CoNe Freiburg

IIF INSTITUT FÜR INFORMATIK FREIBURG

Systeme II

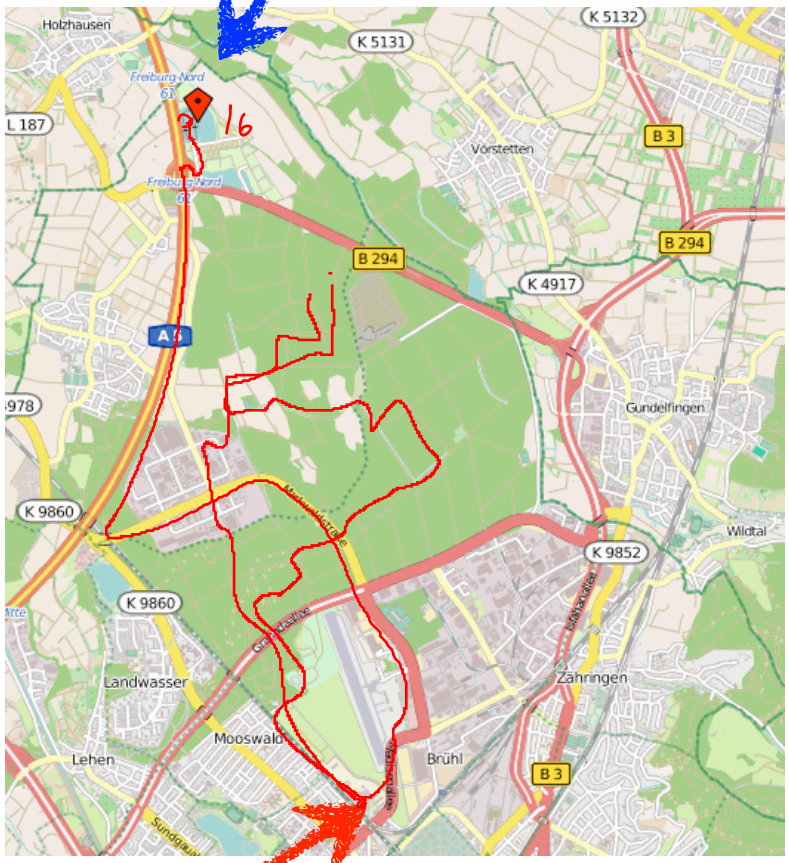
9. Die Struktur des Webs

Christian Schindelhauer
Technische Fakultät
Rechnernetze und Telematik
Albert-Ludwigs-Universität Freiburg
Version 15.07.2013

UNI FREIBURG

Abschlussveranstaltung am Tunisee

- Termin
  - Dienstag, 30.07.2012, 16:00 Uhr am Tunisee
  - oder 15:30 mit Fahrrad am Gebäude 051
- Plan
  - Moderner Dreikampf: Baden,Grillen, Trinken
- Versorgung
  - Getränke werden gestellt (vorher im Forum bestellen)
  - Grill und Grillkohle wird bereit gestellt
  - Essen durch Selbstorganisation



CoNe Freiburg

Der Webgraph

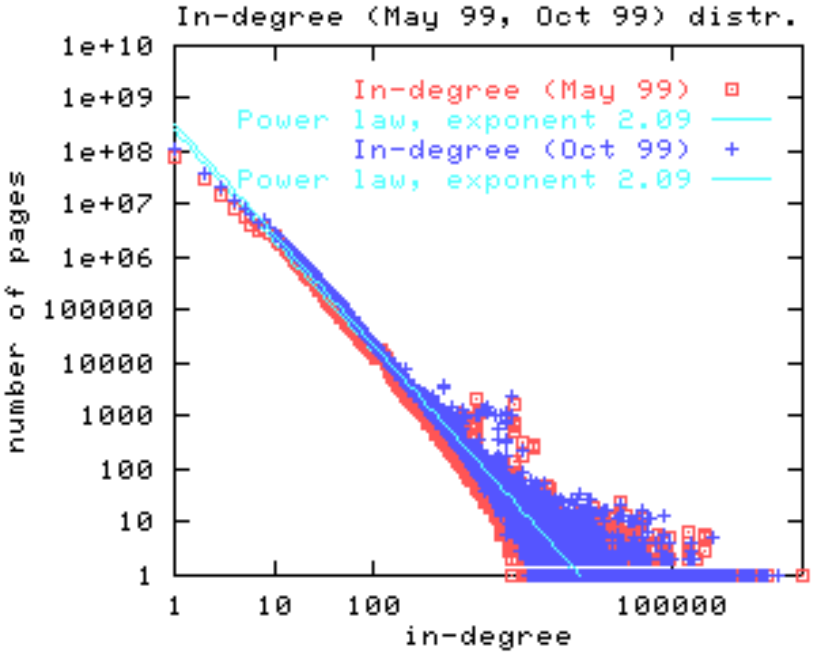
- G_{www}:
  - Statische HTML-Seiten sind **Knoten**
  - Link bezeichnen gerichtete **Kanten**
- Ausgrad eines Knoten:
  - Anzahl der Links einer Web-Seite
- Eingrad eines Knoten
  - Anzahl der Links, die auf eine Web-Seite zeigen
- Gerichteter Pfad von Knoten u nach v
  - Folge von Web-Seiten, indem man den Links folgt
- Ungerichteter Pfad ($u=w_0,w_2,\ldots,w_{m-1},v=w_m$) von Seite u nach v
  - Für alle i gibt es entweder einen Link von w_i nach w_{i+1} oder umgekehrt
- Starke (schwache) Zusammenhangskomponente
  - Maximale Knotenmenge in der zwischen allen Knoten dieser Menge ein (un) gerichteter Pfad besteht

