TASS Projekt 1

Filip Mazur, 300224

Semestr 2022/23Z

1 Zadanie A - analiza grafu za pomocą programu Pajek

Dane 1. - Stacje metra londyńskiego i łączące je linie

- Zbadaj, jaki jest rząd i rozmiar całej sieci, a następnie wyodrębnij największą składową spójną, zbadaj jej rząd i rozmiar

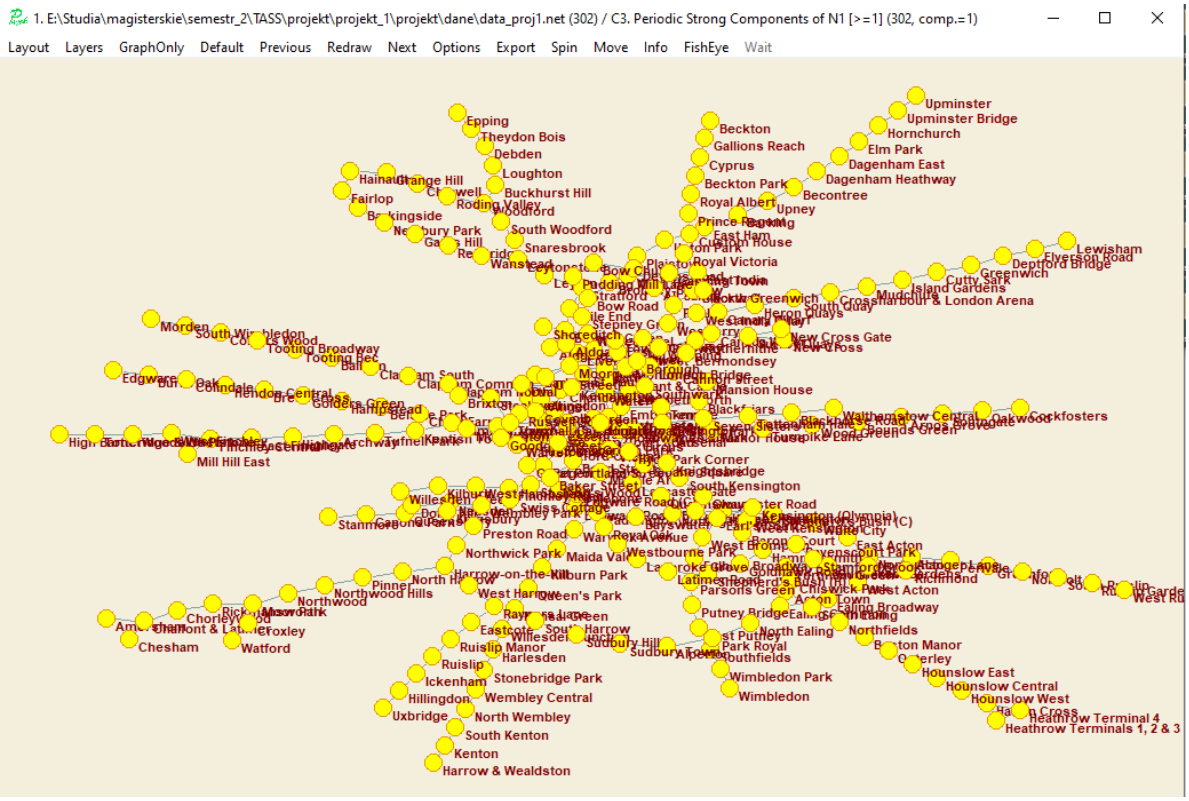
Parametry wczytanego grafu:
- Rząd: 302
- Rozmiar: 350

Składowa spójna
Składowa spójna w grafie nieskierowanym może być wyznaczona jako **strongly connected component** lub **weakly connected component**. Różnice pojęć pojawiają się w przypadku grafów skierowanych. W naszym przypadku mamy do czynienia z grafem nieskierowanym, więc można wyznaczyć największy strongly lubweakly connected component, co będzie największą składową spójną sieci.

W programie wybrałem Network → Create Partition → Components → Weak (o rozmiarze minimum 1). Algorytm zwrócił jeden komponent. Oznacza to, że wczytany graf jest grafem spójnym.

