Wyszukiwanie i Przetwarzanie Informacji WWW

Własności Grafu WWW

Marcin Sydow

PJWSTK

Plan dzisiejszego wykładu:

- graf WWW
- rola analizy linków i jej zastosowania
- rozkład potęgowy
- rozkłady potęgowe w grafie WWW
- Zipf, Pareto i związki z potęgowym
- spójność grafu WWW
- fraktalność grafu WWW
- zjawiska społeczne a graf WWW
- macierz sąsiedztwa grafu WWW
- Podsumowanie Wykładu

Szczególne własności WWW

"The World Wide Web is the only thing I know of whose shortened form takes three times longer to say than its long form."

- Douglas Adams, The Independent on Sunday, 1999

(na razie nie ma polskiej nazwy WWW)

Graf WWW

Definition

Przez graf WWW pewnej kolekcji dokumentów hipertekstowych D rozumiemy skierowany graf G(V,E), gdzie każdy wierzchołek $v \in V$ odpowiada dokumentowi $d \in D$ a skierowana krawędź $(p, q) \in E$ odpowiada hiperlinkowi z dokumentu $p \in D$ do dokumentu $q \in D$.

- usuwa się linki-pętle (postaci (p,p), $p \in D$),
- linki wielokrotne traktuje się pojedynczo.
- czasem nie uwzględnia się linków wewnątrz tego samego hosta, domeny, etc.
- czasami wierzchołkami grafu są **całe hosty** lub domeny.

Analiza Linków

Graf WWW okazał się w praktyce **bardzo użyteczną abstrakcją** WWW.

Dziedziną, która bada własności grafu WWW jest analiza linków WWW (ang. Link Analysis).

Jest to ważny dział Eksploracji sieci WWW (ang. Web Mining) o dużych zastosowaniach praktycznych m.in. w Wyszukiwaniu Informacji w WWW (ang. Web Information Retrieval)

Zastosowania Analizy Linków

Analiza grafu WWW (analiza linków) ma **bezpośrednie zastosowania** w:

Wyszukiwarkach Internetowych, np.:

- Ranking
- Wyszukiwanie dokumentów podobnych
- wykrywanie **chłamu** wyszukiwarkowego (ang. SE spam)

Eksploracji sieci WWW, np.:

- badanie "społecznych" aspektów WWW (ang. community mining)
- odkrywanie praw rządzących rozwojem i strukturą WWW

Pokrewne Zastosowania

Początki analizy linków związane są z Analizą Bibliograficzną (ang. Bibliographic Citation Analysis)

Obecnie, analiza linków jest też powiązana z takimi dziedzinami jak:

- analiza sieci "międzyludzkich" (ang. Social Network Analysis)
- Sieci Zaufania (ang. Trust Networks) - w tym np. systemy reputacyjne w aukcjach Internetowych
- analiza powiązań między pojęciami w ontologiach

Techniki, narzędzia i podejście okazują się być wspólne (lub podobne) dla powyższych dziedzin.

Większość z tych dziedzin ma bezpośredni związek z **rosnącą rolą Internetu**.

Rozmiar i dynamika

W przybliżeniu:

- ponad 11 500 000 000 indeksowalnych dokumentów (Gulli et al., 2005)
- wykładniczy wzrost
- czas połowicznej zmiany: 10 dni

1. "A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, S. Raghavan.", "Searching the Web," ACM Transactions on Internet Technology, 1(1), 2-43, 2001
2. "Gulli, A. and A. Signorini", "The indexable Web is more than 11.5 billion pages", Proceedings of the 14th International World Wide Web Conference. Special Interests, Tracks and Posters, 902-903, 2005

Stopnie wierzchołków

Definition

Stopień wyjściowy wierzchołka v w grafie skierowanym G(V,E):

$$outDeg(v) = |\{u \in V : (v,u) \in E\}|$$

Stopień wejściowy:

$$inDeg(v) = |\{u \in V : (u,v) \in E\}|$$

Interpretacja:
Ilość stron cytowanych przez i **cytujących**, odpowiednio.

Jaki jest rozkład stopni wyjściowych?

Ma to duże znaczenie m.in. dla:

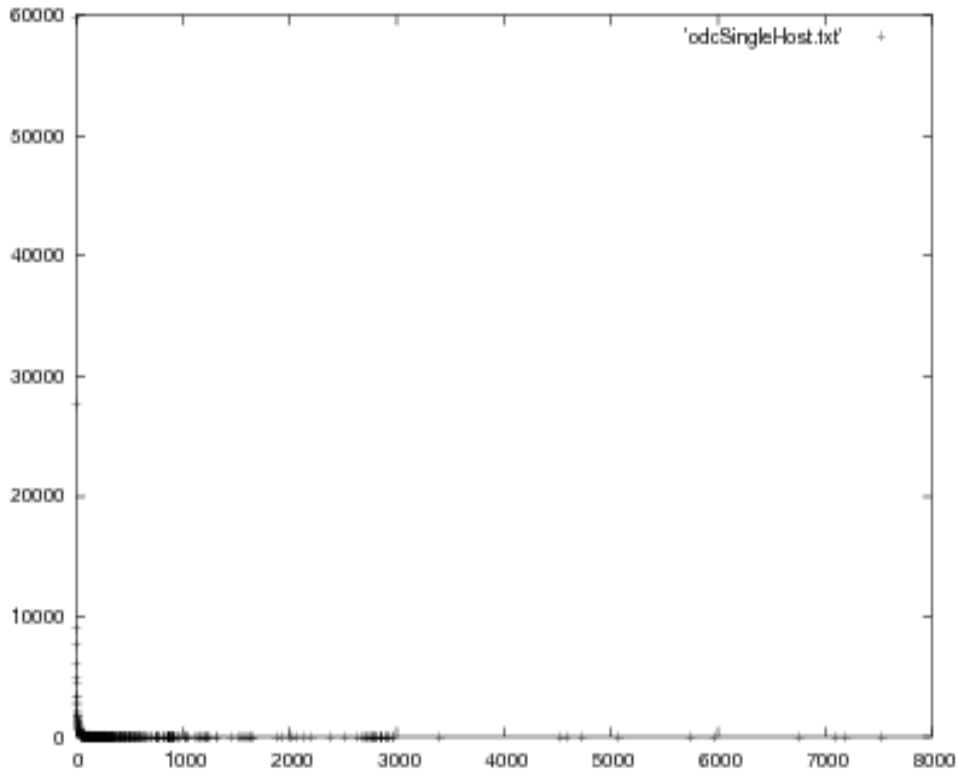
- przechowywania dużych grafów WWW (kompresja)
- obróbki danych na potrzeby analizy linków

Można się domyślić, że jest mało stron o wielu linkach i bardzo dużo stron o niewielu.

Rozkłady takie nazywamy rozkładami o **ciężkich ogonach** (ang. heavy-tailed)

Spróbujmy "zgadnąć" jaki jest to rozkład za pomocą wizualizacji...

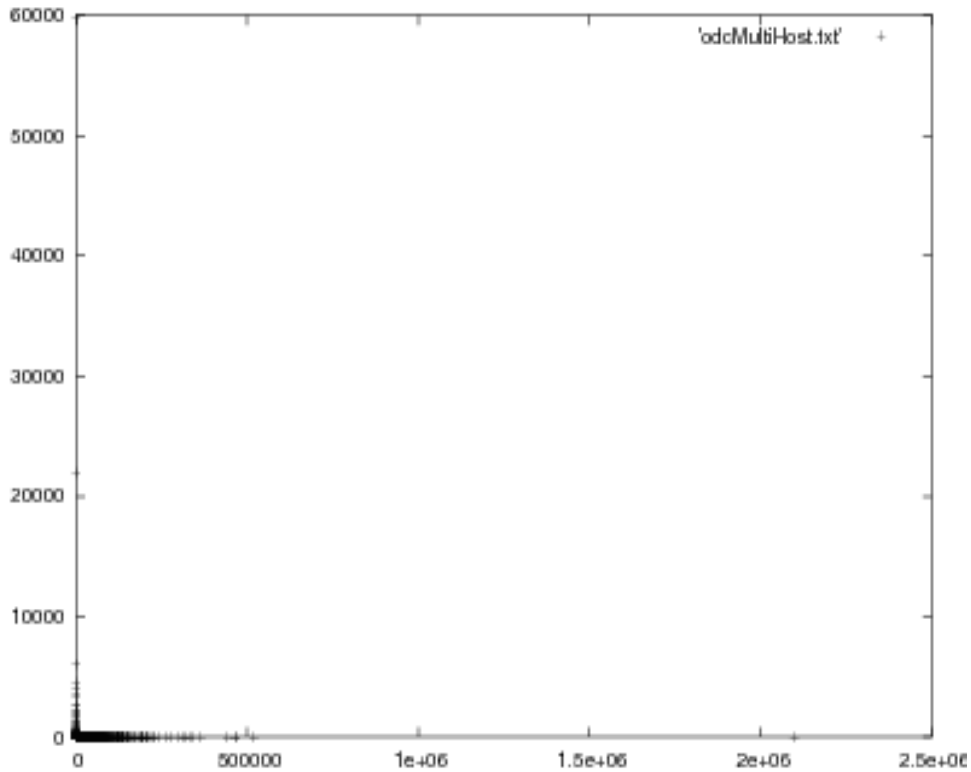
Zobaczmy na wykresie rozkład stopni wyjściowych dla grafu 167604 hostów grafu .pl zebranego w grudniu 2005:



Rysunek: Rozkład stopni wyjściowych w przykładowym grafie 167604 hostów grafu .pl zebranego w grudniu 2005

Rozkład jest **zbyt skośny** aby coś zauważyć na takim wykresie...