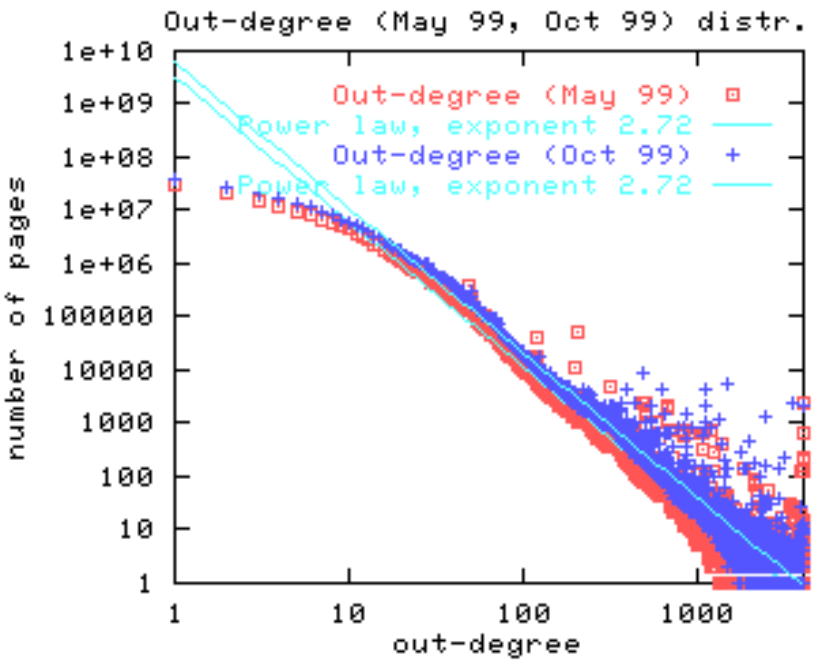
UNI FREIBURG

CoNe Freiburg

Ein- und Ausgradverteilung

- Der Ein- und Ausgrad gehorchen einem Potenzgesetz (power law)
  - d.h. die Häufigkeit von Eingrad i ist proportional zu ~ $1/i^{\alpha}$





- Ergebnisse von
  - Kumar et al 97: 40 Millionen Web-Seiten
  - Barabasi et al 99: Domain *.nd.edu + Web-Seiten in Abstand 3
  - Broder et al 00: 204 million Web-Seiten (Scan Mai und Oktober '99)

CoNe Freiburg

Pareto Verteilung = Verteilung nach Potenzgesetz

- Diskrete Pareto (power law) Verteilung für $x \in \{1,2,3,\ldots\}$

$$\mathbf{P}[X = x] = \frac{1}{\zeta(\alpha) \cdot x^{\alpha}}$$

mit konstanten Faktor $\zeta(\alpha) = \sum_{i=1}^{\infty} \frac{1}{i^{\alpha}}$

auch bekannt als Riemannsche Zeta-Funktion

- "Heavy tail"-Eigenschaft
  - nicht alle Momente $E[X^k]$ sind definiert
  - Der Erwartungswert existiert genau dann wenn $\alpha>2$
  - Varianz und $E[X^2]$ existieren genau dann wenn $\alpha>3$
  - $E[X^k]$ ist genau dann definiert wenn $\alpha>k+1$
- Dichtefunktion der kontinuierlichen Pareto-Verteilung für $x>x_0$

$$f(x) = \frac{\alpha - 1}{x_0} \left(\frac{x_0}{x}\right)^{\alpha}$$

UNI FREIBURG

CoNe Freiburg

Spezialfall: Zipf-Verteilung

- George Kinsley Zipf behauptete, dass die Frequenz des n-häufigsten Word mit Häufigkeit f(n) erscheint,
  - wobei f(n) n = c
- Zipf-Wahrscheinlichkeitsverteilung für x ∈ {1,2,3,…}