Parametry największej składowej spójnej:
- Rząd: 302
- Rozmiar: 350

- Wykreśl największą składową spójną i skomentuj wynik

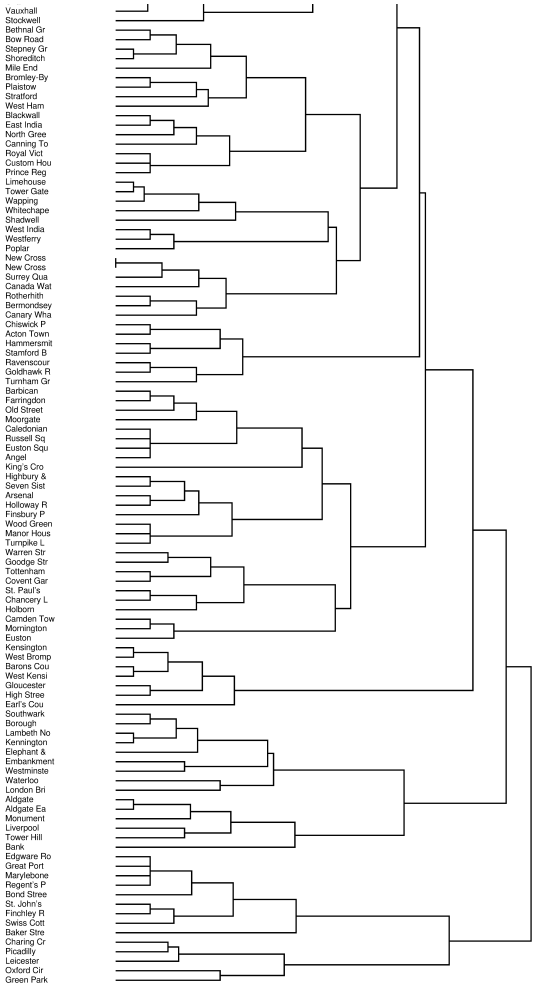


Rysunek 1: Wyznaczona składowa spójna.

Rysunek 1 przedstawia największą składową spójną. Początkowy graf (bezpośrednio wczytany z pliku) wyglądał tak samo, co potwierdza, że wczytany graf jest spójny. Na początku opisu **Zadanie A** podałem źródło dany oraz ich znaczenie. Graf przedstawia linie metra londyńskiego. Londyńskie metro, podobnie jak w innych miastach tworzy graf spójny. Gdy planujemy budowę metra chcemy umożliwić szybki transport z nowego połączenia do dowolnego innego miejsca. Wynikiem tego jest dużo przecięć linii metra. Najwięcej przecięć znajduje się bliżej centrum miasta, jest to zobrazowane na rysunku 1.

- Przeprowadź grupowanie metodą Warda z metryką d1 (odległość dwóch węzłów to liczba sąsiadów połączonych tylko z jednym z nich)

Realizacja:
Aby wykonać grupowanie metodą Warda, stworzyłem najpierw kluster (**Create Complete Cluster**). Należało wykonać tą operację, aby dissimilarities zostały wyznaczone względem wszystkich jednostek. Następnie wykonałem **Operations → Network + Cluster → Dissimilarity* → Network based → d1 → All**. W ten sposób policzone zostały dissimilarities i wygenerowany dendrogram (rys. 2), który zapisałem w formacie .eps.



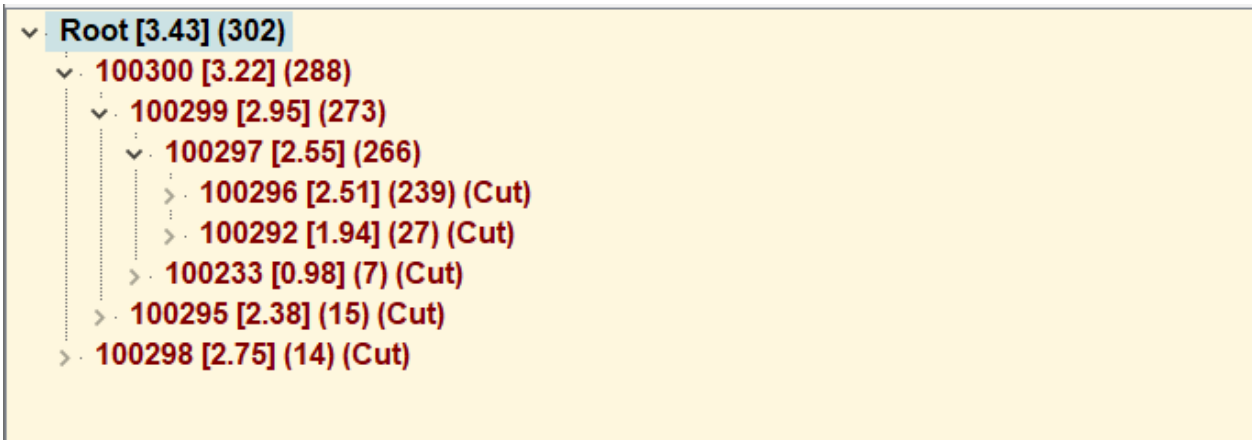
Rysunek 2: Fragment dendrogramu, przedstawiający wynik grupowania hierarchicznego.