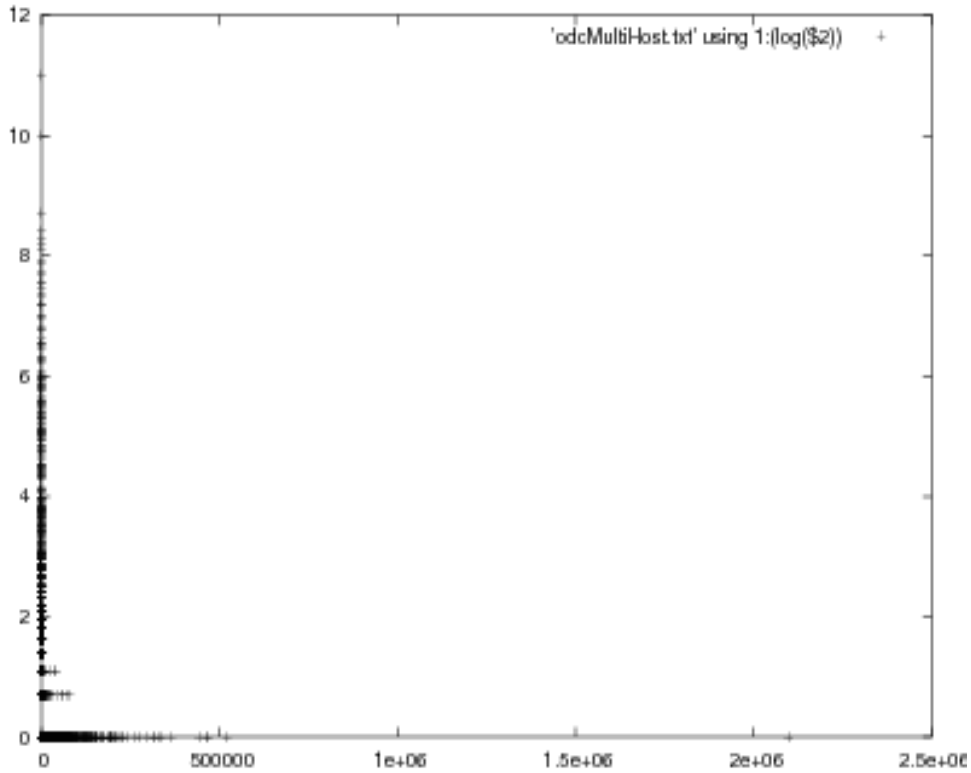
Może wyraźniej będzie jak uwzględnimy wielokrotne linki...



Rysunek: Rozkład stopni wyjściowych w grafie 167604 hostów domeny .pl (grudzień 2005) - uwzględniono wielokrotne linki

Nie jest lepiej...

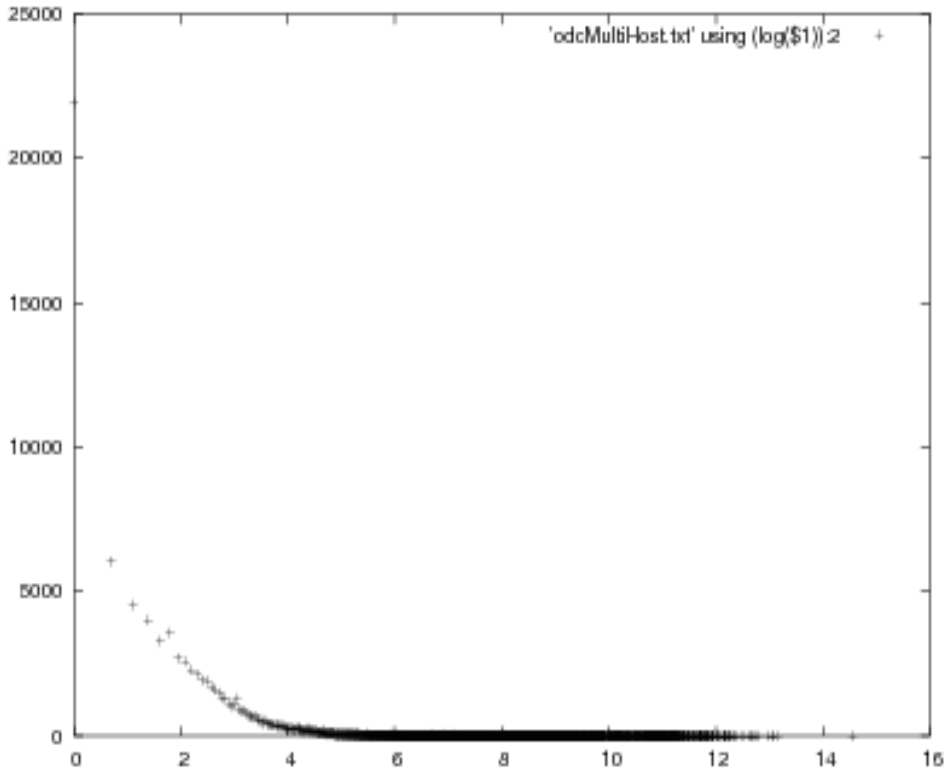
A może by zlogarytmować oś Y?



Rysunek: HostGraf .pl 2005

Wciąż słabo...

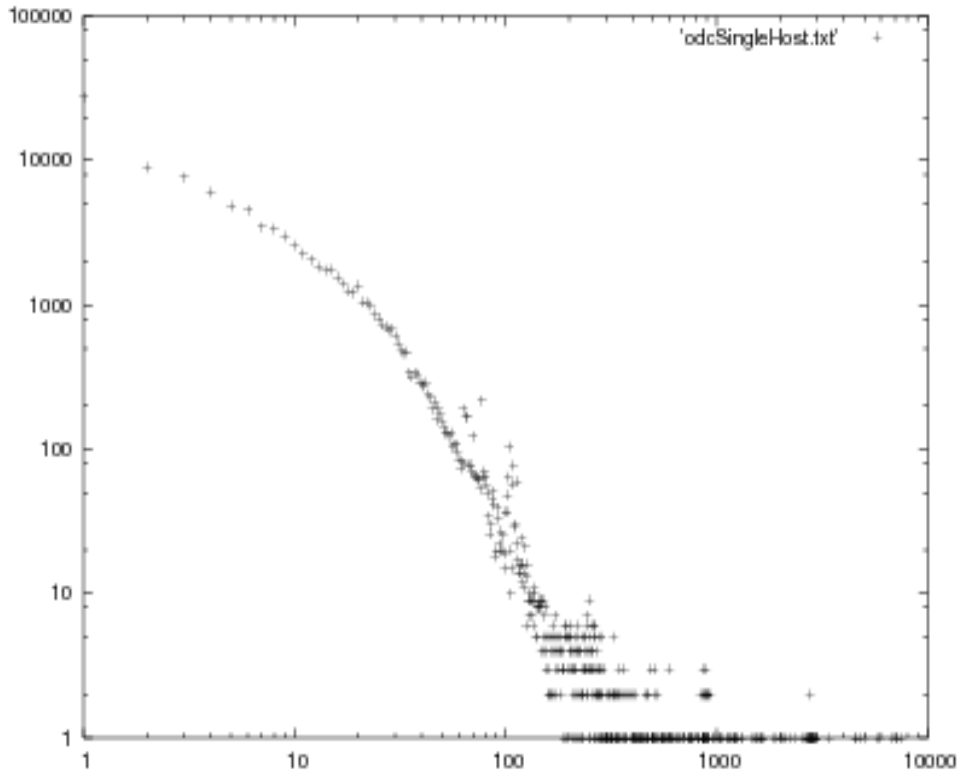
To może zlogarytmować oś X?



Rysunek: HostGraf .pl 2005

Coś się zaczyna dziać...

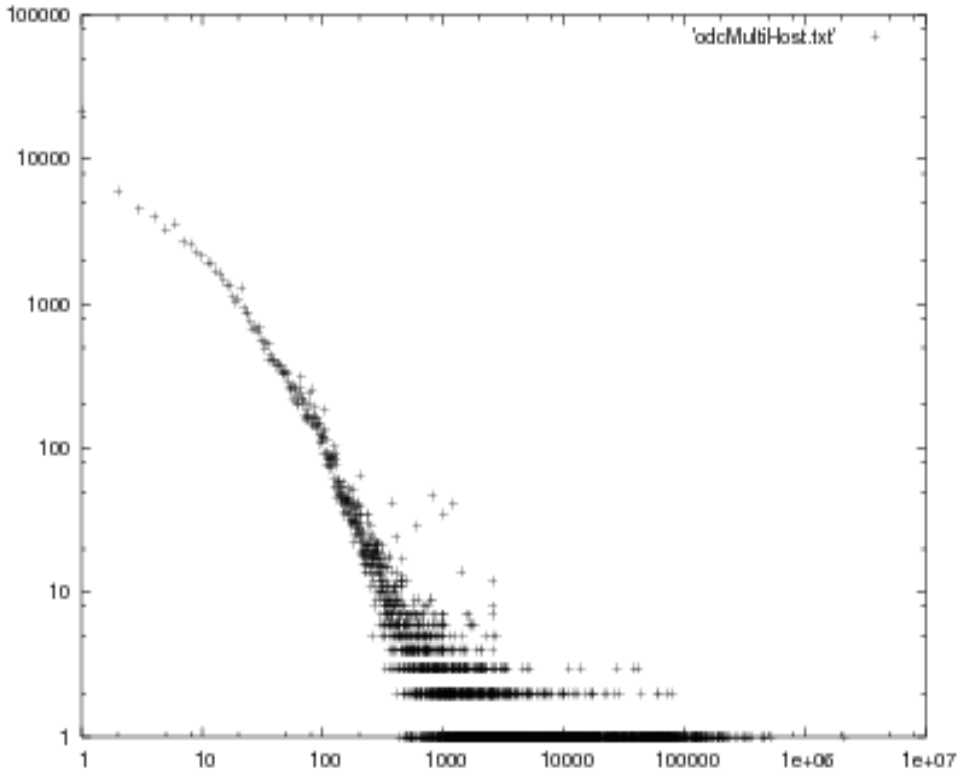
A może by tak zlogarytmować obie osie...



