\,\frac{\exp\big[-(1+o(1))dg(x_0)\big]}{\sqrt{(1+o(1))g''(x_0)d}} \\ &\qquad \times\big(1+O(d^{-1/2})\big)\big(1+O(\log^{-1/2} n)\big)^d \\ &= O(1)\,n^{1+o(1)}d^{-1/2}\exp\big[-(1+o(1))dg(x_0)\big]\,. \end{aligned}$$

Oznacza to, że gdy $d = d(n) = a \log n$, gdzie $a > 1/g(x_0(\eta))$ oczekiwana liczba wierzchołków izolowanych dąży do zera. Zatem, z nierówności Markowa wynika, że prawie na pewno graf $\mathcal{P}_n(d, \eta)$ nie zawiera izolowanych wierzchołków.

Przypuśćmy teraz, że $a < 1/g(x_0(\eta))$. W tym przypadku, oczekiwana liczba wierzchołków izolowanych dąży do nieskończoności. Korzystając z Twierdzenia 5 można udowodnić, że $\mathbb{V}\mathrm{ar}\, Y_{izol}^{(d,\eta,n)} = o((\mathbb{E}Y_{izol}^{(d,\eta,n)})^2)$, co na mocy nierówności Czebyszewa oznacza, że prawie na pewno graf proteuszowy $\mathcal{P}_n(d, \eta)$ zawiera wierzchołek izolowany.

Pokażemy teraz, że dla dostatecznie małego $\delta > 0$ oraz $d(n) > (1/g(x_0(\eta)) - \delta) \log n$, graf proteuszowy $\mathcal{P}_n(d, \eta)$ zawiera dużą składową oraz (być może) pewną liczbę wierzchołków izolowanych. Niech c będzie małą, dodatnią stałą, która zostanie określona później. Rozważmy podgraf H grafu $\mathcal{P}_n(d, \eta)$ indukowany poprzez następujący zbiór wierzchołków

$$V(H) = \{i \in [n] : cn < i \leq n\} .$$

Oznaczmy przez $m = (1 - c)n$ rozmiar tego zbioru.

Na mocy Twierdzenia 5, możemy oszacować prawdopodobieństwo, że dowolny wierzchołek $k = \lceil xn \rceil \in V(H)$ nie posiada sąsiadów w zbiorze $V(H) \setminus \{v_1, \ldots, v_l\}$, dla dowolnej stałej l oraz dowolnych wierzchołków $v_1, \ldots, v_l \in V(H)$, przez (patrz również Twierdzenie 11)

$$\begin{aligned} & (1 + o(1))\Big[1 - \big(1 + O(\log^{-1/2} n)\big) \sum_{i \in V(H)\setminus\{v_1,\ldots,v_l\}, i<k} \frac{(1-\eta)}{n}\Big(\frac{k}{i}\Big)^{\eta}\Big]^d \\ & \qquad \times \prod_{i \in V(H)\setminus\{v_1,\ldots,v_l\}, i>k} \Big[1 - \big(1 + o(1)\big)(1-\eta)\frac{d}{n}\Big(\frac{i}{k}\Big)^{\eta}\Big] \\ = \; & \big(1 + O(\log^{-1/2} n)\big)^d (1 - x + c^{1-\eta}x^{\eta})^d \\ & \qquad \times \exp\Big[- \big(1 + o(1)\big)\frac{1-\eta}{1+\eta} d\big(x^{-\eta} - x\big)\Big] . \end{aligned}$$

Dobierając odpowiednio małą stałą c możemy sprawić by prawdopodobieństwo to było mniejsze niż

$$n^{o(1)} \exp\Big[- 0,75\, d\, g(x_0(\eta))\Big] .$$

Niech k_0 będzie stałą, dla której zachodzi następująca nierówność

$$0,2 \cdot (1-\eta)k_0/g(x_0(\eta)) > 2 \; .$$

Pokażemy najpierw, że w grafie H nie ma składowych o rozmiarze k, gdzie $2 \leq k < k_0$. Istotnie, zauważmy, że prawdopodobieństwo istnienia krawędzi w drzewie rozpinającym składową o rozmiarze $k = O(1) = o(d)$ można oszacować z góry przez $d/(cn)$ (patrz uwaga po Twierdzeniu 6). Zatem prawdopodobieństwo, że istnieje składowa o rozmiarze k, $2 \leq k < k_0$ można, podobnie jak w dowodzie Twierdzenia 15, oszacować z góry przez

$$\begin{aligned}
&\sum_{k=2}^{k_0-1} \binom{m}{k} k^{k-2} \Big(n^{o(1)} \exp\Big[-0,75\, d\, g(x_0(\eta)) \Big] \Big)^k \Big(\frac{d}{cn} \Big)^{k-1} \\
&\quad \leq \sum_{k=2}^{k_0-1} \Big(\frac{en}{k} \Big)^k k^{k-2} n^{o(1)} \exp\Big[-0,75k\, d\, g(x_0(\eta)) \Big] \Big(\frac{d}{cn} \Big)^{k-1} \\
&\quad \leq \sum_{k=2}^{k_0-1} n^{1+o(1)} k^{-2} e^k \Big(\frac{\log n}{cg(x_0(\eta))} \Big)^{k-1} \\
&\qquad \times \exp\Big[-0,75k\, (1/g(x_0(\eta)) - \delta) g(x_0(\eta)) \log n \Big] \\
&\quad \leq n^{1+o(1)} \Big(\frac{e \log n}{cg(x_0(\eta))} \Big)^{k_0} \sum_{k=2}^{k_0-1} k^{-2} \exp\Big[-0,7k\, \log n \Big] \\
&\quad \leq o(n^{-0,3}) \sum_{k=2}^{k_0-1} k^{-2} = o(n^{-0,3}) \; ,
\end{aligned}$$

dobierając dostatecznie mały parametr δ.

Pokażemy teraz, że w grafie H nie ma również składowych o rozmiarze k, $k_0 \leq k \leq 2n/3$. Istotnie, identycznie jak w dowodzie Twierdzenia 15 możemy oszacować z góry prawdopodobieństwo, że żaden wierzchołek w drzewie rozpinającym spójną składową nie jest połączony z wierzchołkami na zewnątrz drzewa przez

$$(1-(1-\eta)d/n)^{(k-2k/\sqrt{d})(m-k)} = (1-(1-\eta)d/n)^{(1+o(1))k(m-k)} \; .$$

Zatem prawdopodobieństwo, że istnieje składowa o rozmiarze k, $k_0 \leq k \leq 2n/3$ można oszacować z góry przez

$$\begin{aligned}
&\sum_{k=k_0}^{2n/3} \binom{m}{k} k^{k-2}(1-(1-\eta)d/n)^{(1+o(1))k(m-k)} \Big(\frac{d}{cn}\Big)^{k-1} \\
&\quad \leq \sum_{k=k_0}^{2n/3} \Big(\frac{en}{k}\Big)^k k^{k-2} \exp\Big(-(1+o(1))(1-\eta)dk(m-k)/n\Big)\Big(\frac{d}{cn}\Big)^{k-1} \\
&\quad \leq \sum_{k=k_0}^{2n/3} nk^{-2}e^k\Big(\frac{d}{c}\Big)^{k-1} \exp\Big(-0,3\cdot(1-\eta)kd\Big) \\
&\quad \leq \sum_{k=k_0}^{2n/3} ne^k\Big(\frac{\log n}{cg(x_0(\eta))}\Big)^k \exp\Big(-0,3\cdot(1-\eta)k(1/g(x_0(\eta))-\delta)\log n\Big) \\
&\quad \leq n\sum_{k=k_0}^{2n/3} \Big(\frac{e\log n}{cg(x_0(\eta))} n^{-0,2\cdot(1-\eta)/g(x_0(\eta))}\Big)^k \\
&\quad \leq n\Big(\frac{e\log n}{cg(x_0(\eta))} n^{-0,2\cdot(1-\eta)/g(x_0(\eta))}\Big)^{k_0}(1+o(1)) = o(n^{-0,9}) \,,
\end{aligned}$$

dobierając dostatecznie małe parametry c oraz δ.

Oznacza to, że prawie na pewno podgraf H składa się wyłącznie z dużej składowej oraz (być może) pewnej liczby wierzchołków izolowanych. Co więcej, liczba wierzchołków izolowanych w H jest mniejsza niż $n^{0.5}$. Istotnie, korzystając z wcześniejszych rozważań, liczbę wierzchołków izolowanych w podgrafie H można oszacować przez

$$n^{1+o(1)} \exp\Big[-0,75\, d\, g(x_0(\eta))\Big] \,,$$

co przy dostatecznie małym parametrze δ jest mniejsze niż $n^{0.3}$. Ostatecznie więc, na mocy nierówności Markowa, prawdopodobieńswo, że w podgrafie H jest więcej niż $n^{0.5}$ wierczhołków izolowanych dąży do zera.

Aby zakończyć dowód wystarczy pokazać, że dowolny wierzchołek $i < cn$, w grafie proteuszowym $\mathcal{P}_n(d,\eta)$, jest połączony z dużą składową w H. Istotnie, oznaczając przez I zbiór wierzchołków izolowanych w H, prawdopodobieństwo, że dowolny wierzchołek $i = \lceil xn \rceil < cn$ nie jest połączony z dużą składową można oszacować przez

$$\begin{aligned}
&\prod_{j\in V(H)\setminus I} \Big[1 - \big(1+o(1)\big)\frac{(1-\eta)}{n}\Big(\frac{j}{i}\Big)^{\eta}\Big]^d \\
&\qquad = (1+o(1))\exp\Big[-\big(1+o(1)\big)\frac{1-\eta}{1+\eta}d\big(x^{-\eta}-x\big)\Big] \\
&\qquad \le (1+o(1))\exp\Big[-\big(1+o(1)\big)\frac{1-\eta}{1+\eta}(1/g(x_0(\eta))-\delta)\big(c^{-\eta}-1\big)\log n\Big] \\
&\qquad = o(n^{-1})\,,
\end{aligned}$$

dobierając dostatecznie małe parametry c oraz δ. □

3.2 Średnica największej składowej

W niniejszym podrozdziale zajmiemy się zagadnieniem istnienia w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ dużej składowej oraz badaniem jej średnicy. Istnienie dużej składowej w grafie posiadającym stały średni stopień oraz fakt, że dowolne dwa wierzchołki w tej składowej są połączone stosunkowo krótką ścieżką pozwalają traktować graf $\mathcal{P}_n(d,\eta)$ jako dobry model grafu internetowego (przypomnijmy również, że graf proteuszowy posiada identyczny, jak w przypadku grafu internetowego, potęgowy rozkład stopni, patrz Twierdzenie 8).

Główne wyniki tej części pracy to Twierdzenia 23 i 25 oraz wynikające z nich Twierdzenie 17. Z Twierdzenia 23 wynika, że istnieje stała $C \in \mathbb{R}_+$ taka, że średnica największej składowej w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ jest mniejsza niż $C\log n$. Na podstawie Twierdzenia 25 wnioskujemy również istnienie stałej $c = c(d,\eta) \in \mathbb{R}_+$ takiej, że średnica największej składowej jest większa niż $c\log n$. W twierdzeniach tych zakładamy, z pewnych względów technicznych, że $d \ge 13$ oraz $\eta \in [0,58; 0,92]$. Zgodnie z Twierdzeniem 8

o rozkładzie stopni w grafie proteuszowym, zakres powyższych parametrów jest wystarczający do modelowania rzeczywistej sieci internetowej.

Oszacowania wynikające z tych dwóch twierdzeń podsumowuje następujący wynik.