- Wykreśl dendrogram i zaproponuj cięcie

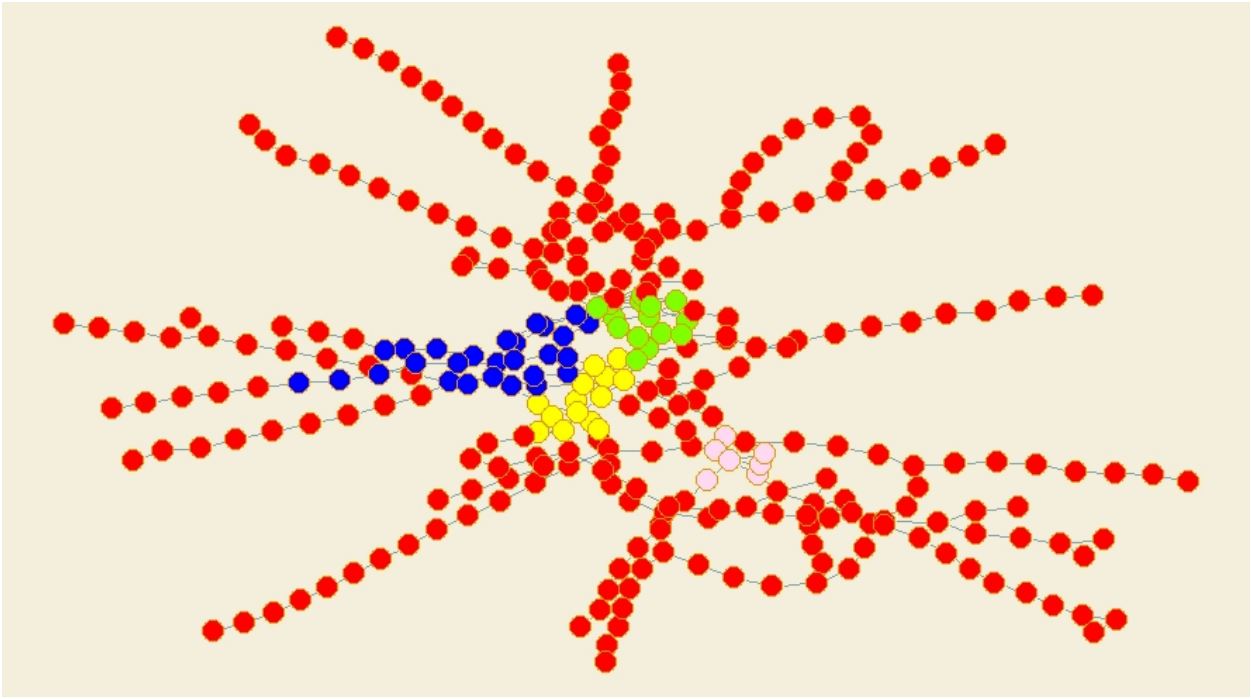
Dendrogram pokazany jest na rysunku 2. Rysunek 3 również pokazuje grupowanie, ale pokazuje tylko grupy, które zdecydowałem się utworzyć. Wykonałem cięcia przedstawione na rys. 3.

- Wykreśl wyodrębnione grupy

Wyodrębnione grupy ilustrują rysunki 4. oraz 5. Rysunek 4 jest odpowiednikiem rysunku 1, ale z naniesionymi kolorami, które wyróżniają 5 grup zgodnie z wykonanymi cięciami przedstawionymi na rysunku 3. Rysunek 5 przedstawia graf w innej postaci. Grupy zostały rozmieszczone w oddzielnych skupiskach. Dodatkowo są wyodrębnione za pomocą kolorów.