Rysunek: HostGraf .pl 2005

Jest różnica.

Podobnie dla linków wielokrotnych:



Rysunek: HostGraf .pl 2005

Po zlogarytmowaniu obu osi wykres gęstości przypomina linię prostą. Funkcja gęstości o takich własnościach odpowiadałaby tzw. *rozkładowi potęgowemu*

Rozkład Potęgowy (ang. Power law)

Linia prosta o ujemnym nachyleniu na wykresie o **zlogarytmowanych** osiach?

$$log(y) = log(c) - a \cdot log(x)$$

Równoważnie:

$$y = \frac{c}{x^a}$$

Rozkład Potęgowy

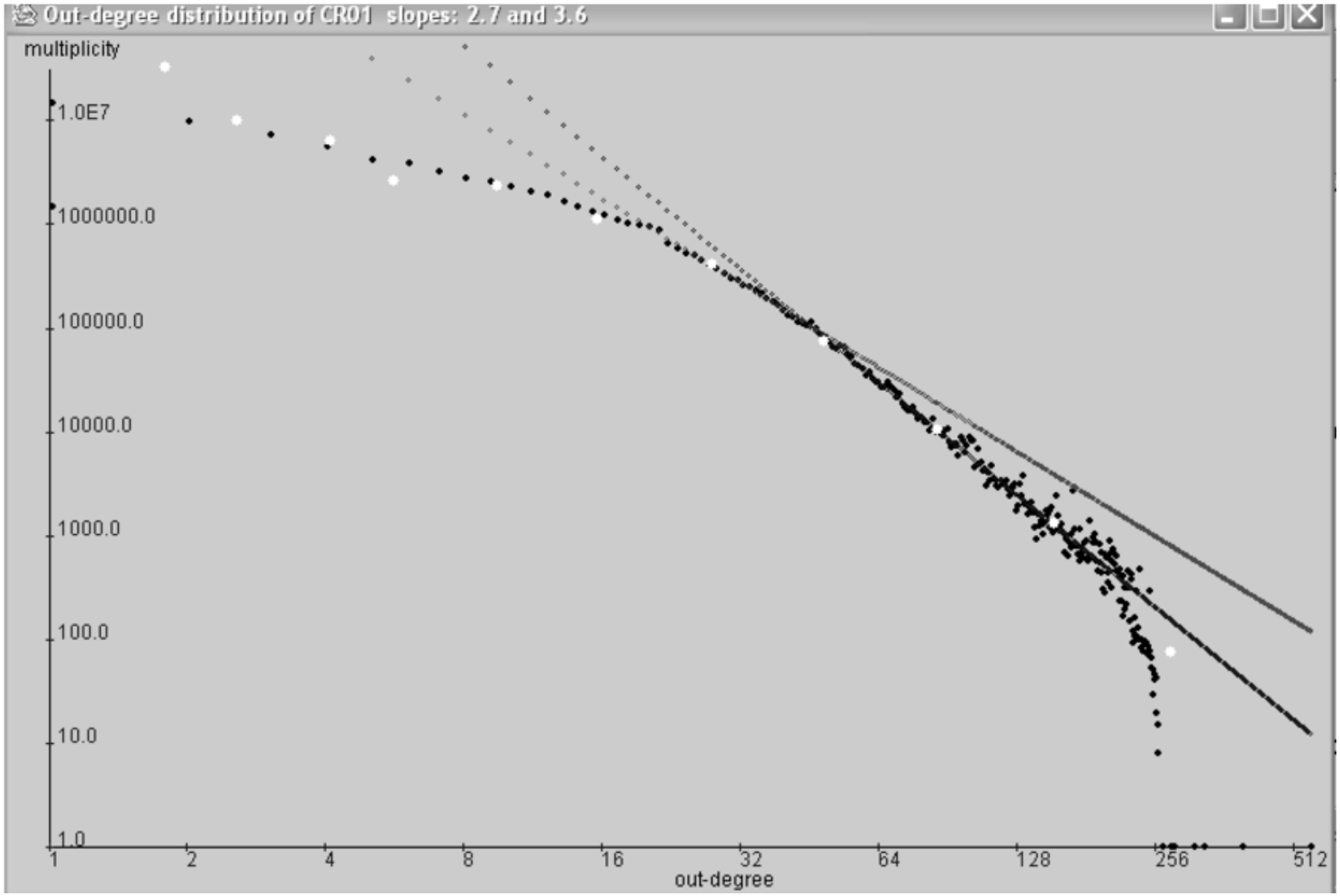
Powiemy, że rzeczywista zmienna losowa X ma rozkład potęgowy jeśli jej funkcja gęstości f dana jest wzorem:

$$f(k) = \frac{c}{k^{\gamma}},$$

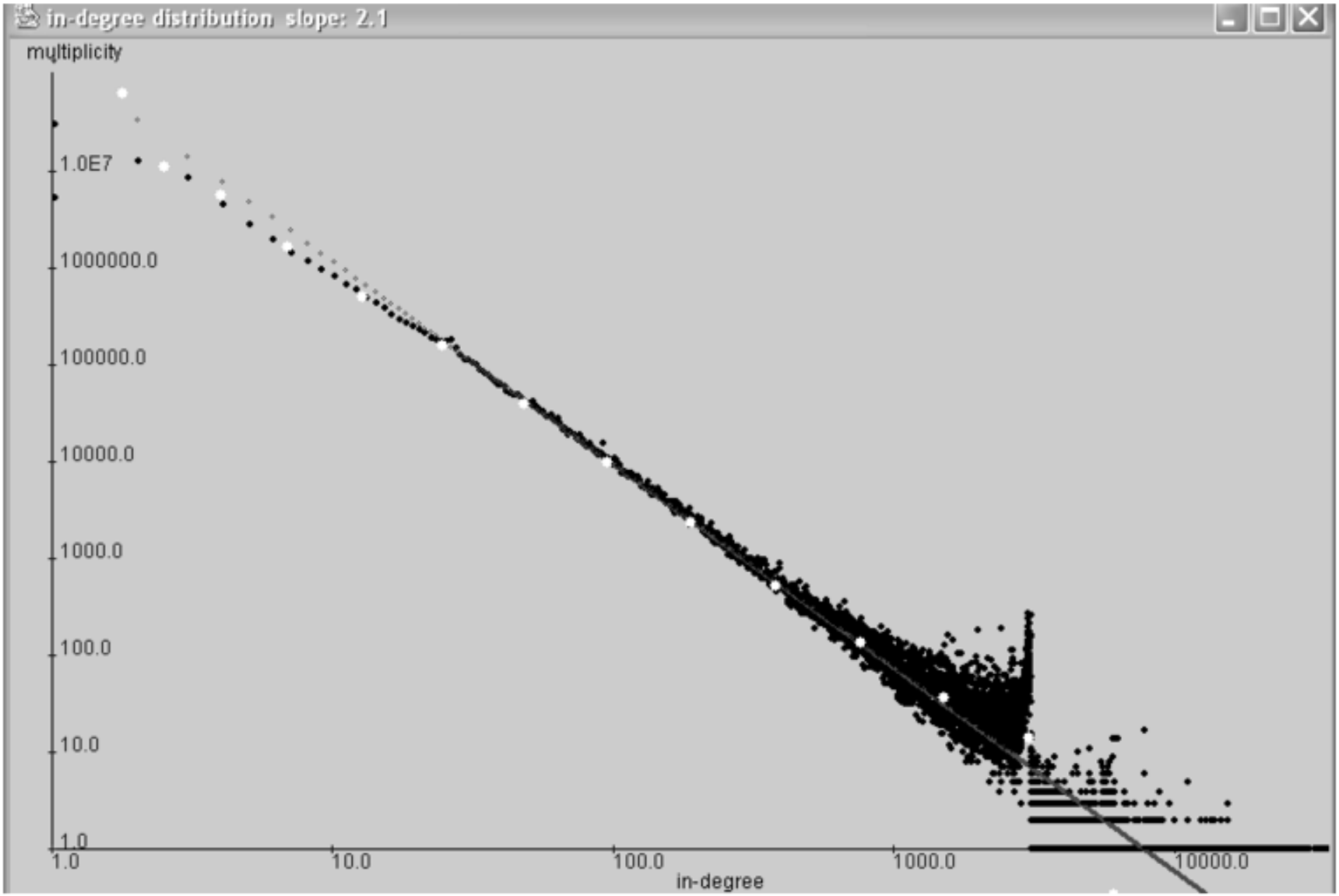
gdzie k jest dodatnią liczbą rzeczywistą, c jest stałym współczynnikiem proporcjonalności.
Parametr γ nazywamy *wykładnikiem* rozkładu.
Zauważmy, że funkcja f po zlogarytmowaniu obu osi wygląda jak linia prosta o ujemnym nachyleniu γ

Kształt wykresu dla grafu hostów .pl z grudnia 2005 to nie przypadek.