Twierdzenie 17. *Niech $d \geq 13$ będzie ustaloną liczbą naturalną i $\eta \in [0,58; 0,92]$. Istnieją stałe $C \in \mathbb{R}_+$ oraz $c = c(d,\eta) \in \mathbb{R}_+$, takie że średnica dużej składowej w grafie proteuszowym $\mathcal{P}_n(d,\eta)$ jest większa niż $c \log n$ oraz mniejsza niż $C \log n$.*

3.2.1 Oszacowanie górne

Niech $d \geq 13$ będzie ustaloną liczbą naturalną oraz $\eta \in [0,58; 0,92]$. Pokażemy, że średnica grafu proteuszowego $\mathcal{P}_n(d,\eta)$ wynosi $O(\log n)$. W tym celu rozważmy proces przeszukiwania wszerz grafu proteuszowego $\mathcal{P}_n(d,\eta)$ startujący z dowolnego wierzchołka v. Przypomnijmy, że proces ten inicjujemy wstawiając wierzchołek v do kolejki Q oraz oznaczając v jako "odwiedzony". W k-tym kroku procesu, pobieramy z kolejki Q wierzchołek $v_k = x_k n$. Przez m_k oznaczmy liczbę wszystkich "odwiedzonych" wierzchołków w tym momencie. Następnie oznaczamy v_k jako "przetworzony" (tym samym każdy wierzchołek "przetworzony" jest traktowany również jako "odwiedzony"), wstawiamy do Q wszystkich "nieodwiedzonych" sąsiadów wierzchołka v_k oznaczając ich jako wierzchołki "odwiedzone". Przez $X_k = X_k(x_k, m_k)$ oznaczmy liczbę wierzchołków dodanych do kolejki Q w k-tym kroku procesu. Zauważmy, że zmienna losowa X_k zależy również od położenia wierzchołka j, z którego "dotarliśmy" do aktualnego wierzchołka (od położenia ojca v_k); jeżeli $j < v_k$, to $X_k(x_k, m_k)$ będziemy oznaczać przez $X_k(x_k, m_k, -)$, a gdy $j > v_k$ przez $X_k(x_k, m_k, +)$. Proces kończymy w momencie, gdy kolejka Q będzie pusta; wszystkie wierzchołki znajdujące się w spójnej składowej zawierającej wierzchołek v zostaną "przetworzone".

W rozdziale tym skorzystamy z dobrze znanych osiągnięć teorii procesów gałązkowych Galtona–Watsona. Rozważmy standardowy proces gałązkowy [3], w którym każda cząstka może produkować inne cząstki tej samej postaci. Niech w chwili początkowej dana będzie jedna cząstka ("pokolenie zerowe"), która w wyniku "podziału" przechodzi z prawdopodobieństwem p_k, $k = 0, 1, 2, \ldots$, w k cząstek tego samego typu, $\sum_{k=0}^{\infty} p_k = 1$. Otrzymane cząstki stanowią "pierwsze pokolenie". Każda z cząstek tego pokolenia zachowuje się dokładnie tak samo, jak cząstka wyjściowa, niezależnie od dotychczasowej historii podziału cząstek i od przyszłych losów innych cząstek. W ten sposób dostajemy "drugie pokolenie" itd. Oznaczmy przez W_l liczbę cząstek w l–tym pokoleniu. Dla opisania ciągu $\{W_l\}_{l=0}^{\infty}$ wprowadzimy do rozważań niezależne od siebie ciągi niezależnych zmiennych losowych o jednakowym rozkładzie $\{Z_j^1\}_{j=1}^{\infty}, \{Z_j^2\}_{j=2}^{\infty}, \ldots$, gdzie zmienne losowe Z_j^i mają rozkład

$$\mathbb{P}(Z_j^i = k) = p_k, k = 0, 1, 2, \ldots$$

Wówczas ciąg $\{W_l\}_{l=0}^{\infty}$ można przedstawić w postaci

$$\begin{aligned} W_0 &= 1 \\ W_1 &= Z_1^1 \\ W_2 &= Z_1^2 + \cdots + Z_{W_1}^2 \\ &\vdots \\ W_l &= Z_1^l + \cdots + Z_{W_{l-1}}^l \, . \end{aligned}$$

Przez degenerację będziemy rozumieć zdarzenie polegające na tym, że poczynając od pewnego l_0 wszystkie $W_l, l > l_0$ są równe 0 (jeżeli $W_{l_0} = 0$, to oczywiście $W_{l_0+1} = W_{l_0+2} = \cdots = 0$). Teoria procesów gałązkowych mówi, że prawdopodobieństwo degeneracji jest równe najmniejszemu pierwiastkowi równania $z = f(z)$, gdzie f jest funkcją tworzącą zmiennej losowej Z_1^1 (patrz na przykład [4, 5]).

Twierdzenie 18. *Rozważmy gałązkowy proces Galtona–Watsona. Niech Z_j^i oznacza liczbę potomków j-tego osobnika z i-tego pokolenia. Załóżmy, że zmienne losowe $\{Z_j^i : i, j = 1, 2, \dots\}$ są niezależne i mają ten sam rozkład prawdopodobieństwa*

$$\mathbb{P}(Z_j^i = k) = p_k, \text{ dla } k = 0, 1, 2, \dots ,$$

gdzie $p_0 > 0$. Przez $f(z)$ oznaczmy funkcję tworzącą zmiennej losowej Z_1^1 zdefiniowaną wzorem

$$f(z) = \sum_{k=0}^{\infty} p_k \cdot z^k .$$

Wówczas, gdy $\mathbb{E}X \leq 1$, to prawdopodobieństwo degeneracji wynosi 1. W przypadku, gdy $\mathbb{E}X > 1$ prawdopodobieństwo degeneracji jest równe z_0, gdzie $z_0 < 1$ jest najmniejszym nieujemnym pierwiastkiem równania $z = f(z)$.

Zauważmy, że proces przeszukiwania wszerz przypomina proces gałązkowy, z tą różnicą, że rozkład zmiennej losowej $X_k(x_k, m_k)$, będącej liczbą nowych wierzchołków, które zostają dołączone do składowej w k–tym kroku procesu, zależy od dotychczasowego jego przebiegu. Pokażemy jednak, że gdy $m_k < n^{2/3}$ to

$$\mathbb{P}(X_k(x_k, m_k) \leq 1) \leq 1/3 .$$

Oznacza to, że zmienną losową $X_k(x_k, m_k)$ można oszacować z dołu przez niezależną zmienną losową X o rozkładzie

$$\begin{aligned} \mathbb{P}(X = 0) &= 1/3 , \\ \mathbb{P}(X = 2) &= 2/3 . \end{aligned}$$

Twierdzenie 19. *Dla każdego $k \in \mathbb{N}$, $x_k \in (0, 1)$ oraz $m_k < n^{2/3}$*

$$\mathbb{P}(X_k(x_k, m_k) \leq 1) \leq 1/3 . \tag{6}$$

Dowód. Dla dowolnych $a, b \in [n]$, $a < b$ waga wierzchołka a jest większa od wagi wierzchołka b. Oznacza to, że suma wag dowolnych m_k wierzchołków jest nie większa niż suma wag $n^{2/3}$ początkowych wierzchołków. Zatem prawdopodobieństwo, że wierzchołek w czasie trwania procesu proteuszowego wybrał wierzchołek spośród wybranych uprzednio m_k wierzchołków jest nie większe niż

$$\frac{\sum_{i=1}^{n^{2/3}} i^{-\eta}}{\sum_{i=1}^{n} i^{-\eta}} = (1+o(1))\frac{\int_0^{n^{-1/3}} t^{-\eta}dt}{\int_0^1 t^{-\eta}dt} = (1+o(1))n^{-1/3(1-\eta)} .$$

Możemy zatem założyć, że powyższa sytuacja nie występuje.

Oznaczmy przez

$$\begin{aligned} w(\eta,x) &= \sum_{l=\lceil x\cdot n\rceil+1}^{n} p^{(1,\eta,n)}(l,\lceil xn\rceil) \\ &= \big(1+o(1)\big)\frac{1-\eta}{n}\sum_{l=\lceil x\cdot n\rceil+1}^{n}\Big(\frac{l}{\lceil xn\rceil}\Big)^{\eta} \\ &= \big(1+o(1)\big)(1-\eta)\int_x^1\Big(\frac{t}{x}\Big)^{\eta}dt \\ &= \big(1+o(1)\big)\frac{1-\eta}{1+\eta}x^{-\eta}(1-x^{1+\eta}) \\ &= \big(1+o(1)\big)\frac{1-\eta}{1+\eta}(x^{-\eta}-x) . \end{aligned}$$

Policzmy najpierw prawdopodobieństwo, że $X_k(x_k, m_k, -) = 0$. Oznacza to, że wierzchołek $i = \lceil x_k n\rceil$ w momencie kiedy był ostatni raz "odświeżany" wybrał ze zbioru wierzchołków $\{1, 2, \ldots, i-1\}$ tylko ustalony wierzchołek j oraz żaden z wierzchołków ze zbioru $\{i+1, i+2, \ldots, n\}$ nie wybrał i.

$$\begin{aligned} \mathbb{P}(X_k(x_k,m_k,-)=0) &= (1+o(1))(1-x)^{d-1}\prod_{l=i+1}^{n}(1-p^{(d,\eta,n)}(l,i)) \\ &= (1+o(1))(1-x)^{d-1}\prod_{l=i+1}^{n}\exp(-d\cdot p^{(1,\eta,n)}(l,i)) \\ &= (1+o(1))(1-x)^{d-1}\exp(-d\sum_{l=i+1}^{n}p^{(1,\eta,n)}(l,i)) \\ &= (1+o(1))(1-x)^{d-1}\exp(-dw(\eta,x)) \end{aligned}$$

Podobnie $X_k(x_k, m_k, +) = 0$ wtedy i tylko wtedy, gdy wierzchołek $i = \lceil x_k n \rceil$ w momencie kiedy był ostatni raz "odświeżany" nie wybrał żadnego wierzchołka ze zbioru $\{1, 2, \ldots, i-1\}$ oraz żaden z wierzchołków ze zbioru $\{i+1, i+2, \ldots, n\}$, za wyjątkiem ustalonego wierzchołka j, nie wybrał i. Mamy zatem

$$\begin{aligned}
\mathbb{P}(X_k(x_k, m_k, +) = 0) &= (1+o(1))(1-x)^d \prod_{\substack{l=i+1 \\ l\neq j}}^{n} (1 - p^{(d,\eta,n)}(l,i)) \\
&= (1+o(1))(1-x)^d \prod_{\substack{l=i+1 \\ l\neq j}}^{n} \exp(-d \cdot p^{(1,\eta,n)}(j,i)) \\
&= (1+o(1))(1-x)^d \exp(-d \sum_{\substack{l=i+1 \\ l\neq j}}^{n} p^{(1,\eta,n)}(j,i)) \\
&= (1+o(1))(1-x)^d \exp(-dw(\eta,x)) \,.
\end{aligned}$$

Stąd

$$\mathbb{P}(X_k(x_k, m_k) = 0) \leq (1+o(1))(1-x)^{d-1} \exp(-dw(\eta,x)) \,.$$

Analogicznie można pokazać, że

$$\begin{aligned}
\mathbb{P}(X_k(x_k, m_k) = 1) &\leq \mathbb{P}(X_k(x_k, m_k, -) = 1) \\
&= (1+o(1))dx(1-x)^{d-2} \prod_{l=i+1}^{n} (1 - p^{(d,\eta,n)}(l,i)) \\
&+ (1+o(1))(1-x)^{d-1} \sum_{s=i+1}^{n} p^{(d,\eta,n)}(s,i) \prod_{\substack{l=i+1 \\ l\neq s}}^{n} (1 - p^{(d,\eta,n)}(l,i)) \\
&= (1+o(1))dx(1-x)^{d-2} \exp(-dw(\eta,x)) \\
&+ (1+o(1))(1-x)^{d-1} d \exp(-dw(\eta,x)) \sum_{s=i+1}^{n} p^{(1,\eta,n)}(s,i) \\
&= (1+o(1))dx(1-x)^{d-2} \exp(-dw(\eta,x)) \\
&+ (1+o(1))(1-x)^{d-1} dw(\eta,x) \exp(-dw(\eta,x)) \,,
\end{aligned}$$

ostatecznie więc

$$\begin{aligned}
\mathbb{P}(X_k(x_k, m_k) \leq 1) &= \mathbb{P}(X_k(x_k, m_k) = 1) + \mathbb{P}(X_k(x_k, m_k) = 0) \\
&\leq (1 + o(1))dx(1-x)^{d-2} \exp(-dw(\eta, x)) \\
&+ (1 + o(1))(1-x)^{d-1} dw(\eta, x) \exp(-dw(\eta, x)) \\
&+ (1 + o(1))(1-x)^{d-1} \exp(-dw(\eta, x)) \\
&= (1 + o(1))\big[dx + dw(\eta, x)(1-x) + (1-x)\big] \\
&\qquad \times (1-x)^{d-2} \exp(-dw(\eta, x)) \, .
\end{aligned}$$

Pokażemy teraz, że dla $d \geq 13$, $\eta \in [0,58; 0,92]$, $m_k < n^{2/3}$, $x_k \in (0,1)$ prawdziwa jest nierówność (6). Niestety, ze względów technicznych, musimy podzielić interesujący nas zakres parametrów na kilka podzakresów i szacować w zależności od występującego przypadku.