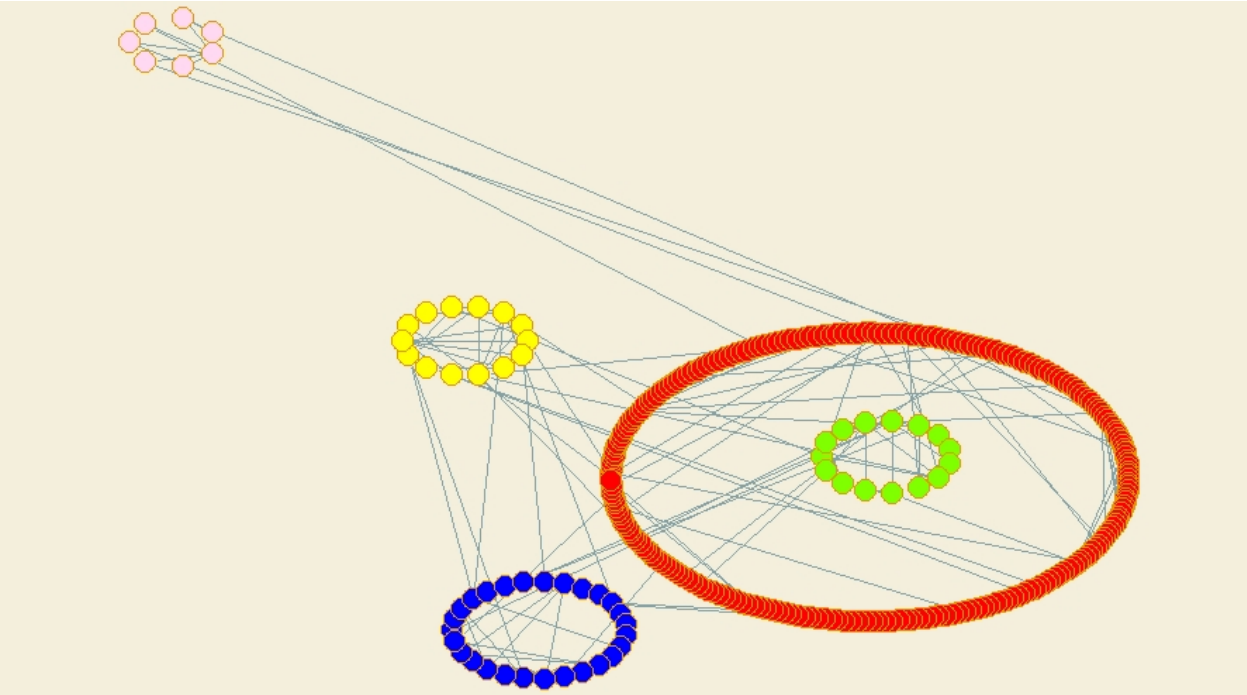
Rysunek 3: Wykonane cięcia.



Rysunek 4: Graf z oznaczonymi poprzez różne kolory grupami.

Grupowanie hierarchiczne ma liczne zalety. Między innymi samemu możemy zdecydować ile grup utworzymy. Płaskie grupowanie dostarcza nam gotowy wynik, którego nie możemy w prosty sposób zmodyfikować. Natomiast metoda hierarchiczna pozwala samodzielnie ustalić jakie grupy chcemy utworzyć. Cięcia jakie wykonałem wyodrębniło 5 grup, z których 4 są bardziej skupione (dużo połączeń wewnątrz grupy i niedaleko rozmieszczone od siebie), ale posiadają też połączenia z innymi grupami. Piąta grupa charakteryzuje się tym,

że duża część jej wierzchołków jest stopnia drugiego i posiada połączenia niemalże tylko wewnątrz swojej grupy. Występują nieliczne połączenia z pozostałymi grupami). Widoczne jest to na rysunku 4.



Rysunek 5: Graf podzielony na grupy. Odseparowane grupy.

2 Zadanie B - analiza grafu za pomocą biblioteki Networkx

Dane 2 - Sieć połączeń lotniczych

- Zbadaj jaki jest rząd i rozmiar całej sieci: pierwotnej oraz po usunięciu pętli i duplikatów krawędzi

Analiza pierwotnej sieci:
rząd: 3425
rozmiar: 19257

Parametry grafu po usunięciu pętli i podwójnych krawędzi:
rząd: 3425
rozmiar: 19256

- Wyodrębnij największą składową spójną, zbadaj jej rząd i rozmiar

Największy spójny komponent
rząd: 3397
rozmiar: 19230

Jak widać graf nie był spójny. Wyznaczona największa spójna ma nieco mniej krawędzi i wierzchołków niż graf po wstępnym przetworzeniu (usunięciu dubli, pętli).