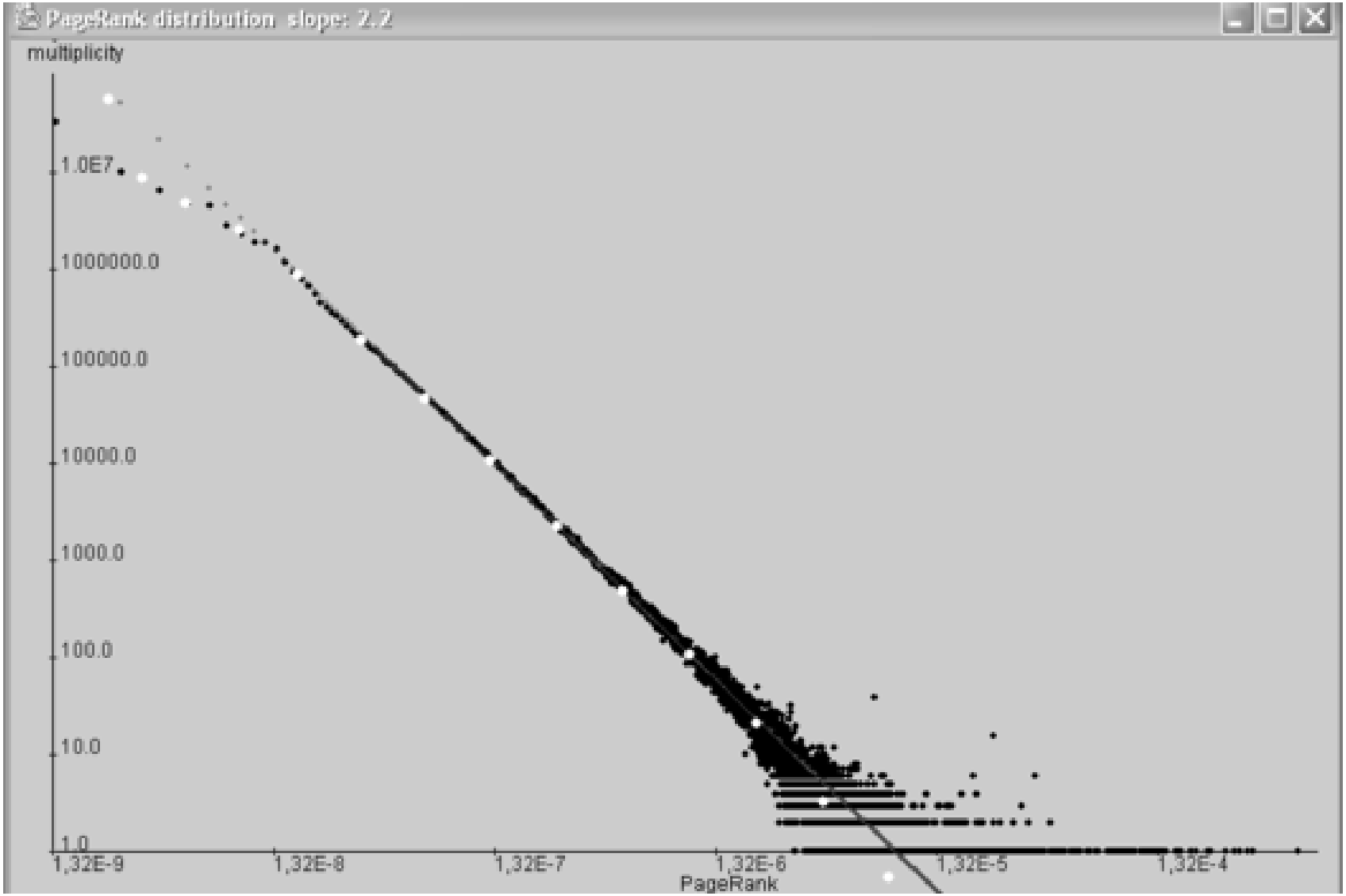
Okazuje się, że w WWW zadziwiająco wiele wielkości ma rozkład potęgowy



Rysunek: Rozkład stopni wyjściowych w przykładowym grafie 80 milionów dokumentów z amerykańskiego WWW (StanfordWebBase/2001)



Rysunek: Rozkład stopni wejściowych w przykładowym grafie 80 milionów dokumentów z amerykańskiego WWW (StanfordWebBase/2001)



Rysunek: Rozkład wartości PageRank, (decay = 0.1) w przykładowym grafie 80 milionów dokumentów z amerykańskiego WWW (StanfordWebBase/2001)

Znanych przykładów jest więcej:

- wielkości hostów lub domen
- aktywność adresów IP w zapytaniach do wyszukiwarek
- wielkości składowych spójnych w grafie WWW
- czestości występowania słów w dokumentach

Skąd taka regularność?

WWW jest dynamicznym tworem kilkuset milionów internautów i trudnej do oszacowania liczby automatów.

Jak jednak widać, WWW nie jest tworem chaotycznym - wręcz przeciwnie - statystycznie rządzą nim **silne ukryte prawa**.

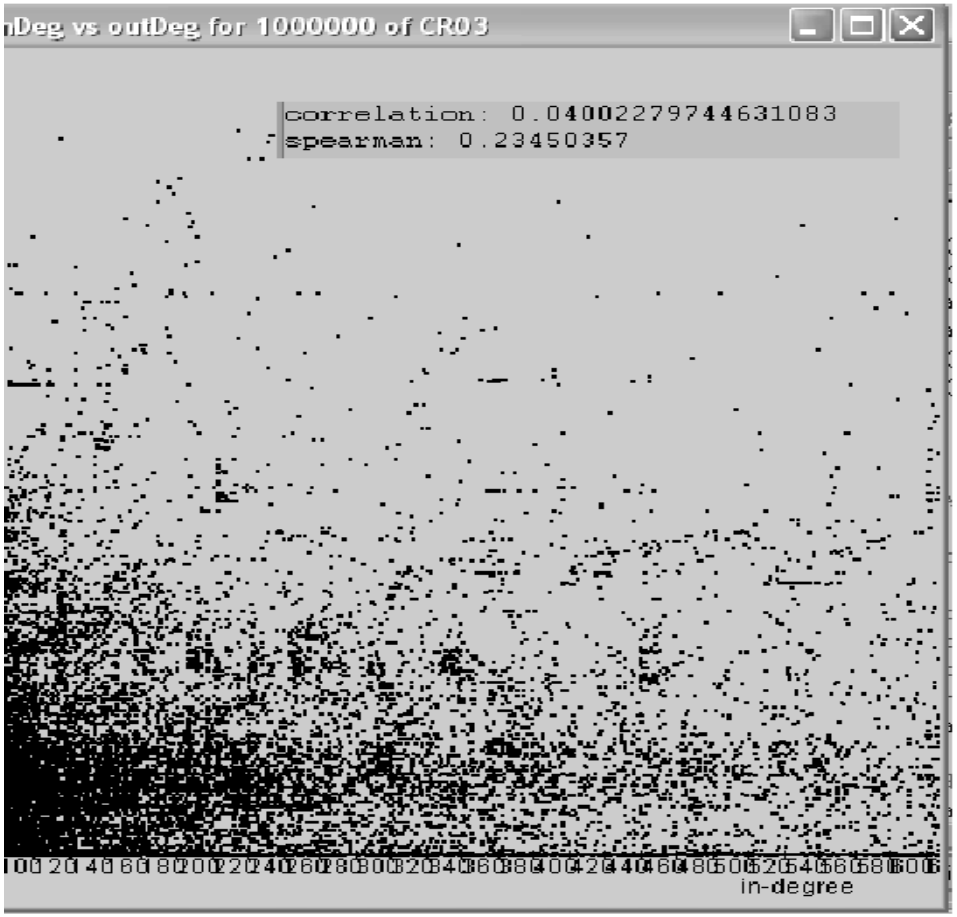
Rozkłady potęgowe obserwuje się także w naukach przyrodniczych i ekonomii w systemach o wykładniczym tempie przyrostu (np. wielkości miast).

Mimo ogromnej różnorodności i dynamiki WWW, wartość wykładnika w rozkładzie stopni wejściowych grafu WWW jest stała od lat i wynosi około 2. **Czy to przypadek**?

PageRank ma bardzo podobny rozkład.

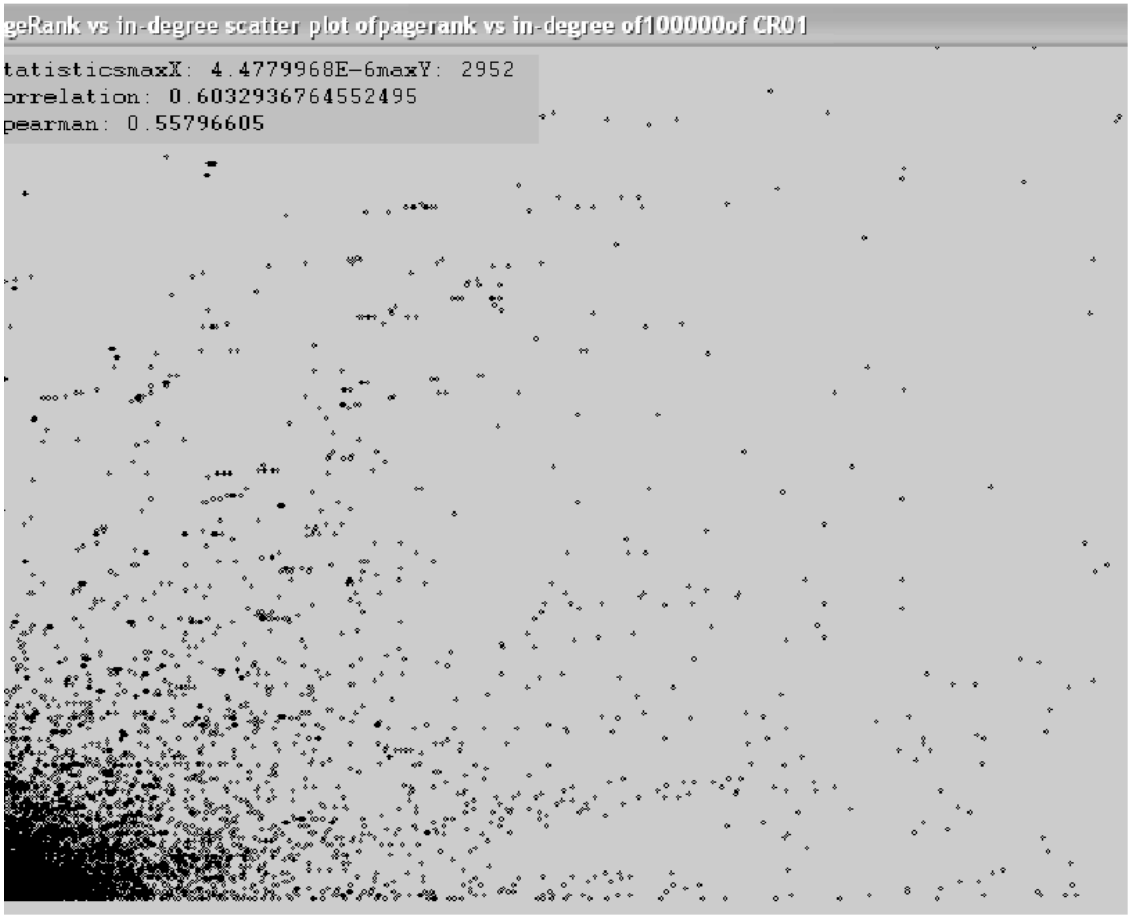
Zależności pomiędzy powyższymi wielkościami