Przypadek 1. Załóżmy, że $x_k \in [0, x_{\max}], \eta \in [\eta_{\min}, \eta_{\max}]$, wtedy

$$\mathbb{P}(X_k(x_k, m_k) \leq 1) \leq (1 + o(1))(dx + dw(\eta, x) + 1) \exp(-dw(\eta, x)) \, .$$

Zauważmy, że funkcja $w(\eta, x)$ (dla ustalonego $\eta \in (0,1)$) jest funkcją malejącą. Oznaczmy przez

$$\hat{w} = \min_{\substack{x \in [0, x_{\max}] \\ \eta \in [\eta_{\min}, \eta_{\max}]}} w(\eta, x) = \big(1 + o(1)\big) \frac{1 - \eta_{\max}}{1 + \eta_{\max}} (x_{\max}^{-\eta_{\min}} - x_{\max}) \, .$$

Zauważmy również, że funkcja $f(x) = xe^{-x}$ jest funkcją malejącą na przedziale $[1, \infty)$, a zatem zachodzi poniższa nierówność (o ile $d\hat{w} \geq 1$)

$$\mathbb{P}(X_k(x_k, m_k) \leq 1) \leq (1 + o(1))(dx_{\max} + d\hat{w} + 1) \exp(-d\hat{w}) \, .$$

W szczególności dla następujących trzech podprzypadków

$$\begin{cases}
x_{\max} = 0,15, \eta_{\min} = 0,58, \eta_{\max} = 0,8 \\
x_{\max} = 0,15, \eta_{\min} = 0,8, \eta_{\max} = 0,9 \\
x_{\max} = 0,15, \eta_{\min} = 0,9, \eta_{\max} = 0,92
\end{cases}$$

zachodzi nierówność (6).

Przypadek 2. Załóżmy, że $x_k \in [x_{\min}, x_{\max}], \eta \in [\eta_{\min}, \eta_{\max}]$. Analogicznie jak poprzednio można pokazać, że gdy $d\hat{w} \geq 1$

$$\begin{aligned}\mathbb{P}(X_k(x_k, m_k) \leq 1) \quad &\leq \quad (1+o(1))[dx_{\max} + d\hat{w}(1-x_{\min}) + (1-x_{\min})] \\ &\quad \times (1-x_{\min})^{d-2} \exp(-d\hat{w}) \ .\end{aligned}$$

W szczególności dla następujących dwóch podprzypadków

$$\begin{cases} x_{\min} = 0,15, x_{\max} = 0,35, \eta_{\min} = 0,58, \eta_{\max} = 0,885 \\ x_{\min} = 0,15, x_{\max} = 0,35, \eta_{\min} = 0,885, \eta_{\max} = 0,92 \end{cases}$$

zachodzi nierówność (6).

Przypadek 3. Załóżmy, że $x_k \in [x_{\min}, 1]$, wtedy

$$\begin{aligned}\mathbb{P}(X_k(x_k, m_k) \leq 1) \quad &\leq \quad (1+o(1))[dx + dw(\eta, x)(1-x) + (1-x)](1-x)^{d-2} \\ &\leq \quad (1+o(1))[dx + d\frac{1}{x}(1-x) + d(1-x)](1-x)^{d-2} \\ &= \quad (1+o(1))\frac{d}{x}(1-x)^{d-2} \leq (1+o(1))\frac{d}{x_{\min}}(1-x_{\min})^{d-2} \ .\end{aligned}$$

W szczególności dla $x_{\min} = 0,35$ zachodzi nierówność (6).

Ostatecznie więc dla $d \geq 13$, $\eta \in [0,58; 0,92]$, $m_k < n^{2/3}$, $x_k \in (0,1)$

$$\mathbb{P}(X_k(x_k, m_k) \leq 1) \leq 1/3 \ .$$

□

Zauważmy, że wartość oczekiwana zmiennej losowej X wynosi $\mathbb{E}X = 4/3$. Skoro zmienną losową $X_k(x_k, m_k)$ można oszacować z dołu przez X, należy się spodziewać, iż z pozytywnym prawdopodobieństwem przeszukując graf proteuszowy wszerz napotkamy dużą składową. Poniższe twierdzenie precyzuje tę obserwację.

Twierdzenie 20. *Niech $d \geq 13$ będzie ustaloną liczbą naturalną oraz $\eta \in [0,58; 0,92]$. Prawdopodobieństwo, że proces przeszukiwania grafu proteuszowego $\mathcal{P}_n(d,\eta)$ wszerz, startujący z dowolnego wierzchołka $v \in [n]$, wymrze wcześniej niż po $n^{2/3}$ krokach, jest mniejsze od $1/2$.*

Dowód. Z Twierdzenia 18 wynika, że prawdopodobieństwo degeneracji procesu gałązkowego, w którym zmienne losowe Z_j^i mają rozkład identyczny jak zmienna losowa X ($p_0 = 1/3$, $p_2 = 2/3$) jest równe najmniejszemu pierwiastkowi równania

$$z = f_X(z) = \frac{1}{3} + \frac{2}{3}z^2 \,,$$

który wynosi $1/2$.

Wykorzystując Twierdzenie 19, które mówi, że zmienna losowa X jest oszacowaniem dolnym zmiennej losowej $X_k(x_k, m_k)$ dla dowolnego $x_k \in (0,1)$, $k \in \mathbb{N}$ oraz $m_k < n^{2/3}$ otrzymujemy tezę twierdzenia. □

W poprzednim twierdzeniu pokazaliśmy, że z prawdopodobieństwem większym niż $1/2$, dowolnie wybrany wierzchołek $v \in [n]$ znajduje się w składowej o rozmiarze większym niż $n^{2/3}$. Rozważmy proces przeszukiwania grafu proteuszowego wszerz startujący z wierzchołka v. Przez $\hat{O}$ oznaczmy liczbę wierzchołków "odwiedzonych", przez $\hat{P}$ zaś liczbę wierzchołków "przetworzonych". Pokażemy teraz, że proces przeszukiwania grafu proteuszowego wszerz zakończy się szybko (wyznaczając składową o rozmiarze nie większym niż $150 \log n$), albo będzie "zagarniał" w szybkim tempie kolejne wierzchołki ($\hat{O} > \frac{7}{6}\hat{P}$).

Twierdzenie 21. *Niech $d \geq 13$, $\eta \in [0,58; 0,92]$. Rozważmy proces przeszukiwania grafu proteuszowego $\mathcal{P}_n(d,\eta)$ wszerz, startujący z dowolnego wierzchołka $v \in [n]$. Przez $\hat{P}$ oznaczmy liczbę wierzchołków "przetworzonych", natomiast przez $\hat{O}$ liczbę wierzchołków "odwiedzonych".*

$$\mathbb{P}\Big(\hat{O} \leq \frac{7}{6}\hat{P} \Big| 150 \log n \leq \hat{O} \leq n^{2/3}\Big) < o(n^{-2}) \,.$$

Dowód. Analogicznie jak w poprzednim dowodzie skorzystamy z faktu, że proces przeszukiwania grafu proteuszowego wszerz jest podobny do procesu gałązkowego, w którym zmienne losowe Z_i^j mają identyczny rozkład jak zmienna losowa X. W każdym kroku procesu gałązkowego wybieramy (po kolei) jedną cząstkę z aktualnego pokolenia (pobieramy wierzchołek z kolejki Q i oznaczamy go jako "przetworzony"), która w wyniki "podziału" tworzy, zgodnie z rozkładem zmiennej losowej X, nowe cząstki (wstawiamy do Q wszystkich "nieodwiedzonych" sąsiadów, oznaczając ich jako wierzchołki "odwiedzone"). Przez W_l oznaczmy liczbę cząstek w l–tym pokoleniu, przez U_l zaś liczbę cząstek do l–tego pokolenia włącznie

$$U_l = l + W_l \ .$$

Pokażemy teraz, że zachodzi następująca nierówność

$$\mathbb{P}\Big(U_l \leq \frac{7}{6}l \Big| 150 \log n \leq l \leq n^{2/3}\Big) < o(n^{-2}) \ ,$$

co na mocy Twierdzenia 19 zakończy dowód.

Dla dowolnego $u < 0$ oraz $t > 0$ zachodzi następująca nierówność

$$\mathbb{P}(X \leq \mathbb{E}X - t) \leq e^{-u(\mathbb{E}X - t)} \cdot \mathbb{E}e^{uX} \ .$$

Istotnie, wykorzystując nierówność Markowa otrzymamy

$$\begin{aligned} \mathbb{P}(X \leq \mathbb{E}X - t) &= \mathbb{P}(X - \mathbb{E}X \leq -t) = \mathbb{P}(u(X - \mathbb{E}X) \geq -ut) \\ &= \mathbb{P}(e^{u(X-\mathbb{E}X)} \geq e^{-ut}) \leq \frac{\mathbb{E}e^{u(X-\mathbb{E}X)}}{e^{-ut}} \\ &= e^{-u(\mathbb{E}X - t)}\mathbb{E}e^{uX} \ . \end{aligned}$$

Oznacza to, że dla dowolnego $u < 0$, $t > 0$ oraz $n \in \mathbb{N}$

$$\begin{aligned} \mathbb{P}\Big(\sum_{i=1}^{n} X \leq n\mathbb{E}X - t\Big) &\leq e^{-u(n\mathbb{E}X - t)}\Big(\mathbb{E}e^{uX}\Big)^n \\ \mathbb{P}\Big(\sum_{i=1}^{n} X \leq \frac{4}{3}n - t\Big) &\leq e^{-u(\frac{4}{3}n - t)}\Big(\frac{1}{3} + \frac{2}{3}e^{2u}\Big)^n \ . \end{aligned}$$

W szczególności dla $u = \frac{1}{2}\log\frac{7}{10}$ $(e^{2u} = \frac{7}{10})$

$$\mathbb{P}\Big(\sum_{i=1}^{n} X \leq \frac{4}{3}n - t\Big) \leq \Big(\frac{10}{7}\Big)^{\frac{1}{2}(\frac{4}{3}n-t)}\Big(\frac{4}{5}\Big)^{n} .$$

Zauważmy, że dla dowolnego $l \in \mathbb{N}$

$$1 + \sum_{j=1}^{l} X = U_l ,$$

co oznacza, że

$$\begin{aligned}
\mathbb{P}\Big(U_l \leq \frac{7}{6}l \Big| 150\log n \leq l \leq n^{2/3}\Big) & \\
&\leq \mathbb{P}\Big(\sum_{j=1}^{l} X \leq \frac{7}{6}l \Big| 150\log n \leq l \leq n^{2/3}\Big) \\
&\leq \Big(\frac{10}{7}\Big)^{\frac{1}{2}\cdot\frac{7}{6}k}\Big(\frac{4}{5}\Big)^{k} < \exp(-0,015k) \\
&\leq \exp(-0,015 \cdot 150\log n) \\
&= n^{-2,25} = o(n^{-2}) .
\end{aligned}$$

□

Z powyższego twierdzenia wynika następujący wniosek.

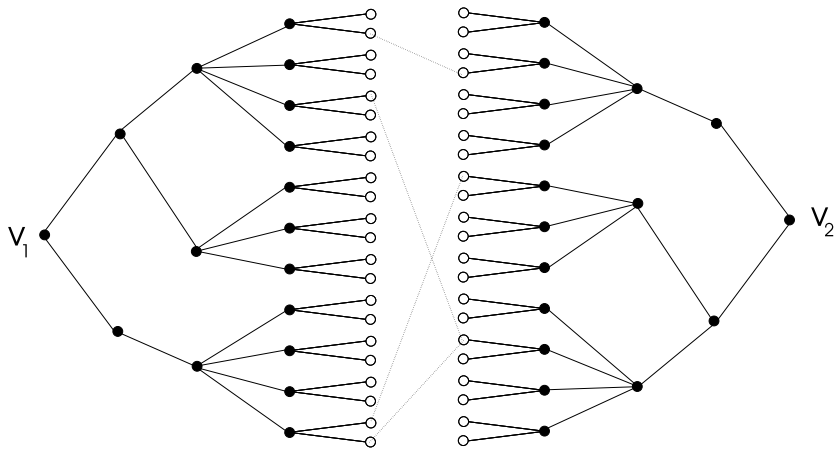
Wniosek 22. *Niech $d \geq 13$, $\eta \in [0,58; 0,92]$. Rozważmy proces przeszukiwania grafu proteuszowego $\mathcal{P}_n(d,\eta)$ wszerz, startujący z dowolnego wierzchołka $v \in [n]$. Przez Z_l oznaczmy liczbę wierzchołków w odległości co najwyżej l od wierzchołka v. Dla dowolnego $l \in \mathbb{N}$ zachodzi następująca nierówność*

$$\mathbb{P}\Big(Z_{l+1} \leq \frac{7}{6}Z_l \Big| 150\log n \leq Z_l = k \leq n^{2/3}\Big) < o(n^{-2}) .$$

Z Twierdzenia 21 wynika, że graf proteuszowy nie ma składowych o rozmiarze k, $150\log n \leq k \leq n^{2/3}$. Pokażemy teraz, że oprócz małych składowych (o rozmiarze nie większym niż $150\log n$), graf posiada jedną dużą składową o średnicy $O(\log n)$.