- Wyznacz aproksymacje średniej długości ścieżki, operując na próbie losowej 100, 1000 i 10 tys. par wierzchołków

Aproksymacja średniej długości ścieżki
Średnia długość ścieżki została przybliżona poprzez wylosowanie ze zwracaniem (ponieważ nie dałoby się wylosować 10 000. par wierzchołków z grafu o rzędzie 3397 - graf ma za mało wierzchołków) 100, 1000 i 10 tys. par wierzchołków i wyznaczenie dla otrzymanych prób średniej minimalnej ścieżki. Dla każdej próby otrzymałem wartości średnią długość około 4,1.

- Wyznacz liczbę rdzeni o największym możliwym rzędzie, o drugim możliwie największym rzędzie o trzecim możliwie największym rzędzie; jakie to są rzędy?

Badanie rdzeni
Trzy rdzenie o największych rzędach to rdzenie rzędu 31, 30 i 29. Oznacza to, że minimalny stopień w podgrafach (wyznaczonych rdzeniach) miał odpowiednio wartość 31, 30 i 29. Znalezione zostały tylko po jednym rdzeniu każdego z podanych rzędów.

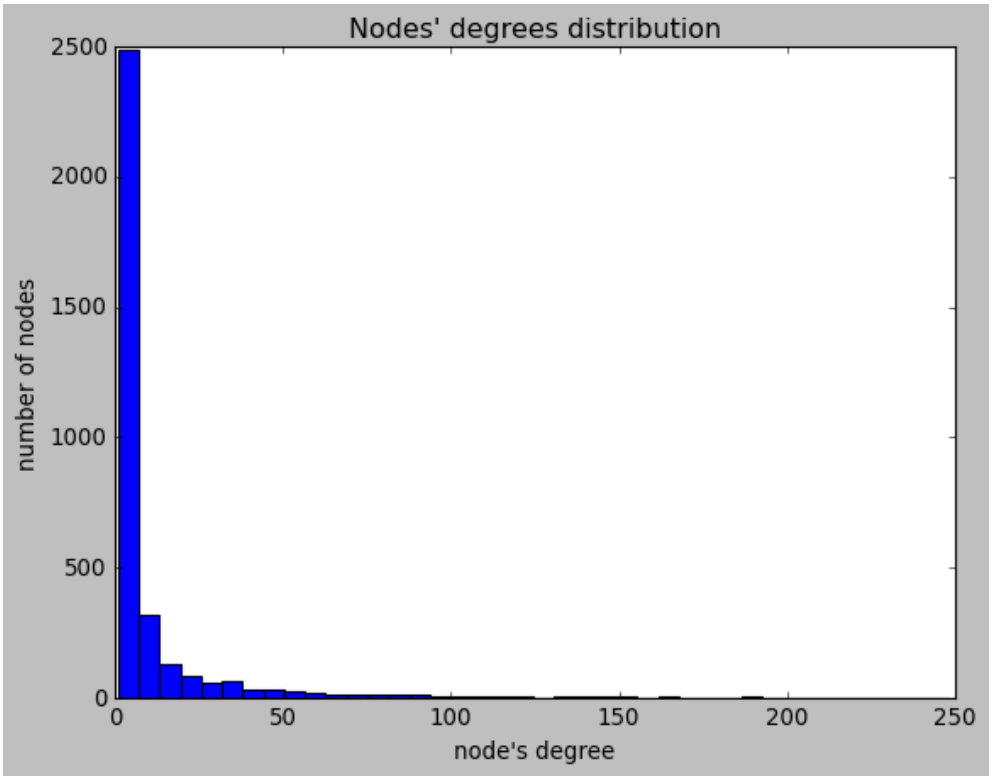
Rdzeń rzędu 31:
93 wierzchołki i 2257 krawędzi
liczba rdzeni: 1

Rdzeń rzędu 30:
138 wierzchołki i 3410 krawędzi
liczba rdzeni: 1

Rdzeń rzędu 29:
149 wierzchołki i 3728 krawędzi
liczba rdzeni: 1

- Wykreśl rozkład stopni wierzchołków

Rozkład stopni wierzchołków
Rysunek 6 przedstawia histogram rozkładu stopni wierzchołków. Jest to najprostszy histogram, jedynie z pogrupowaniem wierzchołków w liniowe zakresy liczby stopni (aby był nieco bardziej czytelny). Natomiast na rysunku 7 pokazany jest rozkład w osiach logarytmicznych. Drugi wykres jest nieco bardziej czytelny. Widać, że liczba wierzchołków o danym stopniu może mieć rozkład potęgowy - przynajmniej początkowy fragment (dla zakresu wierzchołków o stopniu 1-10).