$$\mathbf{P}[X = x] = \frac{c}{x}$$

mit konstantem Faktor c ist nur definiert auf einem endlichen Abschnitt, da

$$\ln n \leq \sum_{i=1}^{n} \frac{1}{i} \leq 1 + \ln n$$

für große n gegen unendlich wächst

- Zipf Verteilungen beziehen sich auf Ränge
  - Der Zipf-Exponent α kann wie die Pareto-Verteilung größer sein als 1, d.h. $f(n) = c/n^{\alpha}$
- Pareto-Verteilungen beziehen sich auf die absolute Größe
  - e.g. Anzahl der Einwohner, Anzahl der Links, etc.

CoNe Freiburg

Pareto-Verteilung

- Beispiele für Power-Laws (= Pareto Verteilungen)

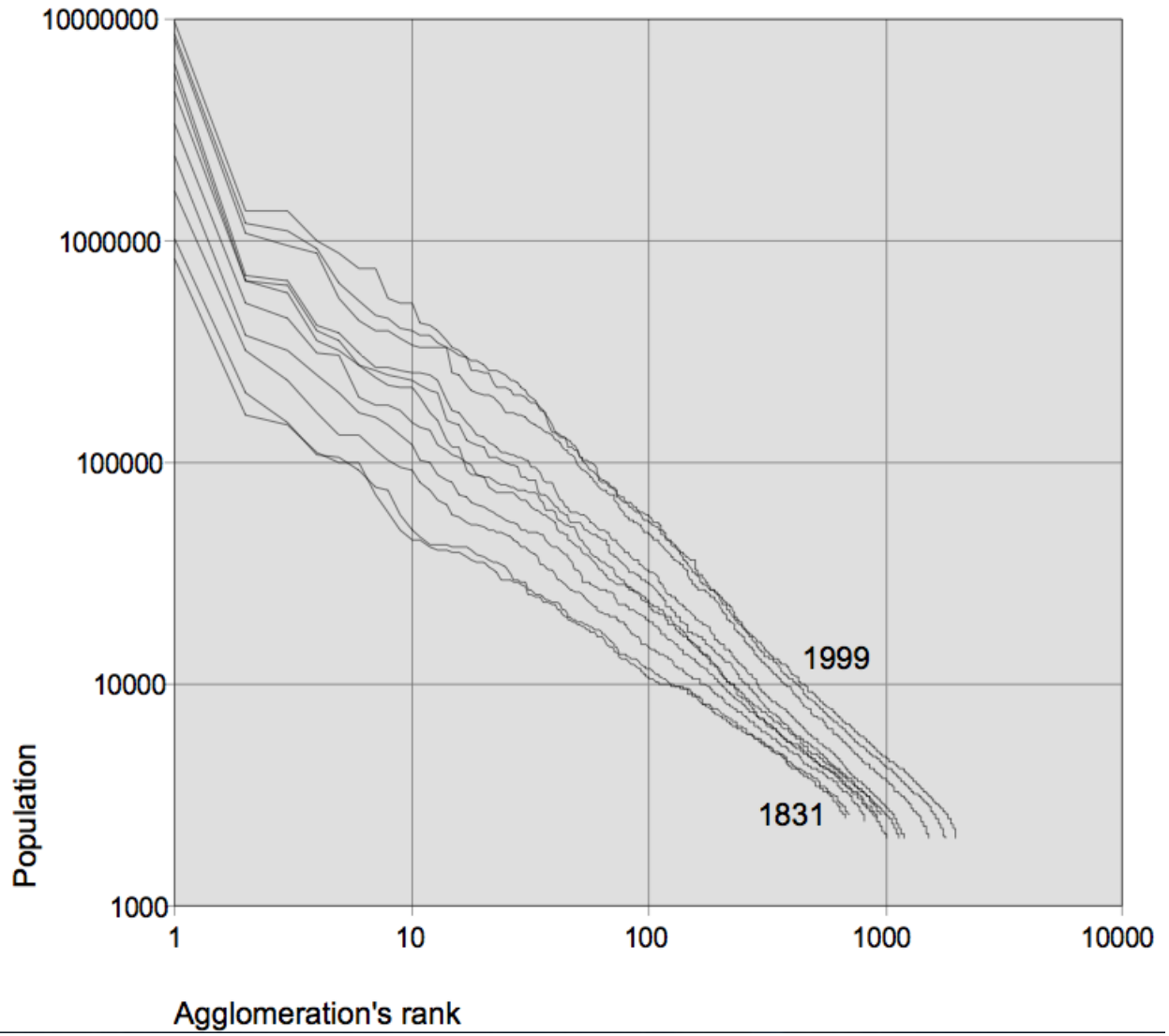
  - Pareto 1897: Einkommensverteilung in der Bevölkerung
  - Yule 1944: Word-Häufigkeit in Sprachen
  - Zipf 1949: Größe von Städten
  - Länge von Molekülketten
  - Dateilängen von Unix-Dateien
  - Initialmasse von Sternen
  - Zugriffshäufigkeit auf Web-Seiten
  - Länge von Telefonanrufen
  - …