Twierdzenie 23. *Niech $d \geq 13$, $\eta \in [0,58; 0,92]$. Graf proteuszowy $\mathcal{P}_n(d,\eta)$ posiada dużą składową o rozmiarze co najmniej $n^{2/3}$; pozostałe składowe mają rozmiary co najwyżej $O(\log n)$. Co więcej, średnica dużej składowej wynosi co najwyżej $O(\log n)$.*

Dowód. Istnienie składowej o rozmiarze co najmniej $n^{2/3}$ wynika z Twierdzenia 20. Pokażemy teraz, że istnieje tylko jedna taka składowa. Rozważmy w tym celu dwa procesy przeszukujące graf wszerz, startujące z wierzchołków v_1 oraz v_2, które nie wymrą wcześniej niż w $n^{2/3}$ krokach. Zatrzymujemy te procesy, gdy liczba "odwiedzonych" wierzchołków będzie równa $n^{2/3}$. Przez $S(v_1)$ oraz $S(v_2)$ oznaczmy zbiory składające się z wierzchołków "odwiedzonych" w procesie startującym z v_1 oraz odpowiednio v_2. Przypuśćmy, że zbiory te są rozłączne. Powyższą sytuację przedstawia następujący rysunek.



$S(v_1)$ $S(v_2)$

Jak nietrudno zauważyć wszyscy sąsiedzi dowolnego "przetworzonego" już wierzchołka $v \in S(v_i)$ (wierzchołki czarne) znajdują się również w zbiorze $S(v_i)$. Natomiast wierzchołki, które nie zostały jeszcze "przetworzone" (wierzchołki białe) mogą połączyć zbiory $S(v_1)$ i $S(v_2)$. Z Twierdzenia 21 wynika, że liczba "odwiedzonych" wierzchołków, które nie zostały jeszcze "przetworzone" wynosi co najmniej $\frac{1}{7}n^{2/3}$. Można zatem wskazać dwa podzbiory $\hat{S}(v_1) \subset S(v_1)$ oraz $\hat{S}(v_2) \subset S(v_2)$, zawierające po $\frac{1}{14}n^{2/3}$ elementów każdy, posiadające tę własność, że wszystkie wierzchołki ze zbioru $\hat{S}(v_1)$ leżą "nad" wierzchołkami $\hat{S}(v_2)$ (dla każdego $i \in \hat{S}(v_1)$, $j \in \hat{S}(v_2)$ $i > j$). Waga dowolnego wierzchołka jest większa od wagi wierzchołka n, ta zaś wynosi $\frac{1-\eta}{n}$. Zatem prawdopodobieństwo, że wierzchołki z $\hat{S}(v_1)$ nie są połączone z wierzchołkami z $\hat{S}(v_2)$ jest mniejsze niż

$$\Big(1 - \frac{n^{2/3}}{14} \cdot \frac{1-\eta}{n}\Big)^{\frac{n^{2/3}}{14}} = (1+o(1)) \exp\Big[-\frac{n^{1/3}(1-\eta)}{196}\Big] = o(n^{-2}) \,,$$

co oznacza, iż istnieje tylko jedna duża składowa (o rozmiarze większym niż $n^{2/3}$). Rozmiar pozostałych składowych wynika z Twierdzenia 21.

Pokażemy teraz, że średnica największej składowej wynosi co najwyżej $O(\log n)$. Weźmy w tym celu dwa dowolne wierzchołki należące do największej składowej v_1 oraz v_2. Z Wniosku 22 wynika, że jeżeli liczba "odwiedzonych" wierzchołków wynosi co najmniej $150 \log n$, to wystarczy $\log_{\frac{7}{6}} n^{2/3} < 5 \log n$ pokoleń, aby dotrzeć do $n^{2/3}$ wierzchołków. Stąd średnica grafu indukowanego przez zbiór $S(v_i)$, $i = 1, 2$ wynosi co najwyżej $155 \log n$. Ostatecznie więc średnica największej składowej wynosi co najwyżej $310 \log n$. □

3.2.2 Oszacowanie dolne

Niech $d \geq 13$ będzie ustaloną liczbą naturalną, $\eta \in [0,58; 0,92]$. Pokażemy, że średnica grafu proteuszowego $\mathcal{P}_n(d,\eta)$ wynosi $\Omega(\log n)$. Zanim jednak zaprezentujemy główne twierdzenie tego podrozdziału, przedstawimy, potrzebne w dowodzie twierdzenia, oszacowania dolne na prawdopodobieństwo $p^{(d,\eta,n)}(j,i)$ istnienia krawędzi pomiędzy wierzchołkami i, j oraz na prawdopodobieństwo, że wierzchołek k jest izolowany, które oznaczymy przez $p_{izol}^{(d,\eta,n)}(k)$.

Lemat 24. *Niech $d \in \mathbb{N}$, $\eta \in (0,1)$. Dla dowolnego $i,j \in [n]$, $j > i$ oraz $k \in [\frac{1}{2}n, \frac{3}{4}n] \cap \mathbb{N}$ zachodzą poniższe nierówności*

$$\begin{aligned} p^{(d,\eta,n)}(j,i) &\geq (1+o(1))\frac{1-\eta}{n}d \\ p_{izol}^{(d,\eta,n)}(k) &\geq \Big(\frac{1}{4\sqrt{e}}\Big)^d. \end{aligned}$$

Dowód. Pierwsza nierówność wynika z faktu, że

$$p^{(d,\eta,n)}(j,i) = (1+o(1))\frac{1-\eta}{n}d\Big(\frac{j}{i}\Big)^{\eta} \geq (1+o(1))\frac{1-\eta}{n}d\,.$$

Niech zatem $k = \lceil xn \rceil$, $x \in [\frac{1}{2}, \frac{3}{4}]$. Na mocy Twierdzenia 11

$$p_{izol}^{(d,\eta,n)}(k) = (1+o(1))(1-x)^d \exp\Big[-(1+o(1))\frac{1-\eta}{1+\eta}d(x^{-\eta}-x)\Big]\,.$$

Oznaczmy przez

$$w(\eta,x) = -\frac{1-\eta}{1+\eta}(x^{-\eta}-x)\,.$$

Zauważmy, że dla ustalonego $\eta_0 \in (0,1)$ funkcja $w(\eta_0,x)$ jest funkcją rosnącą. Istotnie

$$w(\eta_0,x)'_x = \frac{1-\eta_0}{1+\eta_0}(\eta_0 x^{-\eta_0-1}+1) > 0\,,$$

a więc

$$p_{izol}^{(d,\eta,n)}(k) \geq \Big(\frac{1}{4}\Big)^d \exp\Big[-\frac{1-\eta}{1+\eta}d\Big(\Big(\frac{1}{2}\Big)^{-\eta}-\frac{1}{2}\Big)\Big]\,.$$

Można również pokazać, że dla ustalonego $x_0 \in (0,1)$

$$w(\eta,x_0)'_\eta = \frac{x_0^{-\eta}}{(1+\eta)^2}(2-2x_0^{1+\eta}+\log x_0-\eta^2\log x_0)\,.$$

W szczególności

$$\begin{aligned} w\Big(\eta,\frac{1}{2}\Big)'_{\eta} &= \frac{(\frac{1}{2})^{-\eta}}{(1+\eta)^2}\Big[2-2\Big(\frac{1}{2}\Big)^{1+\eta}+\log\frac{1}{2}-\eta^2\log\frac{1}{2}\Big] \\ &\geq \frac{2^{\eta}}{(1+\eta)^2}\Big(2-2\frac{1}{2}+\log\frac{1}{2}\Big)>0\,, \end{aligned}$$

co oznacza, że funkcja $w(\eta,\frac{1}{2})$ jest funkcją rosnącą. Ostatecznie więc

$$p_{izol}^{(d,\eta,n)}(k)\geq\Big(\frac{1}{4}\Big)^d\exp\Big(-d\frac{1}{2}\Big)=\Big(\frac{1}{4\sqrt{e}}\Big)^d\,.$$

□

Aby pokazać, że średnica największej składowej, w grafie proteuszowym $\mathcal{P}_n(d,\eta)$, jest rzędu co najmniej $\log n$ wykażemy, że prawdopodobieństwo istnienia izolowanej ścieżki o długości $O(\log n)$, której pierwszy wierzchołek połączony jest z dużą składową, dąży do 1.

Twierdzenie 25. *Niech $d\geq 13$ będzie ustaloną liczbą naturalną oraz $\eta\in[0,58;0,92]$. Oczekiwana liczba izolowanych ścieżek, w grafie proteuszowym $\mathcal{P}_n(d,\eta)$, o długości*

$$k=k(n)=\frac{\log n}{4d-2\log(1-\eta)}\,,$$

w których pierwszy wierzchołek, należący do zbioru $[1,\frac{1}{2}n)$, połączony jest z dużą składową oraz pozostałe wierzchołki należą do zbioru $[\frac{1}{2}n,\frac{3}{4}n]$ wynosi co najmniej $n^{1/2}$.

Co więcej, prawie na pewno istnieje jedna taka ścieżka.

Dowód. Niech $x_1\in[1,\frac{1}{2}n)$, dla dowolnego $2\leq i\leq k+1$, niech $x_i\in[\frac{1}{2}n,\frac{3}{4}n]$. Przez $A(x_1,x_2,\ldots,x_{k+1})$ oznaczmy zdarzenie polegające na tym, że istnieje izolowana ścieżka $(x_1,x_2,\ldots,x_{k+1})$ o długości k oraz wierzchołek x_1 połączony jest z dużą składową. Dowód rozpoczniemy od oszacowania z dołu prawdopodobieństwa zachodzenia zdarzenia $A(x_1,x_2,\ldots,x_{k+1})$.

Zauważmy, że prawdopodobieństwo, że wierzchołek nie ma sąsiadów, za wyjątkiem być może dwóch bądź jednego ustalonego wierzchołka, jest nie mniejsze niż prawdopodobieństwo bycia wierzchołkiem izolowanym.

W Twierdzeniu 20 pokazano, że dowolny wierzchołek znajduje się w dużej składowej z prawdopodobieństwem nie mniejszym niż 1/2 (proces przeszukiwania grafu wszerz, startujący z tego wierzchołka, nie wymrze przed "zagarnięciem" $n^{2/3}$ wierzchołków). Fakt istnienia w grafie proteuszowym izolowanej ścieżki oraz istnienia krawędzi $\{x_1, x_2\}$ wpływa na prawdopodobieństwo, że wierzchołek x_1 należy do dużej składowej, niemniej wpływ ten nie jest zbyt duży (zwróćmy uwagę, że długość ścieżki wynosi $O(\log n)$ oraz $x_2 > x_1$). Powtarzając rozumowanie przedstawione w dowodzie Twierdzenia 20, można pokazać, że

$$\begin{aligned}\mathbb{P}(x_1 \textit{ należy do dużej składowej} \mid x_1 \textit{ jest początkiem izolowanej ścieżki}) \\ = (1+o(1))\mathbb{P}(x_1 \textit{ należy do dużej składowej}) \geq 1/2(1+o(1))\,.\end{aligned}$$

Podobna własność zachodzi w przypadku postulowania istnienia dwóch ścieżek, również, gdy obie ścieżki mają początek w tym samym punkcie (ten fakt będzie istotny w dalszej części dowodu). Na podstawie powyższych rozważań oraz oszacowań z Lematu 24 otrzymujemy następujące oszacowanie

$$\begin{aligned}\mathbb{P}(A(x_1,x_2,\ldots,x_{k+1})) &= (1+o(1))\prod_{i=1}^{k}\mathbb{P}(x_i \textit{ jest połączony z } x_{i+1}) \\ &\quad \times\prod_{i=2}^{k+1}\mathbb{P}(x_i \textit{ nie ma sąsiadów poza sąsiadami w ścieżce}) \\ &\quad \times\mathbb{P}(x_1 \textit{ należy do dużej składowej}) \\ &\geq (1+o(1))\frac{1}{2}\Big(\frac{1-\eta}{n}d\Big)^k\Big[\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k\,. \qquad (7)\end{aligned}$$