Rysunek 6: Histogram przedstawiający rozkład wierzchołków

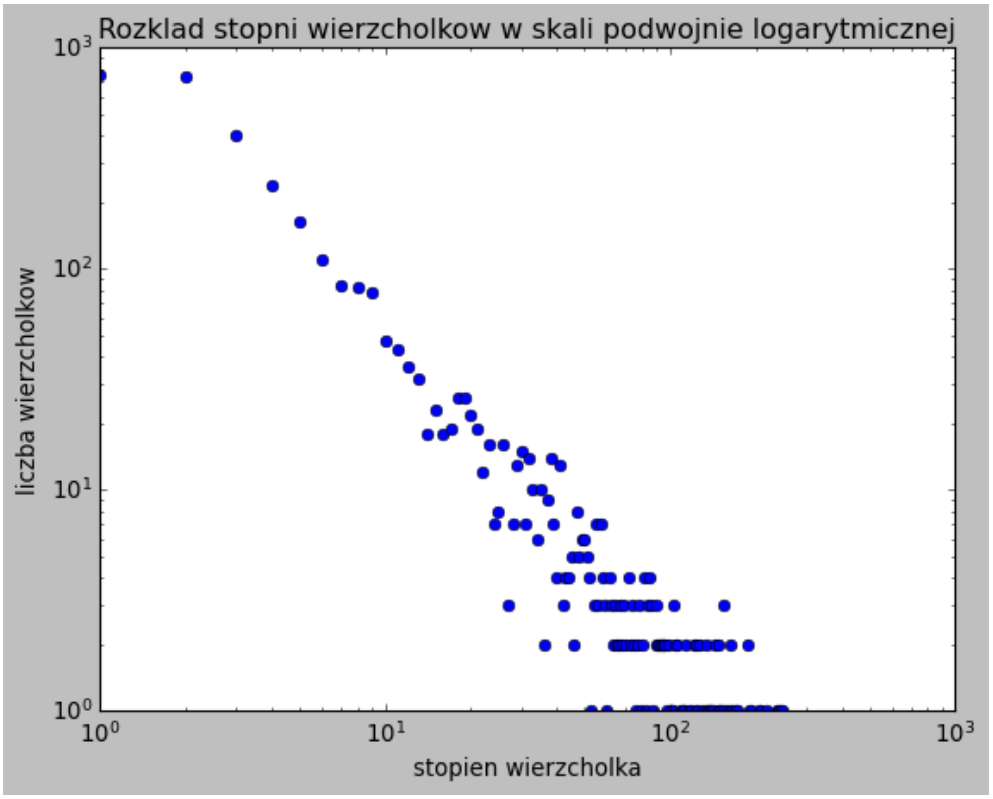
- Wyznacz wykładnik rozkładu potęgowego metodą regresji dla dopełnienia dystrybuanty rozkładu stopni, dla przedziałów rozlokowanych logarytmicznie

Rozwiązanie
Obliczone parametry regresji liniowej:
$y(x) = x \cdot (-1.1179759478312457) + 3.6588427106421406$
$a_0 = 3.6588427106421406$
$a_1 = -1.1179759478312457$