OutDegree ma podobny rozkład do inDegree. Czy są skorelowane?



Rysunek: Zależność stopni wyjściowych i wejściowych w grafie 50 milionów dokumentów z amerykańskiego WWW z roku 2003 (pomiary na grafie Stanford

Zależność PageRank i in-degree



Rysunek: Zależność stopni wejściowych i wartości PageRank w grafie 80 milionów dokumentów z amerykańskiego WWW z roku 2001 (pomiary na grafie Stanford WebBase)

Rozkłady Potęgowe w Przyrodzie - Tekst

Prawo Zipfa:

1. weźmy dowolny (dostatecznie długi) tekst w języku naturalnym
2. policzmy częstości słów i posortujmy je nierosnąco.

Rozkłady Potęgowe w Przyrodzie - Tekst

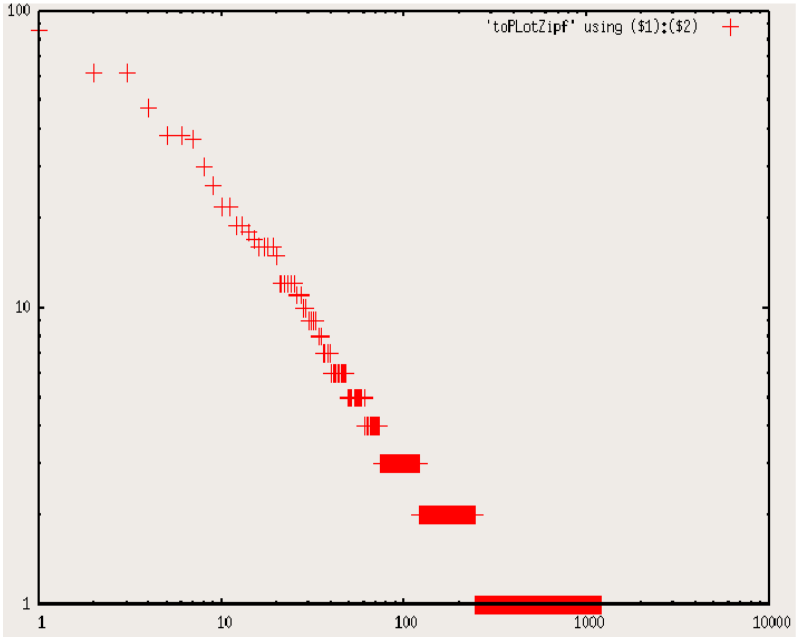
Prawo Zipfa:

1. weźmy dowolny (dostatecznie długi) tekst w języku naturalnym
2. policzmy częstości słów i posortujmy je nierosnąco. Co się okazuje?

Rozkłady Potęgowe w Przyrodzie - Tekst

Prawo Zipfa:

1. weźmy dowolny (dostatecznie długi) tekst w języku naturalnym
2. policzmy częstości słów i posortujmy je nierosnąco. Co się okazuje?
3. r-ta częstość wynosi mniej więcej c/r, gdzie c jest stałą!



Rysunek: Przykład: częstości wyrazów **tego wykładu**, uporządkowane nierosnąco (i osie zlogarytmowane). Prosta o ujemnym nachyleniu - **rozkład Zipfa**.

Rozkłady Potęgowe w Przyrodzie c.d

- rozmiar r-tego największego miasta w danym kraju (bez centralnego planowania) (też Zipf)
- wielkość zarobków r-tego najlepiej zarabiającego pracownika w populacji (Pareto)
- ilość odwiedzin strony WWW w danej domenie (hoście) (Potęgowy)
- ilość linków cytujących r-tą najpopularniejszą stronę (Potęgowy)

Wielość nazw

Pojawia się **pozornie** wiele podobnych rozkładów:

- Zipf (wielkość r-tej wartości): $y \approx r^{-b}$
- Pareto (oryginalnie: "ilu ludzi zarabia więcej niż x?") $P(X > x) \approx x^{-k}$
- Rozkład potęgowy: $P(X = x) \approx x^{-a}$

Jakie są związki między tymi trzema rozkładami?

Potęgowy $\leftrightarrow$ Pareto

Zbadajmy związek pomiędzy rozkładem Potęgowym a Pareto:

Pareto z wykładnikiem k:
$P(X > x) = (\frac{m}{x})^k$, dla m,k$>$ 0, $x >= m$, m - minimalne zarobki

Wobec tego **dystrubuantą** tego rozkładu jest:
$F(x) = P(X \leq x) = 1 - (\frac{m}{x})^k$

Funkcja **gęstości**[1] tego rozkładu jest **różniczką**[2] dystrybuanty:
$p_X(x) = k \cdot m^k \cdot x^{-(k+1)}$

Jak widać, odpowiada to rozkładowi potęgowemu z wykładnikiem $k + 1$.

[1] lub funkcja prawdopodobieństwa - dla rozkładu dyskretnego
[2] w rozkładzie dyskretnym odpowiada temu operator **różnicowy**

Zipf <-> Pareto

Zbadajmy teraz związek rozkładu Zipfa z rozkładem Pareto:

Zipf: “r-ta co do wielkości wartość ma wielkość n”

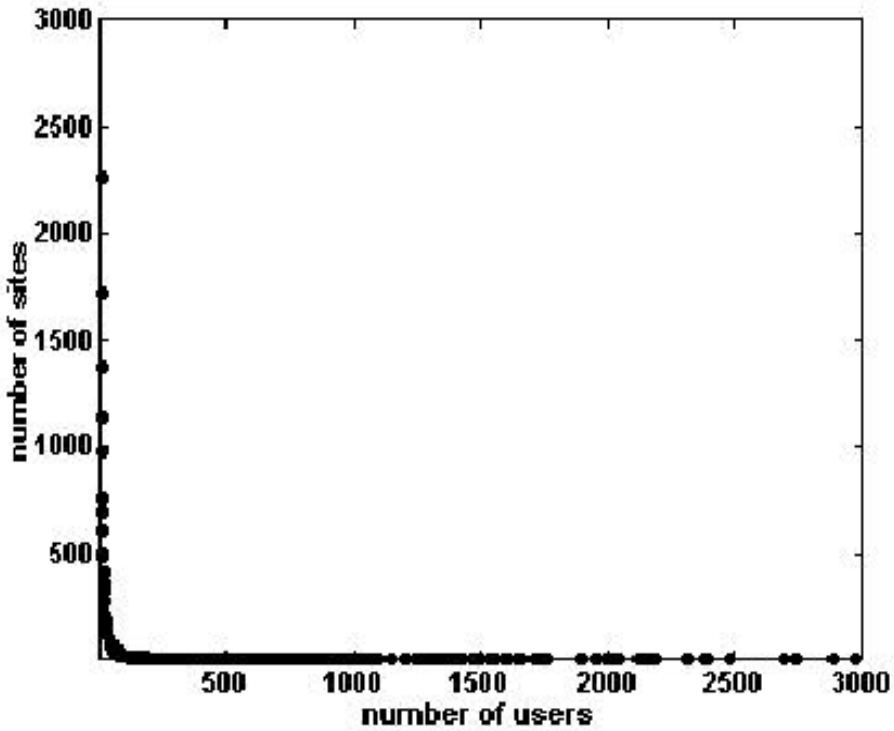
jest równoważne powiedzeniu: “r wartości jest niemniejszych niż n”

Wystarczy więc **odwrócić** znaczenie zmiennych r i n aby otrzymać rozkład Pareto:

(Zipf:) $n \approx r^{-b} \Leftrightarrow r \approx n^{-1/b}$ (Pareto)