Niech $Y(x_1,x_2,\ldots,x_{k+1})$, $x_1 \in [1,\frac{1}{2}n)$, $x_i \in [\frac{1}{2}n,\frac{3}{4}n]$, $2 \leq i \leq k+1$ będzie rodziną zmiennych losowych zdefiniowanych w następujący sposób

$$Y(x_1,x_2,\ldots,x_{k+1}) = \begin{cases}1 & \text{gdy zachodzi zdarzenie } A(x_1,x_2,\ldots,x_{k+1}) \\ 0 & \text{w przeciwnym razie .}\end{cases}$$

Wtedy liczba izolowanych ścieżek wynosi

$$Y = \sum_{1\le x_1<1/2n} \sum_{1/2n\le x_2,\ldots,x_{k+1}\le 3/4n} Y(x_1, x_2, \ldots, x_{k+1}) \ .$$

Oszacujmy najpierw wartość oczekiwaną zmiennej losowej Y.

$$\begin{aligned}
\mathbb{E}Y &= \sum_{1\le x_1<1/2n} \sum_{1/2n\le x_2,\ldots,x_{k+1}\le 3/4n} \mathbb{P}(A(x_1, x_2, \ldots, x_{k+1})) \\
&\ge \frac{n}{2}\binom{n/4}{k} k!(1+o(1))\frac{1}{2}\Big(\frac{1-\eta}{n}d\Big)^k\Big[\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k \\
&\ge \frac{n}{5}\Big(\frac{n}{4}\Big)^k\Big(\frac{1-\eta}{n}d\Big)^k\Big[\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k \\
&= \frac{n}{5}\Big[\frac{1-\eta}{4}d\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k \\
&\ge n\Big[(1-\eta)\Big(\frac{1}{4\sqrt{e}}\Big)^d\Big]^k \\
&\ge n\exp\Big[\Big(\log(1-\eta)-1,9d\Big)k\Big] \\
&= n^{1-\frac{1,9d-\log(1-\eta)}{4d-2\log(1-\eta)}} \ge n^{1/2}
\end{aligned}$$

Niech $\hat{x} = (x_1, x_2, \ldots, x_{k+1})$, $\hat{y} = (y_1, y_2, \ldots, y_{k+1})$. Pokażemy, że zmienna losowa $Y = \sum_{\hat{x}} Y(\hat{x})$ jest skoncentrowana wokół swojej wartości oczekiwanej. W tym celu policzmy wariancję

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\, Y &= \mathbb{V}\mathrm{ar}\Big(\sum_{\hat{x}} Y(\hat{x})\Big) \\
&= \sum_{\hat{x}} \mathbb{V}\mathrm{ar}\, Y(\hat{x}) + \sum_{\hat{x},\hat{y}} \mathbb{C}\mathrm{ov}\Big(\mathrm{Y}(\hat{\mathrm{x}}), \mathrm{Y}(\hat{\mathrm{y}})\Big) \ .
\end{aligned}$$

Zauważmy, że $\mathbb{E}Y(\hat{x}) = \mathbb{E}Y(\hat{x})^2$ (zmienna losowa $Y(\hat{x})$ przyjmuje wyłącznie wartości 0 lub 1). Możemy zatem łatwo oszacować pierwszy składnik

$$\begin{aligned}
\sum_{\hat{x}} \mathbb{V}\mathrm{ar}\, Y(\hat{x}) &= \sum_{\hat{x}} \big(\mathbb{E}Y(\hat{x})^2 - (\mathbb{E}Y(\hat{x}))^2\big) \le \sum_{\hat{x}} \mathbb{E}Y(\hat{x})^2 \\
&= \sum_{\hat{x}} \mathbb{E}Y(\hat{x}) = \mathbb{E}\Big(\sum_{\hat{x}} Y(\hat{x})\Big) = \mathbb{E}Y \ ,
\end{aligned}$$

drugi zaś oszacujemy korzystając z wprowadzonych wcześniej zdarzeń $A(\hat{x})$

$$\begin{aligned}\sum_{\hat{x},\hat{y}} \mathbb{C}\mathrm{ov}(\mathrm{Y}(\hat{\mathrm{x}}),\mathrm{Y}(\hat{\mathrm{y}})) &= \sum_{\hat{x},\hat{y}} \left[\mathbb{E}Y(\hat{x})Y(\hat{y}) - \mathbb{E}Y(\hat{x})\mathbb{E}Y(\hat{y})\right] \\ &= \sum_{\hat{x},\hat{y}} \left[\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) - \mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))\right] .\end{aligned}$$

Zauważmy, że dwie ścieżki $\hat{x} = (x_1, x_2, \ldots, x_{k+1})$ oraz $\hat{y} = (y_1, y_2, \ldots, y_{k+1})$ mogą być albo wierzchołkowo rozłączne albo stykać się w punkcie x_1. Te dwa przypadki należy rozważyć osobno. Przypuśćmy najpierw, że ścieżki $\hat{x}$, $\hat{y}$ są rozłączne. Zwróćmy uwagę, że ze względu na rozłączność wierzchołkową ścieżek, istnienie krawędzi w drugiej ścieżce nie wpływa znacząco na prawdopodobieństwo istnienia ustalonej krawędzi z pierwszej ścieżki oraz na prawdopodobieństwo bycia wierzchołkiem izolowanym. Prawdopodobieństwa te mogą różnić się co najwyżej o czynnik $(1 + o(1))$ (patrz Twierdzenie 5). Zatem na podstawie (7) mamy

$$\begin{aligned}\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) &= (1+o(1)) \prod_{i=1}^{k} \mathbb{P}(x_i \textit{ jest połączony z } x_{i+1}) \\ &\quad \times \prod_{i=1}^{k} \mathbb{P}(y_i \textit{ jest połączony z } y_{i+1}) \\ &\quad \times \prod_{i=2}^{k+1} \mathbb{P}(x_i \textit{ nie ma sąsiadów poza sąsiadami w ścieżce}) \\ &\quad \times \prod_{i=2}^{k+1} \mathbb{P}(y_i \textit{ nie ma sąsiadów poza sąsiadami w ścieżce}) \\ &\quad \times \mathbb{P}(x_1 \textit{ należy do dużej składowej}) \\ &\quad \times \mathbb{P}(y_1 \textit{ należy do dużej składowej}) \\ &= (1+o(1))\mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y})) .\end{aligned}$$

Również w przypadku, gdy ścieżki $\hat{x}$ oraz $\hat{y}$ mają dokładnie jeden wierzchołek wspólny ($z = x_1 = y_1$), prawdopodobieństwa po prawej stronie równości (7) mogą różnić się co najwyżej o czynnik $(1+o(1))$. Zwróćmy bowiem uwagę, że sąsiedzi wspólnego wierzchołka a posiadają większe numery ($x_2 > a$, $y_2 > a$).

Zatem, gdy ścieżki są połączone mamy

$$\begin{aligned}
\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) &= (1+o(1)) \prod_{i=1}^{k} \mathbb{P}(x_i \textit{ jest połączony z } x_{i+1}) \\
&\quad \times \prod_{i=1}^{k} \mathbb{P}(y_i \textit{ jest połączony z } y_{i+1}) \\
&\quad \times \prod_{i=2}^{k+1} \mathbb{P}(x_i \textit{ nie ma sąsiadów poza sąsiadami w ścieżce}) \\
&\quad \times \prod_{i=2}^{k+1} \mathbb{P}(y_i \textit{ nie ma sąsiadów poza sąsiadami w ścieżce}) \\
&\quad \times \mathbb{P}(a = x_1 = y_1 \textit{ należy do dużej składowej}) \\
&= \frac{(1+o(1))\mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))}{\mathbb{P}(a = x_1 = y_1 \textit{ należy do dużej składowej})} \\
&= O(1)\mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y})) \, .
\end{aligned}$$

Oznacza to, że

$$\begin{aligned}
\sum_{\hat{x},\hat{y}} \mathbb{C}\mathrm{ov}(\mathrm{Y}(\hat{\mathrm{x}}), \mathrm{Y}(\hat{\mathrm{y}})) &= \sum_{\hat{x},\hat{y}} \big[\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) - \mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))\big] \\
&= \sum_{\hat{x}} \Bigg[\sum_{\hat{y}, y_1 \neq x_1} \big[\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) - \mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))\big] \\
&\qquad + \sum_{\hat{y}, y_1 = x_1} \big[\mathbb{P}(A(\hat{x}) \cap A(\hat{y})) - \mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))\big] \Bigg] \\
&= \mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y})) \sum_{\hat{x}} \Bigg[\sum_{\hat{y}, y_1 \neq x_1} o(1) + \sum_{\hat{y}, y_1 = x_1} O(1) \Bigg] \\
&= (1+o(1)) \sum_{\hat{x},\hat{y}} o\Big(\mathbb{P}(A(\hat{x}))\mathbb{P}(A(\hat{y}))\Big) \\
&= o\Big(\sum_{\hat{x}} \mathbb{P}(A(\hat{x})) \sum_{\hat{y}} \mathbb{P}(A(\hat{y}))\Big) \\
&= o\Bigg(\Big(\mathbb{E} \sum_{\hat{x}} Y(\hat{x}) \Big)^2 \Bigg) = o\Big((\mathbb{E}Y)^2 \Big) \, .
\end{aligned}$$

Ostatecznie więc $\mathbb{V}\mathrm{ar}\, Y = o\Big((\mathbb{E}Y)^2\Big)$, co na mocy nierówności Czebyszewa oznacza, że prawie na pewno istnieje choć jedna taka ścieżka. □

3.3 Czas powrotu

W dotychczasowych rozważaniach badaliśmy własności grafu proteuszowego $\mathcal{P}_n(d,\eta)$, w tym rozdziale rozpatrywać będziemy natomiast własności procesu proteuszowego $\mathfrak{P}_n(d,\eta) = \{(\mathcal{P}_n^t(d,n),\sigma_t)\}_{t=0}^{\infty}$. Przedmiotem naszych zainteresowań będzie, nie mający odpowiednika w innych modelach struktur losowych, "czas powrotu", to jest czas po którym graf proteuszowy odzyskuje "typową" dla siebie własność, którą w trakcie procesu proteuszowego utracił. Zacznijmy od formalnej definicji tej wielkości.

Niech $\mathcal{A}$ będzie dowolną własnością taką, że graf $\mathcal{P}_n(d,\eta)$ posiada własność $\mathcal{A}$ z prawdopodobieństwem $1-o(1)$, lecz definiując $\tau(\mathcal{A})$ jako

$$\tau(A) = \min\{t : \mathcal{P}_n^t(d,\eta) \textit{ nie posiada własności } \mathcal{A}\}\,,$$

otrzymujemy $\Pr(\tau(\mathcal{A}) < \infty) = 1$. Oznacza to, że z prawdopodobieństwem równym jeden, w pewnym kroku procesu proteuszowego $\mathfrak{P}_n(d,\eta)$ własność $\mathcal{A}$ przestanie występować przez jakiś czas. Czasem powrotu $\mathrm{rec}(\mathcal{A})$ dla własności $\mathcal{A}$ nazwiemy zmienną losową zdefiniowaną wzorem

$$\mathrm{rec}(\mathcal{A}) = \min\{t > \tau(\mathcal{A}) : \mathcal{P}_n^t(d,\eta) \textit{ posiada } \mathcal{A}\} - \tau(\mathcal{A})\,.$$

Zauważmy, że po upływie $O(n\log n)$ kroków procesu prawie na pewno każdy z wierzchołków $\mathcal{P}_n(d,\eta)$ zostanie co najmniej raz wybrany, a zatem, ponieważ $\mathcal{P}_n(d,\eta)$ ma własność $\mathcal{A}$ prawie na pewno, z prawdopobieństwem $1-o(1)$ zachodzi

$$\mathrm{rec}(\mathcal{A}) = O(n\log n)\,.$$

Powyższe ograniczenie górne, prawdziwe jest dla każdej własności $\mathcal{A}$ i dowolnego wyboru parametrów d oraz η, który gwarantuje, że graf proteuszowy ma własnośc $\mathcal{A}$ prawie na pewno. Następujące twierdzenie pokazuje, że czas powrotu $\mathrm{rec}(\mathcal{C})$ dla spójności grafu jest nieco krótszy.