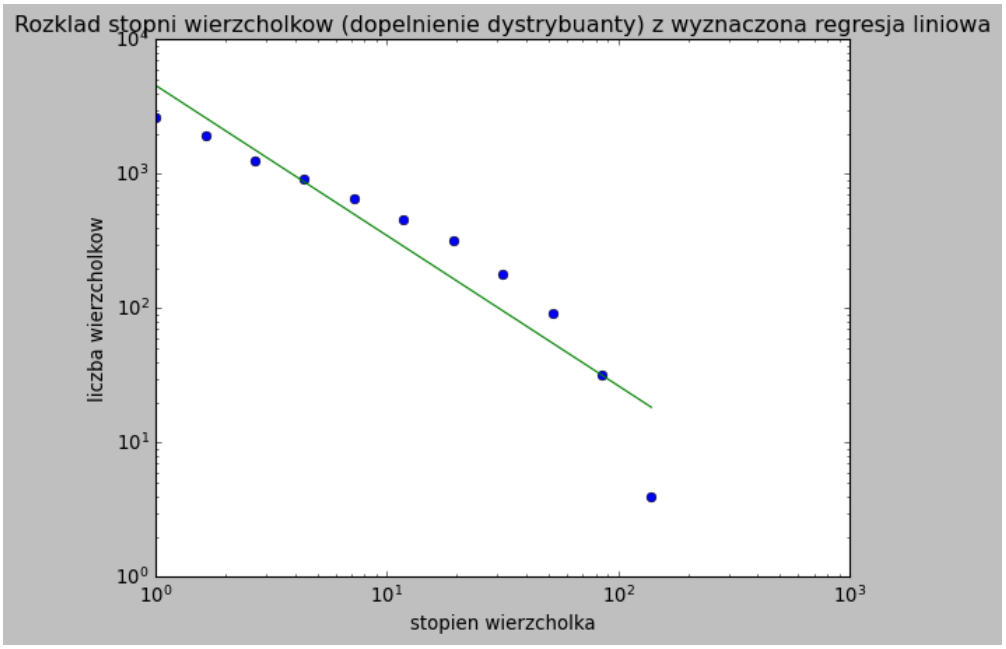
Na rysunku 8 widać wyznaczony rozkład potęgowy oraz dopasowaną prostą (za pomocą regresji liniowej). Widać, prosta regresji dość dobrze opisuje prawie wszystkie punkty wykresu. Jedynie ostatni jest stosunkowo daleki od wyznaczonej prostej.

- Wyznacz wykres Hilla

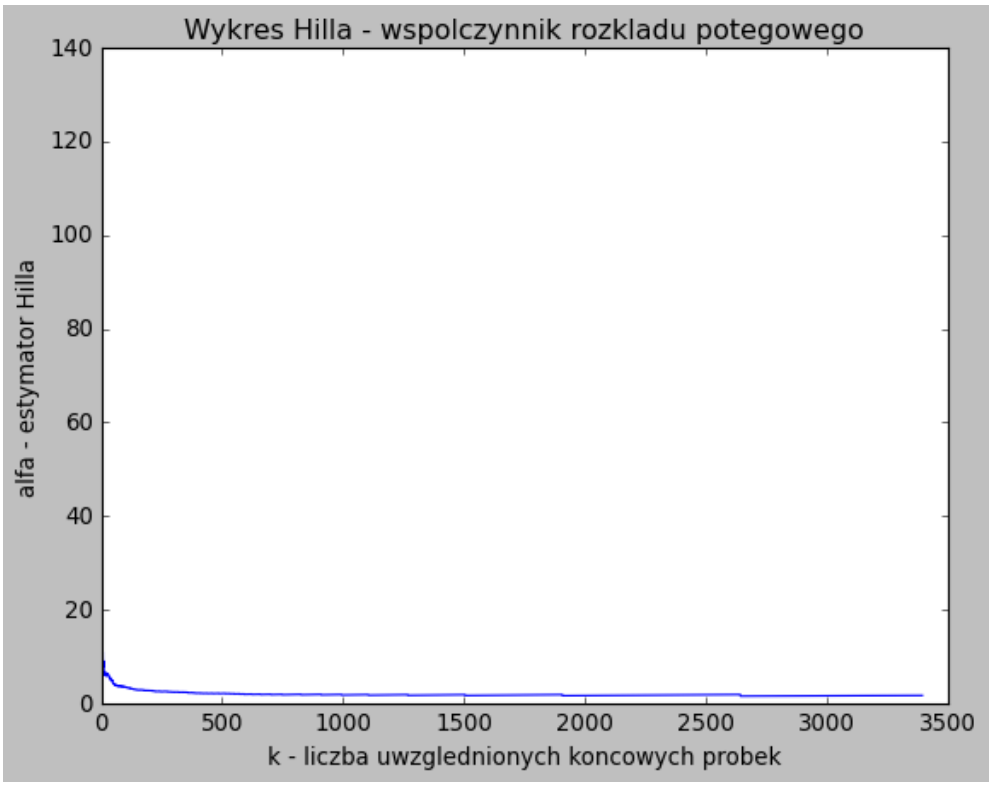
Rozwiązanie
Wykres Hilla to zależność estymatora Hilla od liczby uzwględnionych końcowych próbek. Estymator Hilla oszacowuje wartość parametru α w rozkładzie potęgowym rozkładu wierzchołków $f(d) = d^{-\alpha}$. Funkcja f(d) mówi ile jest wierzchołków o danym stopniu d w analizowanym grafie. Ponizej przedstawiony jest wzór na