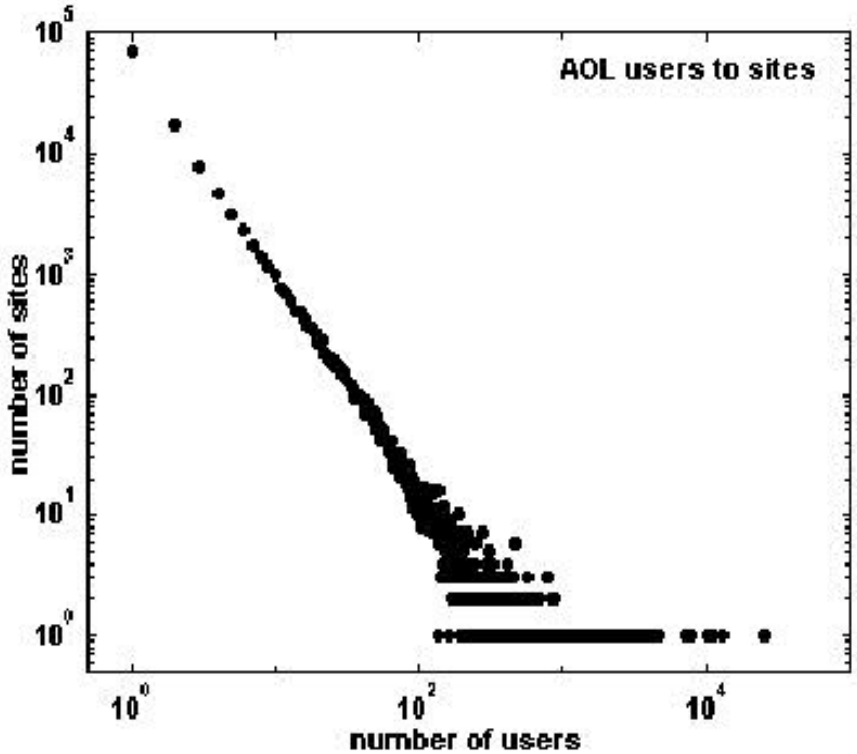
Przykład - AOL

Dla ilustracji powyższych zależności pomiędzy rozkładami zobaczmy ilość odwiedzin stron na serwisie AOL.



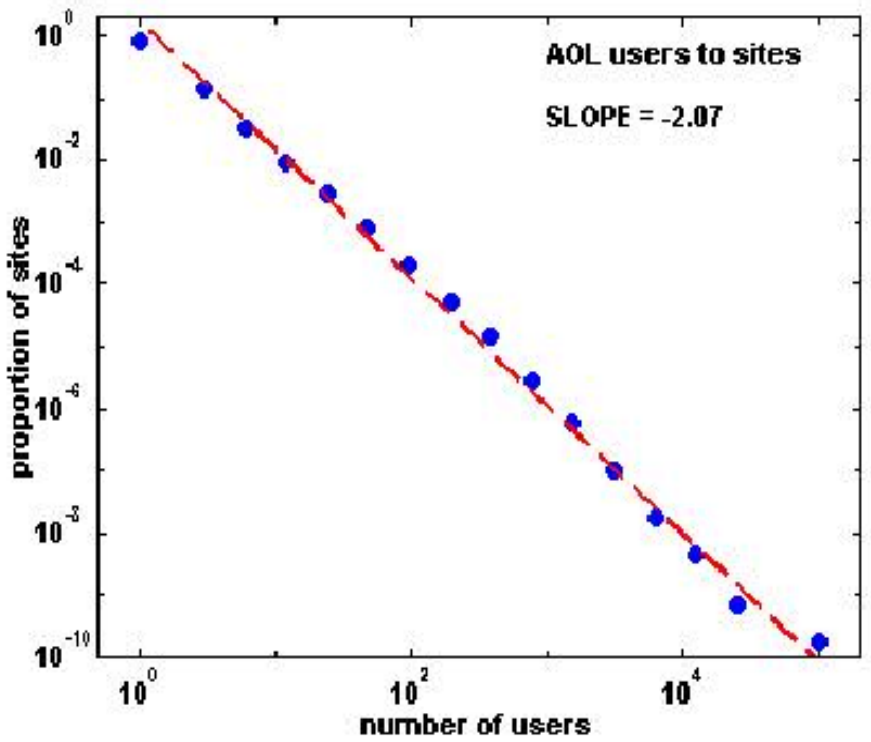
Rysunek: Rozkład odwiedzin użytkowników na poszczególnych stronach AOL (skale liniowe) (źródło: L.Adamic "Zipf, Power Laws and Pareto - a ranking tutorial")

Przykład - AOL, po zlogarytmowaniu osi (Rozkład Potęgowy)



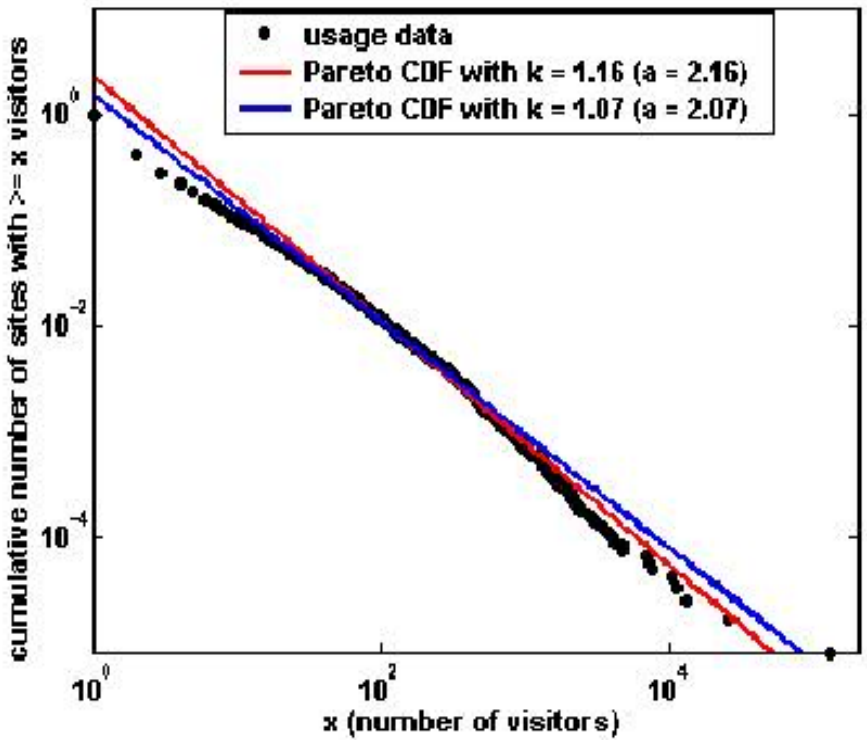
Rysunek: Rozkład odwiedzin użytkowników na poszczególnych stronach AOL (skale logarytmiczne) (źródło: L.Adamic "Zipf, Power Laws and Pareto - a ranking tutorial")

Przykład - AOL, po użyciu wykładniczych "koszyków" (ang. bins)



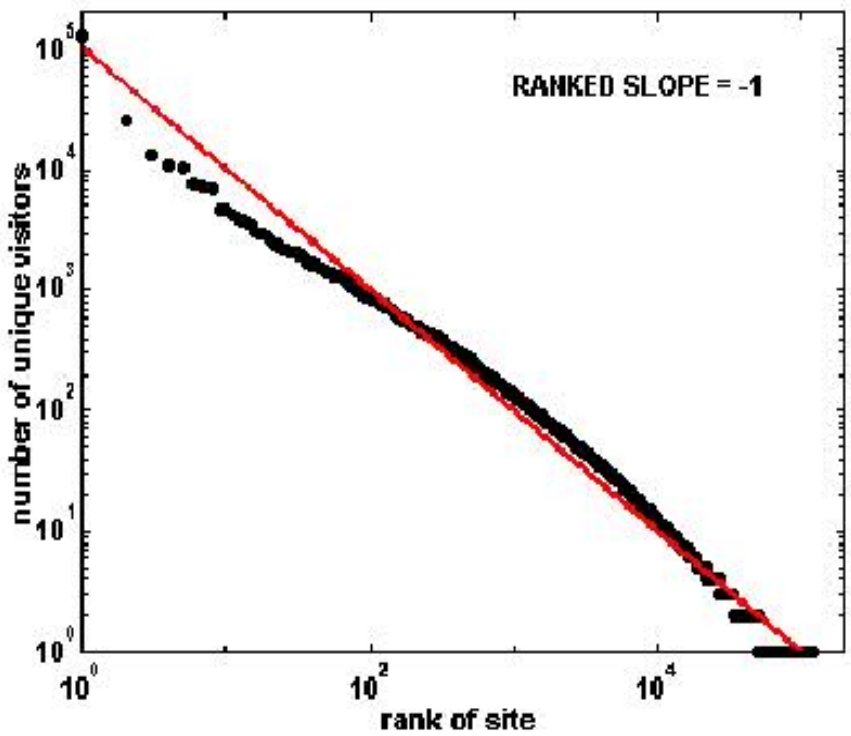
Rysunek: Rozkład odwiedzin użytkowników na poszczególnych stronach AOL (skale logarytmiczne) - **wykładnik = -2.07** (źródło: L.Adamic "Zipf, Power Laws and Pareto - a ranking tutorial")

Przykład - AOL - dystrybuanta (Pareto)



Rysunek: Rozkład odwiedzin użytkowników na poszczególnych stronach AOL (skale logarytmiczne) - dystrybuanta - **Pareto** ≈ 1.1 (źródło: L.Adamic "Zipf, Power Laws and Pareto - a ranking tutorial")

Przykład - AOL, uporządkowane (Zipf)



Rysunek: Strony AOL uporządkowane wg. popularności (**Zipf**) (źródło: L.Adamic "Zipf, Power Laws and Pareto - a ranking tutorial")

Wszystko powiązane

Jak widać, wszystkie te rozkłady są ze sobą ściśle powiązane i ilustrują to samo zjawisko.