Twierdzenie 26. *Niech $\eta \in (0,1)$ oraz $d = a\log n$, gdzie $a > 1/g(x_0)$, gdzie $x_0 = x_0(\eta) \in (0,1)$ będzie punktem, w którym funkcja $g : (0,1) \longrightarrow \mathbb{R}$, określona wzorem*

$$g(x) = \frac{1-\eta}{1+\eta}(x^{-\eta} - x) - \log(1-x)\ ,$$

posiada minimum, tj. $x_0(\eta)$ jest rozwiązaniem równania

$$(1-\eta)\eta x^{-1-\eta} + 1 - \eta = \frac{1+\eta}{1-x}\ .$$

Wtedy

$$\mathrm{rec}(\mathcal{C})\frac{(1-\eta)a}{(x_0)^\eta}\frac{\log n}{n} \xrightarrow{D} Z\ ,$$

gdzie zmienna losowa Z, posiada rozkład wykładniczy, to znaczy, jej gęstość dana jest wzorem

$$f_Z(z) = \begin{cases} e^{-z} & dla\ z \geq 0 \\ 0 & dla\ z < 0\,. \end{cases}$$

Dowód. Pokażemy najpierw, że prawie na pewno w momencie $\tau(\mathcal{C})$, gdy graf proteuszowy po raz pierwszy staje się niespójny, składa się on z dużej składowej oraz pojedynczego wierzchołka izolowanego i, takiego, że

$$\sigma_{\tau(\mathcal{C})}(i) = (1+o(1))x_0 n\ .$$

Przypomnijmy, że nie jest to wierzchołek o najmniejszym oczekiwanym stopniu w grafie proteuszowym, ale taki, dla którego prawdopobieństwo bycia izolowanym jest największe (patrz Twierdzenie 11). Dodajmy, że funkcja g jest funkcją ciągłą posiadającą w punkcie x_0 minimum, a zatem wystarczy pokazać, że

$$g(\sigma_{\tau(\mathcal{C})}(i)/n) \leq (1+o(1))g(x_0)\ .$$

Rozważmy pierwsze $n\log^2 n$ kroków procesu proteuszowego. Oznaczmy przez $\rho_1(\varepsilon)$, $\rho_2(\varepsilon)$ oraz ρ_3 prawdopodobieństwa, że w tym okresie w procesie proteuszowym pojawi się odpowiednio:

- izolowany wierzchołek i dla którego $g(\sigma(i)/n) \in [\, g(x_0), (1+\varepsilon)g(x_0))$,
- izolowany wierzchołek i dla którego $g(\sigma(i)/n) \geq (1+\varepsilon)g(x_0)$,
- składowa o rozmiarze k, $2 \leq k \leq 2n/3$.

W badaniu zachowania powyższych wielkości przydatne będzie oszacowanie prawdopodobieństwa $\rho(i,j,t)$, że wierzchołek i stał się wierzchołkiem izolowanym w kroku t procesu proteuszowego. Może się to zdażyć tylko wtedy, gdy wybrano jedynego sąsiada wierzchołka i, wierzchołek j. Korzystając z Twierdzenia 5 możemy oszacować $\rho(i,j,t)$ przez

$$\begin{aligned} &o(n\exp(-\log^{3/2} n)) + (1+o(1))\frac{1}{n}p(\ell_j,\ell_i) \\ &\qquad \times \prod_{r=1}^{\ell_i-1}\Big[1-(1+O(\log^{-1/2} n))\frac{1-\eta}{n}\Big(\frac{\ell_i}{r}\Big)^{\eta}\Big]^d \\ &\qquad \times \prod_{s=\ell_i+1,s\neq\ell_j}^{n}\Big[1-(1+O(\log^{-1/2} n))\frac{1-\eta}{n}d\Big(\frac{s}{\ell_i}\Big)^{\eta}\Big] \\ &= n^{-2+o(1)}\Big(\frac{\ell_j}{\ell_i}\Big)^{\eta}\exp\Big(-(1+o(1))g\Big(\frac{\ell_i}{n}\Big)d\Big) \end{aligned}$$

dla $\ell_i < \ell_j$, oraz

$$\begin{aligned} &o(n\exp(-\log^{3/2} n)) + (1+o(1))\frac{1}{n}p(\ell_i,\ell_j) \\ &\qquad \times \prod_{r=1,r\neq\ell_j}^{\ell_i-1}\Big[1-(1+O(\log^{-1/2} n))\frac{1-\eta}{n}\Big(\frac{\ell_i}{r}\Big)^{\eta}\Big]^d \\ &\qquad \times \prod_{s=\ell_i+1}^{n}\Big[1-(1+O(\log^{-1/2} n))\frac{1-\eta}{n}d\Big(\frac{s}{\ell_i}\Big)^{\eta}\Big] \\ &= n^{-2+o(1)}\Big(\frac{\ell_i}{\ell_j}\Big)^{\eta}\exp\Big(-(1+o(1))g\Big(\frac{\ell_i}{n}\Big)d\Big) \end{aligned}$$

dla $\ell_i > \ell_j$, gdzie $\ell_i = \sigma_t(i)$ oraz $\ell_j = \sigma_t(j)$ oznaczają miejsce w permutacji odpowiednio wierzchołka i oraz j w t-tym kroku procesu proteuszowego.

Niech $\varepsilon > 0$. Oznaczmy przez $\mathbf{A}_t(i)$ zdarzenie, że wierzchołek i stał się wierzchołkiem izolowanym w t-tym kroku procesu oraz w tym momencie

$$g(\sigma_t(i)/n) \in [g(x_0), (1+\varepsilon/4)g(x_0)) \ ,$$

przez $\mathbf{A}_t = \bigcup_{i=1}^n \mathbf{A}_t(i)$ oznaczmy zdarzenie, że w t-tym kroku procesu proteuszowego pojawił się wierzchołek izolowany o tej własności. Podobnie niech $\mathbf{B}'_t(i)$ oraz $\mathbf{B}'_t$ będą analogicznymi zdarzeniami, lecz w tym momencie żądamy, aby

$$g(\sigma_t(i)/n) > (1+\varepsilon)g(x_0) \ .$$

Szacując prawdopodobieństwo występowania zdarzenia $\mathbf{A}_t(i)$ otrzymujemy

$$\begin{aligned}
\mathbb{P}(\mathbf{A}_t(i)) &= \sum_{j=1, j\neq i}^{n} \rho(i,j,t) \\
&= \Big[\sum_{\ell_j<\ell_i} \Big(\frac{\ell_i}{\ell_j}\Big)^{\eta} + \sum_{\ell_j>\ell_i} \Big(\frac{\ell_j}{\ell_i}\Big)^{\eta} \Big] n^{-2+o(1)} \exp\Big(-(1+o(1)) g\Big(\frac{\ell_i}{n}\Big) d \Big) \\
&= \Big[\ell_i^{\eta} n \int_0^{\ell_i/n} (xn)^{-\eta} dx + \ell_i^{-\eta} n \int_{\ell_i/n}^{1} (xn)^{\eta} dx \Big] n^{-2-(1+o(1))ag(\ell_i/n)} \\
&= \Big[\ell_i^{\eta} n^{1-\eta} \frac{(\ell_i/n)^{1-\eta}}{1-\eta} + \ell_i^{-\eta} n^{1+\eta} \frac{1-(\ell_i/n)^{1+\eta}}{1+\eta} \Big] n^{-2-(1+o(1))ag(\ell_i/n)} \\
&= \Big[\frac{\ell_i}{1-\eta} + \frac{n}{1+\eta} (n/\ell_i)^{\eta} - \frac{\ell_i}{1+\eta} \Big] n^{-2-(1+o(1))ag(\ell_i/n)} \\
&= \Theta(n) n^{-2-(1+o(1))ag(\ell_i/n)} = n^{-1-(1+o(1))ag(\ell_i/n)}.
\end{aligned}$$

Otrzymujemy zatem następujące oszacowania

$$n^{-1-(1+o(1))(1+\varepsilon/4)ag(x_0)} \leq \ \mathbb{P}(\mathbf{A}_t(i)) \ \leq n^{-1-(1+o(1))ag(x_0)}$$

oraz

$$\mathbb{P}(\mathbf{B}'_t(i)) \ \leq \ n^{-1-(1+o(1))(1+\varepsilon)ag(x_0)} \ .$$

Z ostatniej nierówności otrzymujemy natychmiast

$$\mathbb{P}(\mathbf{B}'_t) \leq \sum_{i=1}^{n} \mathbb{P}(\mathbf{B}'_t(i)) \leq n^{-(1+o(1))(1+\varepsilon)ag(x_0)}.$$

Pokażemy teraz, że dla dowolnych $i, i' \in [n]$, takich, że $\ell_i < \ell_{i'}$ zdarzenia $\mathbf{A}_t(i)$ oraz $\mathbf{A}_t(i')$ zależą od siebie w niewielkim stopniu. Podobnie jak poprzednio możemy oszacować prawdopodobieństwo $\rho(i, i', j, t)$, że wierzchołki i oraz i' stały się wierzchołkami izolowanymi, ponieważ w t-tym kroku procesu proteuszowego, wybrano jedynego ich sąsiada, wierzchołek j przez

$$n^{-3+o(1)}\Big(\frac{\max\{\ell_i,\ell_j\}}{\min\{\ell_i,\ell_j\}}\Big)^{\eta}\Big(\frac{\max\{\ell_{i'},\ell_j\}}{\min\{\ell_{i'},\ell_j\}}\Big)^{\eta} \\ \times \exp\Big(-(1+o(1))g\Big(\frac{\ell_i}{n}\Big)d\Big)\exp\Big(-(1+o(1))g\Big(\frac{\ell_{i'}}{n}\Big)d\Big)\,.$$

Oznacza to, że gdy $\eta < 1/2$

$$\begin{aligned}
\mathbb{P}(\mathbf{A}_t(i)\cap\mathbf{A}_t(i')) &= \sum_{j=1,j\neq i,j\neq i'}^{n}\rho(i,i',j,t)\\
&= \Big[\sum_{\ell_j<\ell_i}\Big(\frac{\ell_i}{\ell_j}\Big)^{\eta}\Big(\frac{\ell_{i'}}{\ell_j}\Big)^{\eta}+\sum_{\ell_i<\ell_j<\ell_{i'}}\Big(\frac{\ell_j}{\ell_i}\Big)^{\eta}\Big(\frac{\ell_{i'}}{\ell_j}\Big)^{\eta}+\sum_{\ell_{i'}<\ell_j}\Big(\frac{\ell_j}{\ell_i}\Big)^{\eta}\Big(\frac{\ell_j}{\ell_{i'}}\Big)^{\eta}\Big] \qquad (8)\\
&\qquad\times n^{-3+o(1)}\exp\Big(-(1+o(1))g\Big(\frac{\ell_i}{n}\Big)d\Big)\exp\Big(-(1+o(1))g\Big(\frac{\ell_{i'}}{n}\Big)d\Big)\\
&= \Big[\ell_i^{\eta}\ell_{i'}^{\eta}n\int_0^{\ell_i/n}(xn)^{-2\eta}dx+(\ell_{i'}-\ell_i)\Big(\frac{\ell_{i'}}{\ell_i}\Big)^{\eta}+\ell_i^{-\eta}\ell_{i'}^{-\eta}n\int_{\ell_{i'}/n}^{1}(xn)^{2\eta}dx\Big]\\
&\qquad\times n^{-3-(1+o(1))ag(\ell_i/n)-(1+o(1))ag(\ell_{i'}/n)}\\
&= \Big[\ell_i^{\eta}\ell_{i'}^{\eta}n^{1-2\eta}\frac{(\ell_i/n)^{1-2\eta}}{1-2\eta}+(\ell_{i'}-\ell_i)\Big(\frac{\ell_{i'}}{\ell_i}\Big)^{\eta}+\ell_i^{-\eta}\ell_{i'}^{-\eta}n^{1+2\eta}\frac{1-(\ell_{i'}/n)^{1+2\eta}}{1+2\eta}\Big]\\
&\qquad\times n^{-3-(1+o(1))ag(\ell_i/n)-(1+o(1))ag(\ell_{i'}/n)}\\
&= \Big[\frac{\ell_i}{1-2\eta}\Big(\frac{\ell_{i'}}{\ell_i}\Big)^{\eta}+(\ell_{i'}-\ell_i)\Big(\frac{\ell_{i'}}{\ell_i}\Big)^{\eta}+\frac{n}{1+2\eta}\Big(\frac{n}{\ell_i}\Big)^{\eta}\Big(\frac{n}{\ell_{i'}}\Big)^{\eta}-\frac{\ell_{i'}}{1+2\eta}\Big(\frac{\ell_{i'}}{\ell_i}\Big)^{\eta}\Big]\\
&\qquad\times n^{-3-(1+o(1))ag(\ell_i/n)-(1+o(1))ag(\ell_{i'}/n)}\\
&= \Theta(n)n^{-3-(1+o(1))ag(\ell_i/n)-(1+o(1))ag(\ell_{i'}/n)}\\
&= n^{-1-(1+o(1))ag(\ell_i/n)}n^{-1-(1+o(1))ag(\ell_{i'}/n)}n^{o(1)}\\
&= \mathbb{P}(\mathbf{A}_t(i))\mathbb{P}(\mathbf{A}_t(i'))n^{o(1)}.
\end{aligned}$$