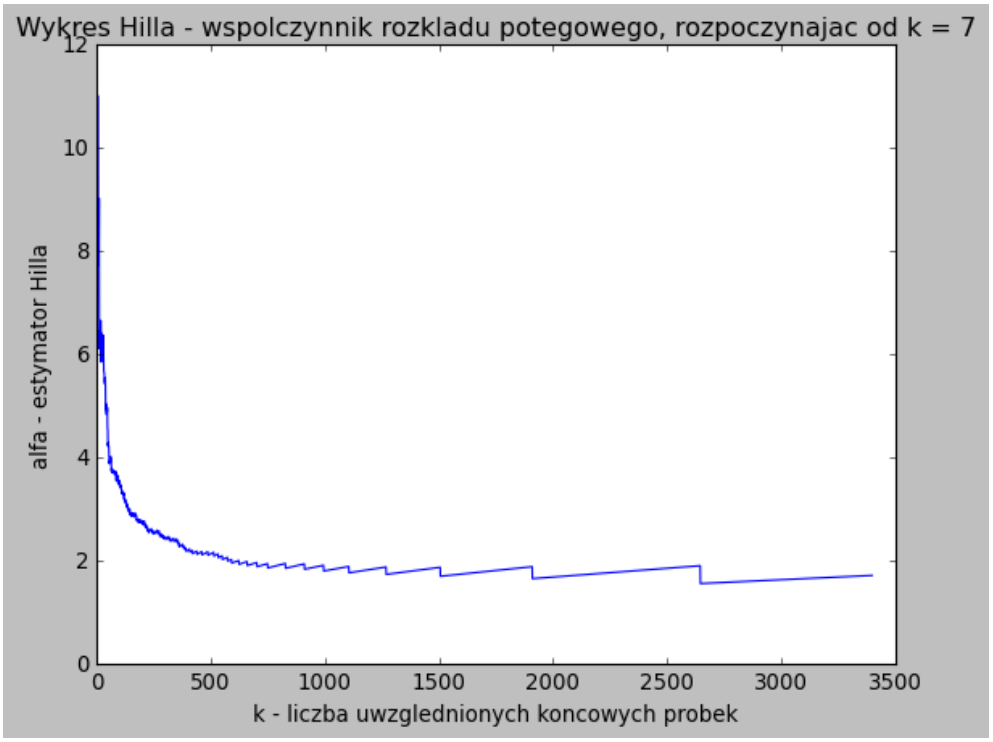
Rysunek 7: Wykres przedstawia rozkład wierzchołków w skali podwójnie logarytmicznej w zależności od ich stopnia.



Rysunek 8: Wykres rozkładu potęgowego dla dopełnienia dystrybuanty rozkładu stopni, przedziały rozklokowane logarytmicznie. Zielona linia przedstawia regresję liniową rozkładu.



Rysunek 9: Wykres Hilla (cała dziedzina).



Rysunek 10: Wykres Hilla (dla $k > 6$).

estymator Hilla.

$$\hat{a} = 1 + \hat{y}_k$$
$$\hat{y}_k = \frac{1}{k} \cdot \sum_{i=0}^{k-1} ln \frac{d_{N-i}}{d_{N-k}}$$

Pokazałem dwa wykresy Hilla, pierwszy ilustruje całą zależność (dla wszystkich k), drugi uwzględnia wartości od k = 7, aby lepiej widoczny był kształt wykresu, pomijając pierwsze elementy (rys. 9 i 10). Po analizie wykresu Hilla można zauważyć, że wartość estymatora stabilizuje się, od pewnego k (około k=500). Warty uwagi jest też kształt schodkowy, a skok zaczyna się powiększać wraz ze wzrostem k. Dla k około 3000 skok nie jest jeszcze znacząco duży i można by przyjąć wyestymowany współczynnik około $\alpha = 2$.

3 Dodatek

Repozytorium: https://github.com/MatFil99/TASS_p1