Znajomość rozkładów i zależności jest cenna

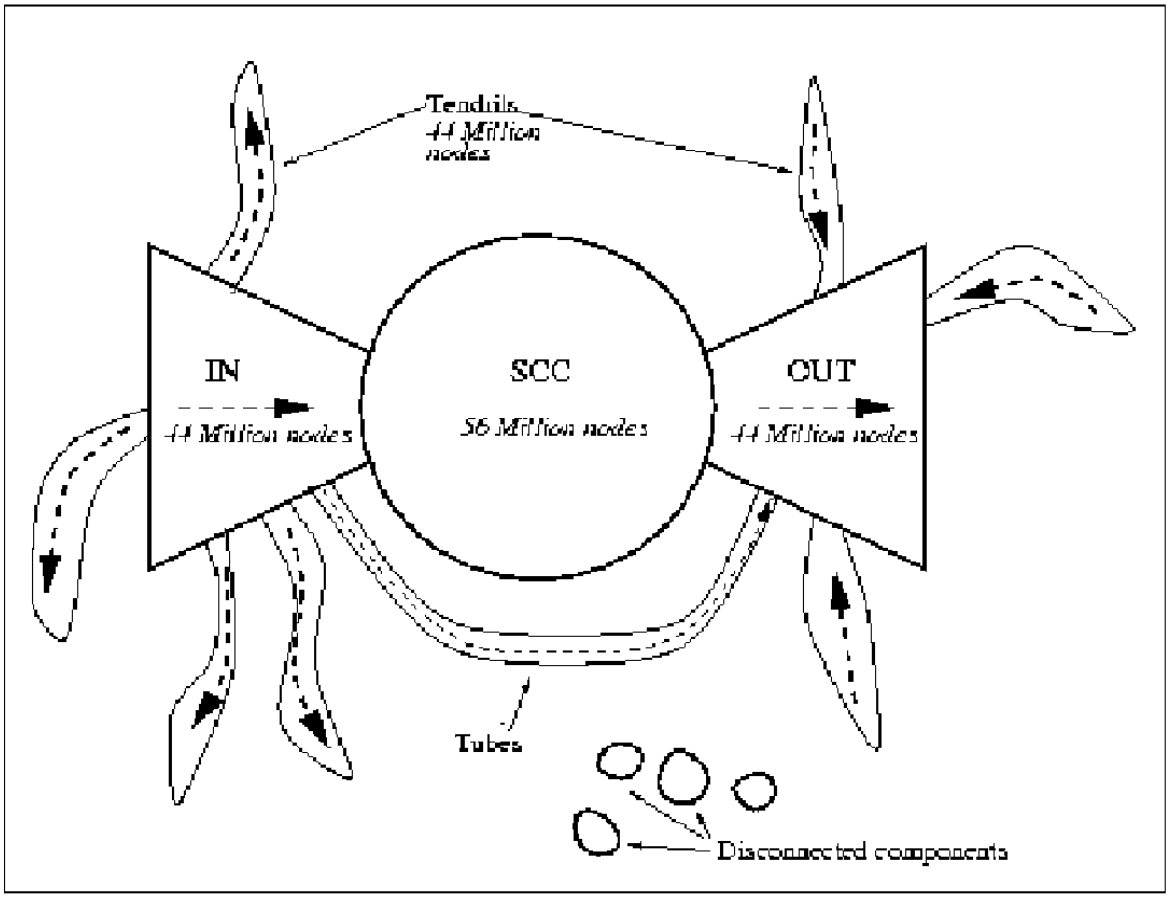
Znajomość rozkładu wielkości i zależności ma kluczowe znaczenie w:

- kompresji danych
- prawidłowej analizie danych
- projektowaniu struktur danych
- projektowaniu algorytmów

Spójność grafu WWW

- W kontekście koncepcji losowego internauty interesujące są pytania dotyczące silnej spójności grafu WWW.
- W wielu losowych grafach istnieje zjawisko tzw. *małego świata* (ang. *"small-world" phenomenon*) - *średnica* jest logarytmiczna $O(log(N))$ a średnia odległość niska.
- Graf WWW jest daleki od posiadania takiej właściwości - nie jest silnie spójny. Nie jest nawet słabo spójny.

"mucha" (ang. bow-tie)



Rysunek: Struktura badanego grafu

Spójność grafu WWW

Pomiar w 2000 roku na 203M dokumentów.

- największa silnie spójna składowa (SCC) tylko ok.25% badanego grafu.
- największa słabo spójna składowa (WCC) - 90%.
- przeciętna odległość (tylko dla odpowiednich par) - 16
- średnica SCC - 28
- średnica WCC - 500
- rozkład wielkości składowych silnie spójnych - potęgowy

Spójność grafu WWW

Podobne (w sensie zaprzeczenia zjawiska “małego świata”) wyniki osiągano w innych pomiarach.

1. Kleinberg, J. and R.Kumar and P.Raghavan and S.Rajagopalan and A.Tomkins, "The Web as a graph: measurements, models and methods", Proceedings of the 5th Annual International Computing and Combinatorics Conference, 1999
2. Broder, A. and R.Kumar and F.Maghoul and P.Raghavan and S.Rajagopalan and R.Stata and A.Tomkins and J.Wiener, "Graph Structure in the Web.", Proceedings of the 9th WWW Conference, 2000
3. Randall, K. and R.Stata and R.Wickremesinghe and J.Wiener, “The Link Database: Fast Access to Graphs of the Web”, Proc. of the Data Compression Conference, 2002

Samopodobieństwo grafu WWW

- Pomiar z 2001 roku. Różne podziały grafu WWW, ze względu na następujące kategorie:
  - zawierające pewne słowa kluczowe
  - mające ten sam host
  - będące w danym rejonie geograficznym
- Mierzono rozkłady stopni oraz wielkości silnie spójnych składowych.
- Każda część miała takie same właściwości statystyczne jak inne i jak cały graf.

1. Dill, S. and R.Kumar and K.McCurley and S.Rajagopalan and D.Sivakumar and A.Tomkins,"Self-Similarity in the Web", Proceedings of the 27th International Conference on Very Large Databases, 2001

Szczególne cechy grafów Intranetu

Grafy *Intranetów* mają trochę inne własności statystyczne.
np. SCC stanowiło tylko 10% całości grafu
Mimo to, rozkłady stopni są takie jak wszędzie.

1. Fagin, R. and R.Kumar and K.McCurley and J.Novak and D.Sivakumar and J.Tomlin and D.Williamson, "Searching the Workplace Web", Proc. of the 12th International WWW Conference, 2003

Społeczne aspekty grafu WWW

Badanie grafu WWW może służyć w wykrywaniu (a nawet przewidywaniu powstania!) nowych grup zainteresowań użytkowników.
Co ciekawe, w celach tych wystarczają czysto kombinatoryczne metody (np. oparte na identyfikowaniu klik dwudzielnych)

1. Kumar, R. and P.Raghavan and S.Rajagopalan and A.Tomkins, "Trawling the Web for Emerging Cyber-Communities", Proceedings of the 8th WWW Conference, 403-416, 1999
2. Gibson, D. and J.Kleinberg and P.Raghavan, "Inferring Web communities from link topology.", Proceedings of the 9th ACM Symposium on Hypertext and Hypermedia, 1998

Struktura Blokowa grafu WWW

Pomiar z 2001 roku.

- linki wewnątrz domen: 83.9% (95.2%)
- linki wewnątrz hostów: 79.1% (93.6%)

Macierz sąsiedztwa, odpowiednio poindeksowana ma strukturę blokową.
Np. poindeksowanie leksykograficzne po odwróconych domenach daje zagnieżdżoną strukturę blokową (domeny główne, poddomeny, hosty, ...).

1. Kamvar, S. and T.Haveliwala and C.Manning and G.Golub, "Exploiting the Block Struture of the Web for Computing PageRank", Stanford University Technical Report, 2003

Struktura Blokowa grafu WWW - aspekty praktyczne

Po przeindeksowaniu, strukturę blokową można wykorzystać do przyspieszenia obliczeń algorytmów rankingowych.

- możliwość równoległego obliczania algorytmów rankingowych na oddzielnych blokach macierzy
- redukcja kosztów we/wy (lokalność odwołań)

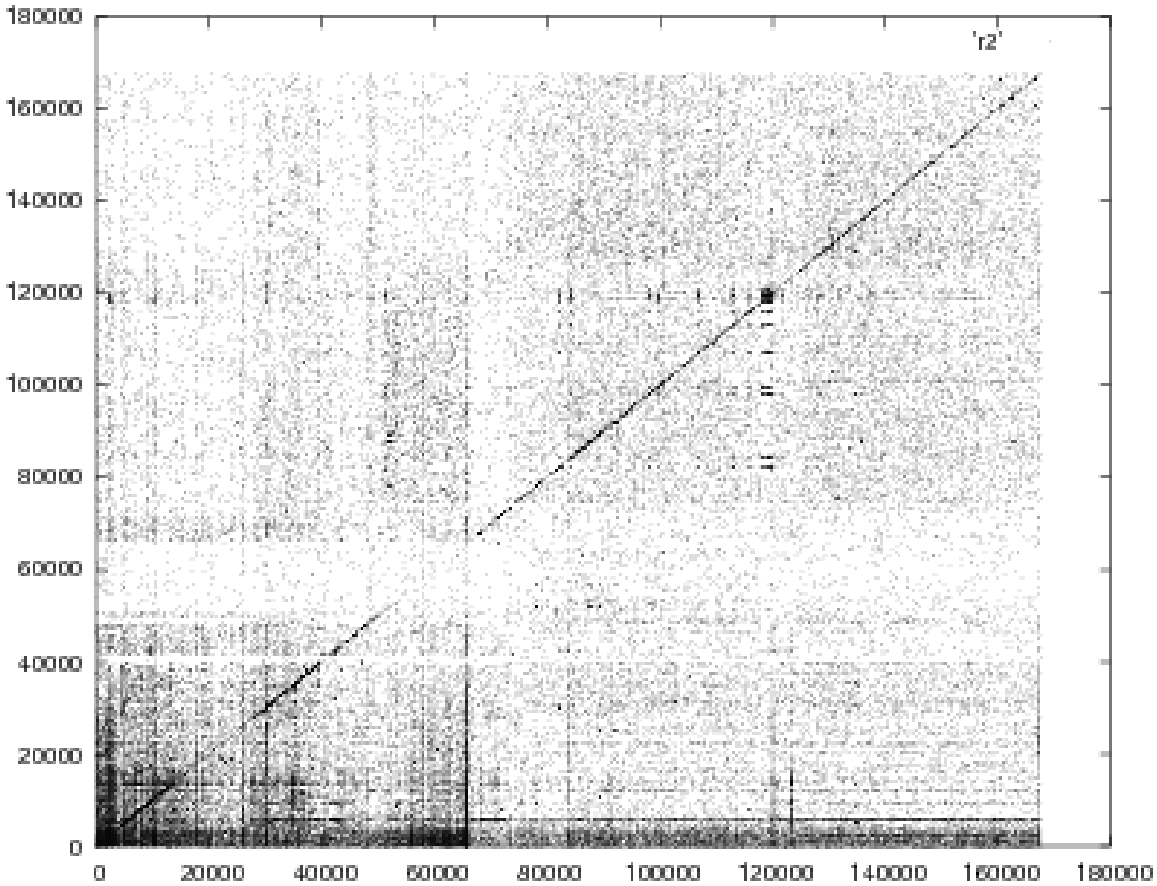
Osiąga się nawet *2-krotne* przyspieszenie obliczania PageRank dzięki wykorzystaniu tej techniki.