W przypadku, gdy $\eta \geq 1/2$ sumy występujące w (8) nie można przybliżać całkami i w tym przypadku

$$\Big[\sum_{\ell_j<\ell_i}\Big(\frac{\ell_i}{\ell_j}\Big)^{\eta}\Big(\frac{\ell_{i'}}{\ell_j}\Big)^{\eta}+\sum_{\ell_i<\ell_j<\ell_{i'}}\Big(\frac{\ell_j}{\ell_i}\Big)^{\eta}\Big(\frac{\ell_{i'}}{\ell_j}\Big)^{\eta}+\sum_{\ell_{i'}<\ell_j}\Big(\frac{\ell_j}{\ell_i}\Big)^{\eta}\Big(\frac{\ell_j}{\ell_{i'}}\Big)^{\eta}\Big]=\Theta(n^{2\eta})\ .$$

Ostatecznie więc

$$\mathbb{P}(\mathbf{A}_t(i)\cap\mathbf{A}_t(i'))=\mathbb{P}(\mathbf{A}_t(i))\mathbb{P}(\mathbf{A}_t(i'))n^{o(1)}\Theta(1+n^{2\eta-1})\ .$$

Aby oszacować $\mathbb{P}(\mathbf{A}_t)$ skorzystamy z nierówności Bonferroniego

$$\begin{aligned}\mathbb{P}(\mathbf{A}_t)&=\mathbb{P}\Big(\bigcup_{i=1}^{n}\mathbf{A}_t(i)\Big)\\&\geq\sum_{i=1}^{n}\mathbb{P}(\mathbf{A}_t(i))-\sum_{1\leq i<i'\leq n}\mathbb{P}(\mathbf{A}_t(i)\cap\mathbf{A}_t(i'))\\&=\sum_{i=1}^{n}\mathbb{P}(\mathbf{A}_t(i))-\sum_{1\leq i<i'\leq n}\mathbb{P}(\mathbf{A}_t(i))\mathbb{P}(\mathbf{A}_t(i'))n^{o(1)}\Theta(1+n^{2\eta-1})\\&\geq n^{-(1+o(1))(1+\varepsilon/4)ag(x_0)}\Big(1-n^{-(1+o(1))ag(x_0)}\Theta(1+n^{2\eta-1})\Big)\\&=(1-o(1))n^{-(1+o(1))(1+\varepsilon/4)ag(x_0)}\geq n^{-(1+\varepsilon/3)ag(x_0)}\ .\end{aligned}$$

Zauważmy, że fakt pojawienia się w kroku t_1 wierzchołka izolowanego nie wpływa znacząco na prawdopodobieństwo pojawienia się kolejnego wierzchołka izolowanego w kroku t_2 ($t_2 > t_1$). Pojawienie się wcześniej wierzchołków izolowanych może wpłynąć jedynie na wagi wierzchołków, lecz wpływ ten, zgodnie z Twierdzeniem 2, nie jest zbyt duży. Można pokazać, że dla dowolnych $t_1 < t_2$

$$\mathbb{P}(\mathbf{A}_{t_1}\cap\mathbf{A}_{t_2})=\mathbb{P}(\mathbf{A}_{t_1})\mathbb{P}(\mathbf{A}_{t_2})n^{o(1)}\ .$$

Ostatecznie więc

$$\rho_2(\varepsilon)=\sum_{t=1}^{n\log^2 n}\mathbb{P}(\mathbf{B}'_t)\leq n^{1-(1+o(1))(1+\varepsilon)ag(x_0)}$$

oraz

$$\begin{aligned}\rho_1(\varepsilon) &\geq \mathbb{P}\Big(\bigcup_{t=1}^{n\log^2 n}\mathbf{A}_t\Big)\\ &\geq \sum_{t=1}^{n\log^2 n}\mathbb{P}(\mathbf{A}_t) - \sum_{1\leq t_1<t_2\leq n\log^2 n}\mathbb{P}(\mathbf{A}_{t_1}\cap\mathbf{A}_{t_2}) \geq n^{1-(1+\varepsilon/2)ag(x_0)}\,.\end{aligned}$$

Co więcej można pokazać, że

$$\rho_3 = n^{1+o(1)}[\mathbb{P}(\mathbf{A}(i))]^2 \leq n^{1-(1+o(1))2ag(x_0)} \leq \rho_2(\varepsilon).$$

Rozważmy teraz $n^{(1+\frac{3}{4}\varepsilon)ag(x_0)}\log^2 n$ początkowych kroków procesu proteuszowego. Prawdopodobieństwo pojawienia się w tym czasie wierzchołka izolowanego i dla którego

$$g(\sigma(i)/n) \geq (1+\varepsilon)g(x_0)$$

wynosi $o(n^{-\frac{\varepsilon}{5}ag(x_0)})$. Również prawie na pewno nie pojawi się podczas tego okresu składowa o rozmiarze większym lub równym 2.

Niech $\mathbf{D}_k$, $k=0,1,\ldots,k_0$, gdzie $k_0 = n^{(1+\frac{3}{4}\varepsilon)ag(x_0)-1}/3$ będzie zdarzeniem polegającym na tym, że pomiędzy $2kn\log^2 n$ a $(2k+1)n\log^2 n$ krokiem pojawił się wierzchołek izolowany i, dla którego

$$g(\sigma(i)/n) \in [g(x_0),(1+\varepsilon)g(x_0))\,. \tag{9}$$

Niech $\mathbf{F}$ oznacza zdarzenie, że każdy wierzchołek w grafie proteuszowym został wybrany co najmniej raz w czasie $t\in((2k-1)n\log^2 n, 2kn\log^2 n)$, dla każdego $k=1,\ldots,k_0$. Zauważmy, że $\mathbb{P}(\mathbf{F}')\leq k_0 n\exp(-\log^2 n/2)\leq \exp(-\log^{3/2} n/2)$, oraz $\mathbb{P}(\mathbf{D}_k)=\rho_2(\varepsilon)$, co więcej warunkując przez $\mathbf{F}$, wszystkie zdarzenia $\mathbf{D}_k$ są niezależne. Ponieważ $k_0\rho_1(\varepsilon)\to\infty$ przy $n\to\infty$, otrzymujemy, że $\mathbb{P}\Big(\bigcup_{k=0}^{k_0}\mathbf{D}_k\Big)\to 1$. Oznacza to, że $\tau(\mathcal{C}) = n^{(1+o(1))ag(x_0)}$ oraz, że w momencie $\tau(\mathcal{C})$ graf proteuszowy składa się z dużej składowej oraz pojedynczego wierzchołka izolowanego i_0, dla którego $\sigma(i)=(1+o(1))x_0 n$.

Pokażemy teraz, że po kolejnych $\Theta(n/\log n)$ krokach procesu graf znów będzie spójny. Prawdopodobieństwo, że wybierzemy wierzchołek i_0 w trakcie tego okresu dąży do 0, gdy $n \to \infty$. Argumentując podobnie jak w przypadku oszacowań liczb $\rho_1(\varepsilon)$, $\rho_2(\varepsilon)$ oraz ρ_3 możemy pokazać, że prawdopodobieństwo, że w tym czasie pojawią się inne wierzchołki izolowane bądź składowe dąży do 0. Oznacza to, że graf proteuszowy $\mathcal{P}_n(d,\eta)$ ma szansę ponownie zostać spójnym dzięki temu, że inny wierzchołek wybierze wierzchołek i_0 na swojego sąsiada. Ponieważ waga wierzchołka i_0 zmieni się nieznacznie podczas $\Theta(n/\log n)$ kroków, prawdopodobieństwo, że dla dowolnego $z \geq 0$

$$\operatorname{rec}(\mathcal{C}) \geq z\frac{(x_0)^\eta}{(1-\eta)a}\frac{n}{\log n} = z\frac{(x_0)^\eta}{1-\eta}\frac{n}{d}\,,$$

wynosi

$$\begin{aligned}\Big[1-(1+o(1))(1-\eta)\frac{d}{n}\Big(\frac{n}{x_0 n}\Big)^\eta\Big]^{z\frac{(x_0)^\eta}{1-\eta}\frac{n}{d}} &\\ &= \Big[1-(1+o(1))(1-\eta)\frac{d}{n}(x_0)^{-\eta}\Big]^{z\frac{(x_0)^\eta}{1-\eta}\frac{n}{d}}\\ &= e^{-(1+o(1))z}\\ &= (1+o(1))e^{-z},\end{aligned}$$

co kończy dowód. □

4 Symulacje

Podczas pisania rozprawy doktorskiej przeprowadzono szereg symulacji. Zazwyczaj potwierdzały one wcześniejsze rozważania teoretyczne, ale zdarzało się również, że pomagały w wyborze właściwej techniki dowodzenia zaobserwowanych własności. W niniejszym rozdziale przedstawiono wybrane wyniki. Ze względu na możliwość modelowania rzeczywistej sieci internetowej, skupiono się na dwóch szczególnych wartościach parametru η ($\eta_{\text{out}} = 0{,}59$ i $\eta_{\text{in}} = 0{,}91$) oraz w przypadku, gdy średni stopień w grafie proteuszowym nie jest zbyt duży ($d = 10$).

4.1 Środowisko

Do symulacji wykorzystano klaster składający się z 16 komputerów (Pentium III 500MHz – 8 komputerów; Celeron 466 – 8 komputerów). Na maszynach zainstalowano system operacyjny Red Hat Linux 7.3 [31] oraz pakiet Mosix [21, 12, 11, 30]. Pakiet ten jest cały czas udoskonalany przez prof. Amnona Baraka pracującego na Hebrew University. Jego nazwa to skrót od "Multicomputer Operating System for Unix". Oprogramowanie to, choć darmowe, przewyższa wydajnością wiele systemów tego typu i pozwala połączyć ze sobą aż 65.536 komputerów. Dzięki umiejętności sprawdzania obciążenia procesorów w sieci heterogenicznej, do jego budowy można stosować komputery o różnej mocy obliczeniowej, zarówno te wyposażone w nowoczesne i szybkie, jak i starsze i mniej wydajne procesory. Skonfigurowany i uruchomiony klaster "zachowuje się" jak wieloprocesorowy komputer. Jest to podstawowa cecha, która wyróżnia go spośród pozostałych rozwiązań klastrowych. Na tym też polega prostota jego obsługi. Nie ma tutaj zcentralizowanego serwera, który zarządza wszystkimi procesami, a zatem programy działające na tego typu klastrze należy pisać identycznie jak w przypadku maszyny wieloprocesorowej [14].

Program generujący graf proteuszowy oraz badający jego strukturę wewnętrzną został napisany w większości w języku C. O jego wyborze zadecydowała szybkość działania programu – badane grafy są stosunkowo duże. Aby jednak ułatwić wyprowadzanie danych oraz dynamiczne rezerwowanie pamięci skorzystano z udogodnień języka C++.

4.2 Podstawowe własności

Eksperymenty dowiodły, że w rzeczywistych sieciach współczynnik skupienia jest znacząco większy niż d/n, gdzie d jest średnim stopniem w grafie [29]. Współczynnik skupienia dla grafów proteuszowych jest istotnie większy od d/n, lecz przyznać należy, że daleko mu do wartości obserwowanych dla grafu internetowego, czy podobnych do niego *"small world" graphs*. Dla $n = 100.000$, $d = 10$, w przypadku, gdy $\eta = \eta_{\text{out}} = 0{,}59$ współczynnik skupienia wynosi $C^{\mathcal{P}_n(d,\eta_{\text{out}})} = 4,2 \cdot 10^{-4}$, zaś dla $\eta = \eta_{\text{in}} = 0{,}91$ różnica jest jeszcze większa $C^{\mathcal{P}_n(d,\eta_{\text{in}})} = 1,0099 \cdot 10^{-2}$ $(d/n = 10^{-5})$.