Struktura Blokowa grafu WWW

Na koniec zróbmy więc mały eksperymencik...

Struktura Blokowa grafu WWW

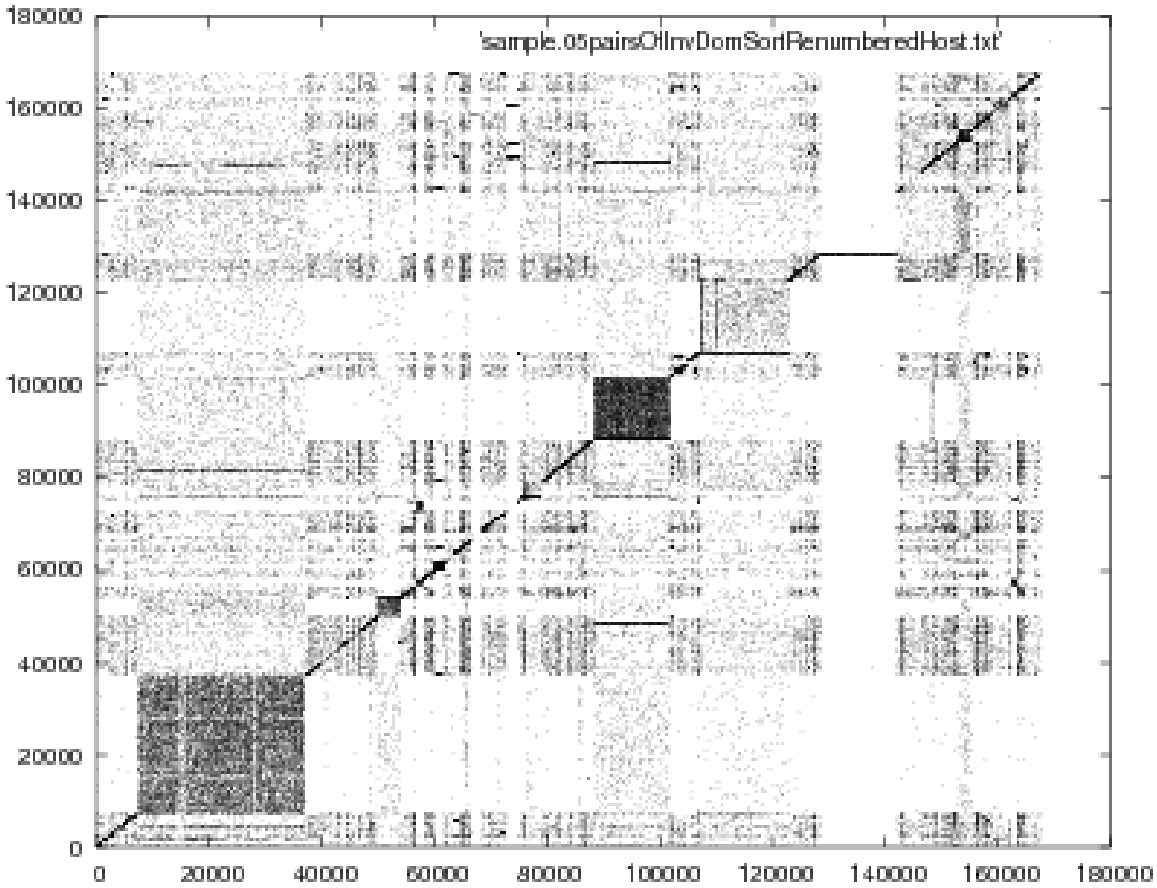
Na koniec zróbmy więc mały eksperymencik...



Rysunek: Macierz sąsiedztwa grafu hostów .pl z 2005 (kolejność crawlowania)

Teraz posortujmy po domenach od końca...

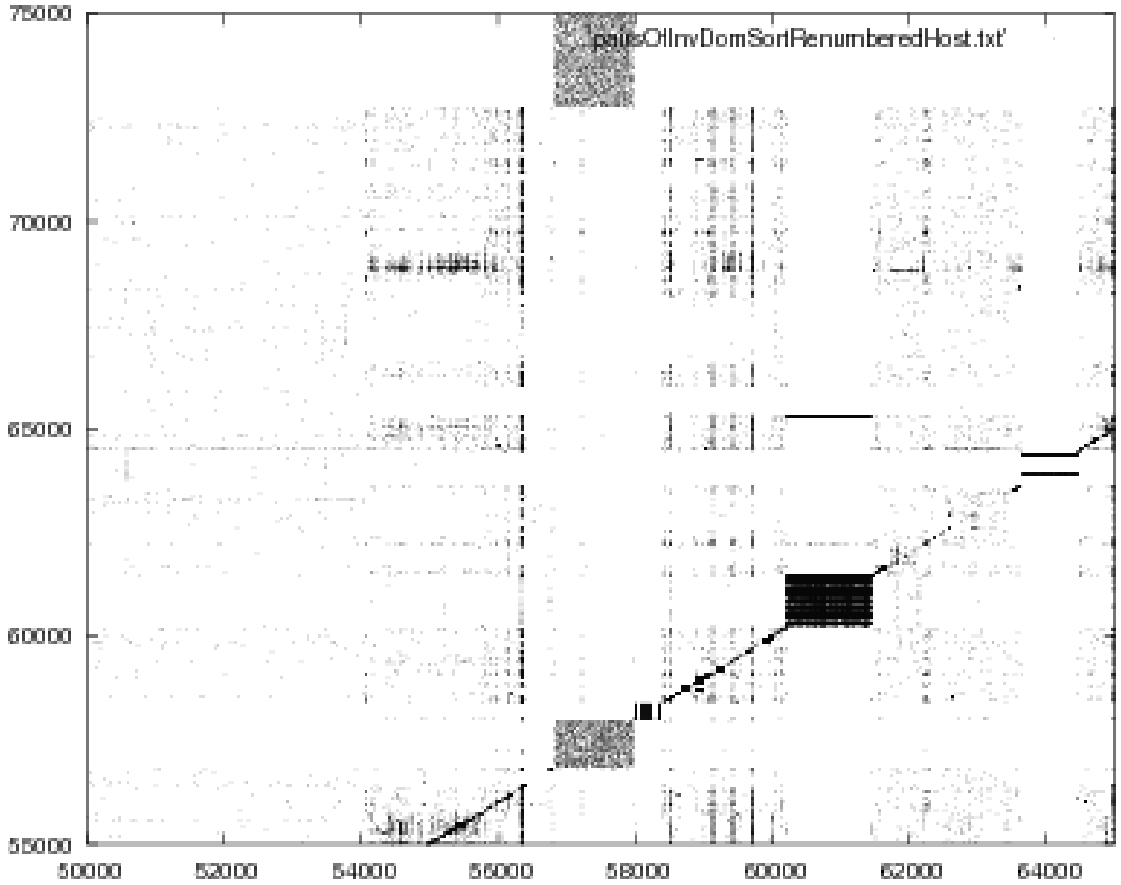
Teraz posortujmy po domenach od końca...



Rysunek: Ta sama macierz, ale po przeindeksowaniu...

Dominujące bloki to: blog.pl eblog.pl mylog.pl (zróbmy mały zoomik...)

Wizualizacja macierzy sąsiedztwa cd.



Rysunek: Powiększony interesujący fragment macierzy sąsiedztwa

enocleg.pl(57000) filmweb.pl(61000) info.pl (72500)

Struktura Blokowa grafu WWW - aspekty praktyczne

Wykorzystanie przeindeksowania ma więc też dużą wartość w wizualizacji i analizie danych WWW.
Wspomaganie w wykrywaniu:

- autorytetów (linie poziome)
- koncentratorów (linie pionowe)
- spamu

Na zaliczenie tego wykładu:

1. graf WWW
2. zastosowania analizy linków
3. jaki jest rozkład stopni wejściowych
4. co to jest rozkład potęgowy
5. rozkład Zipfa i Pareto, powiązania z potęgowym
6. gdzie jeszcze obserwuje się takie zjawiska? (3 przykłady)
7. podstawowe własności spójności grafu WWW
8. na czym polega "fraktalność" grafu WWW
9. na czym polega "struktura blokowa" grafu WWW (i do czego można ją wykorzystać)

Dziękuję za uwagę.