Kolejną własnością grafu jest rozmiar jego k- rdzenia (ang. *k-core*). k- rdzeń, to największy podgraf, w którym minimalny stopień wynosi co najmniej k. Oto rozmiary k- rdzeni grafów proteuszowych składających się z 200.000 wierzchołków, $d = 10$, w przypadku, gdy $\eta = \eta_{\text{out}} = 0{,}59$ oraz $\eta = \eta_{\text{in}} = 0{,}91$.

k	rozmiar k- rdzenia $\eta = 0{,}59$	rozmiar k- rdzenia $\eta = 0{,}91$
1	199.800 (99,9%)	196.960 (98,5%)
2	198.971 (99,5%)	189.102 (94,6%)
3	196.261 (98,1%)	172.552 (86,3%)
4	189.020 (94,5%)	142.924 (71,5%)
5	171.460 (85,7%)	83.691 (41,8%)
6	0	0

4.3 Składowe oraz ich średnice

W programie wykorzystano przeszukiwanie grafu proteuszowego wszerz w celu wyznaczenia rozmiarów składowych (dla $n = 200.000$). Ze względu na złożoność problemu wyznaczania średnicy, jej wartość obliczono w przypadku mniejszego grafu (dla $n = 100.000$). Z rozważań teoretycznych (patrz Twierdzenie 17) wynika, że średnica grafu proteuszowego (w przypadku, gdy $\eta \in [0{,}58; 0{,}92]$) wynosi $\Theta(\log n)$. Wykonane symulacje potwierdzają tę obserwację, gdzie zarówno dla $\eta = \eta_{out} = 0{,}59$ jak i dla $\eta = \eta_{in} = 0{,}91$ średnica wynosi 9 ($\log 100.000 \approx 11{,}5$). W tabeli poniżej przedstawiono w kolejnych kolumnach rozmiar składowej, liczbę składowych o danym rozmiarze, średnicę oraz numery wierzchołków na przykładowej ścieżce o długości równej średnicy składowej.

$n = 200.000$, $d = 10$, $\eta_{out} = 0{,}59$	
rozmiar	liczba składowych
199.977	1
1	23

$n = 200.000$, $d = 10$, $\eta_{in} = 0{,}91$	
rozmiar	liczba składowych
199.547	1
2	4
1	445

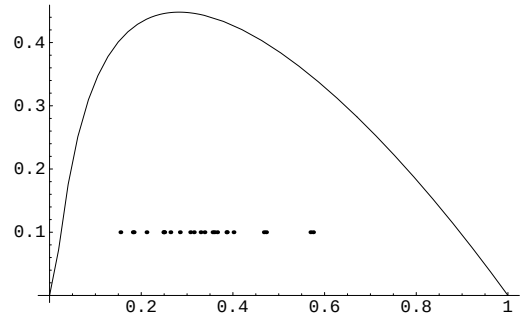
$n = 100.000$, $d = 10$, $\eta_{out} = 0{,}59$			
rozmiar	liczba	średnica	wierzchołki
99.990	1	9	39531-41663-25696-831-48428- -4405-76143-83288-23121-30371
1	10	0	

$n = 100.000$, $d = 10$, $\eta_{in} = 0{,}91$			
rozmiar	liczba	średnica	wierzchołki
99.817	1	9	21123-44104-22082-42435-1- -99828-15-77587-10745-46403
2	1	1	25472-28262
1	181	0	

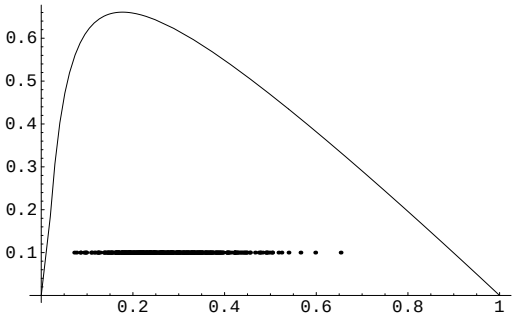
Zwróćmy uwagę, że najdłuższa ścieżka w grafie proteuszowym w przypadku, gdy $\eta = \eta_{\text{in}} = 0{,}91$ przechodzi przez wierzchołek 1 (wierzchołek "najstarszy"), później "skacze" do wierzchołka 99828 (wierzchołek bardzo "młody"), by ponownie wrócić do początku, do wierzchołka 15. W przypadku, gdy $\eta = \eta_{\text{out}} = 0{,}59$ również można zaobserwować podobną tendencję.

4.4 "Kryzys wieku średniego"

W Rozdziale 2.3 znajdują się rozważania na temat prawdopodobieństwa, że ustalony wierzchołek, w grafie proteuszowym $\mathcal{P}_n(d, \eta)$, jest wierzchołkiem izolowanym. Pokazano, że dla każdego $\eta \in (0{,}1)$ istnieje $x_0(\eta) \in (0, 1)$, takie, że prawdopodobieństwo bycia wierzchołkiem izolowanym jest największe dla wierzchołka o numerze $(1 + o(1))x_0(\eta)n$. Stosunkowo łatwo można pokazać, że w interesujących nas przypadkach, gdy $\eta = \eta_{\text{out}} = 0{,}59$ oraz $\eta = \eta_{\text{in}} = 0{,}91$ maksimum znajduje się odpowiednio w punktach $x_0(0{,}59) \approx 0{,}282$ oraz $x_0(0{,}91) \approx 0{,}177$.



$\eta = 0{,}59$



$\eta = 0{,}91$

Na powyższych wykresach przedstawiono prawdopodobieństwo bycia wierzchołkiem izolowanym w zależności od położenia wierzchołka w grafie proteuszowym (dla $d = 1$). W wyniku przeprowadzonych symulacji (dla $n = 200.000$ oraz $d = 10$) otrzymaliśmy zbiór wierzchołków izolowanych. Na wykresach naniesiono również te wartości x_i, dla których wierzchołki $\lceil x_i n \rceil$ są wierzchołkami izolowanymi.

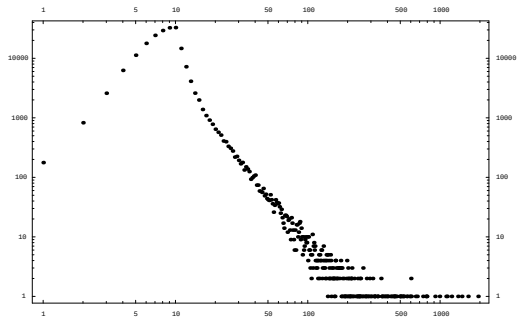
4.5 Rozkład stopni

Rozkład stopni w grafie proteuszowym jest rozkładem potęgowym (patrz Rozdział 1.2). Zgodnie z Twierdzeniem 8 rozkład stopni w grafie proteuszowym $\mathcal{P}_n(d, \eta)$ również jest rozkładem potęgowym (potęga zależy od wyboru parametru η), tzn.

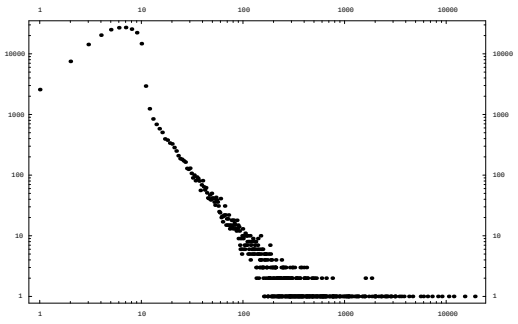
$$P(\deg(i) = k) \sim k^{-1-1/\eta} \text{ dla każdego } i \in [n]\,.$$

Oznacza to, że dla parametru $\eta = \eta_{\text{out}} = 0{,}59$ oraz $\eta = \eta_{\text{in}} = 0{,}91$ powinniśmy otrzymać identyczne rozkłady jak w grafie internetowym, odpowiednio dla stopni wyjściowych (potęga wynosi $-2{,}7$) oraz stopni wejściowych (potęga wynosi $-2{,}1$).

Symulacje przeprowadzone dla powyższych parametrów η oraz w przypadku, gdy $n = 200.000$, $d = 10$ potwierdzają teoretyczne rozważania. Na poniższych wykresach przedstawiono liczbę wierzchołków (oś Y) posiadających dany stopień (oś X). Wykresy te przedstawiono w skali logarytmiczno– logarytmicznej. Liniowy rozkład punktów w tej skali świadczy o właściwym rozkładzie (rozkładzie potęgowym). Współczynniki nachylenia prostych są bardzo bliskie oczekiwanych wartości wynikających z Twierdzenia 8.



$\eta = 0{,}59$



$\eta = 0{,}91$

W przypadku, gdy $\eta = \eta_{\text{out}} = 0{,}59$ maksymalny stopień wyniósł 1.936, natomiast w grafie proteuszowym dla $\eta = \eta_{\text{in}} = 0{,}91$ wierzchołek o największym stopniu posiadał aż 19.144 sąsiadów.

Literatura

[1] W. Aiello, F. Chung, L. Lu, *A random graph model for massive graphs*, Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, 2000, 171–180.

[2] R. Albert, H. Jeong, A. Barabási, *Diameter of the World-Wide Web*, Nature **401**, (1999) 130–131.

[3] K.B. Athreya, P.E. Ney, *Branching procsses*, Die Grundlehren der mathematishen Wissenschaften, Band 196, Springer, Berlin, 1972.

[4] P. Billingsley, *Prawdopodobieństwo i miara*, PWN, Warszawa, 1987.

[5] A.A. Borowkow, *Rachunek prawdopodobieństwa*, PWN, Warszawa, 1975.

[6] B. Bollobás, *Random Graphs*, Cambridge University Press, Cambridge, 2001.

[7] B. Bollobás, C. Borgs, J. Chayes, O. Riordan, *Directed scale-free graphs*, Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms, 2003, 132–139.

[8] B. Bollobás, O. Riordan, J. Spencer, G.E. Tusnády, *The degree sequence of a scale-free random graph process*, Random Structures and Algorithms **18** (2001) 279–290.

[9] D. Clark, *Face-to-Face with Peer-to-Peer Networking*, IEEE Computer **34** (2001) 18–21.

[10] C. Cooper, A. Frieze, *A general model of web graphs*, Random Structures and Algorithms **22** (2002) 311–335.

[11] A. Barak, O. La'adan, *The MOSIX Multicomputer Operating System for High Performance Cluster Computing* , Journal of Future Generation Computer Systems **13** (1998) 361–372.

[12] A. Barak, O. La'adan, A. Shiloh, *Scalable Cluster Computing with MOSIX for Linux*, Proc. Linux Expo'99, Raleigh, 1999, 95–100.

[13] A. Broder, R. Kumar, F. Maghoul, P. Rahaghavan, S. Rajagopalan, R. State, A. Tomkins, J. Wiener, *Graph structure in the web*, Proceedings 9th International World-Wide Web Conference (WWW), 2000, 309–320.

[14] J.S. Gray, *Arkana – komunikacja między procesami w Unixie*, RM 1999.

[15] J. Gross, J. Yellen, *Graph Theory and its applications*, CRC Press, 1999.

[16] J.W. Grossman, *The Evolution of the Mathematical Research Collaboration Graph*, Proceedings of 33rd Southeastern Conference on Combinatorics, Congressus Numerantium **158** (2002) 201–212.

[17] S. Janson, T. Łuczak, A. Ruciński, *Random Graphs*, Wiley, New York, 2000.

[18] M. Kocken, *The Small World*, Ablex (Norwood, NJ), 1989.

[19] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, *Stochastic models for the Web graph*, Proceedings of the 41th IEEE Symp. on Foundations of Computer Science (FOCS), 2000, 57–65.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, *Trawling the Web for emerging cyber-communities*, Proceedings of 8th International World-Wide Web Conference (WWW), 1999, 1481–1493.

[21] S. McClure, R. Wheeler, *MOSIX: How Linux Clusters Solve Real World Problems*, Proceedings of the 2000 USENIX Annual Tech. Conf., San Diego, 2000, 49–56.

[22] T. Łuczak, P. Prałat, *Protean graphs*, praca złożona do druku.

[23] S. Milgram, *The small world problem*, Psychology Today **2** (1967) 60–67.

[24] G. Pandurangan, P. Raghavan, E. Upfal, *Building low-diameter P2P networks*, Proceedings of the 42th IEEE Symp. on Foundations of Computer Science (FOCS), 2001, 492–499.

[25] T. Remes, *Six Degrees of Rogers Hornsby*, New York Times, August 17, 1997.

[26] B. Tjaden, G. Wasson, `http://www.cs.virginia.edu/oracle/`, 1997.

[27] D.J. Watts, *Small Worlds*, Princeton University Press, Princeton, 1999.

[28] D.J. Watts, *Six Degrees: The Science of a Connected Age*, W.W. Norton & Company, 2003.

[29] D.J. Watts, S.H. Strogatz, *Collective dynamics of "small-world" networks*, Nature **393** (1998) 440–442.

[30] Dokumentacja pakietu Mosix, `http://www.mosix.org`.

[31] Dokumentacja systemu Red Hat Linux 7.3, `http://www.redhat.com`.