CoNe Freiburg

Größe von Städten (I)

Scaling Laws and Urban Distributions, Denise Pumain, 2003

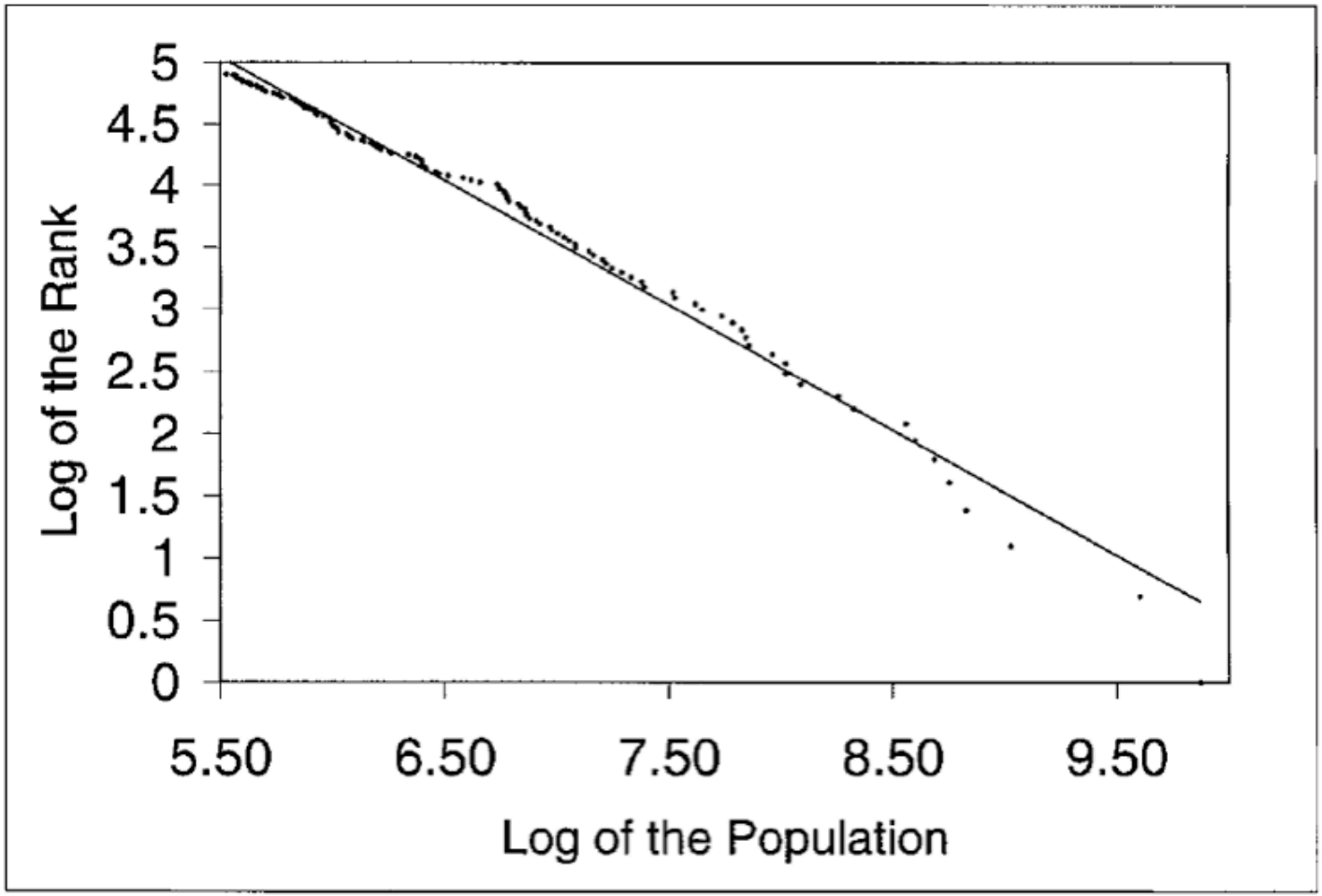
Figure 1 The hierarchical differentiation in urban systems:
Rank-size distribution of French agglomerations (1831-1999)



Zipf Verteilung

CoNe Freiburg

Größe von Städten (II)



Zipf Verteilung

FIGURE I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

UNI FREIBURG

CoNe Freiburg

Zipf's Law and the Internet

Lada A. Adamic, Bernardo A. Huberman, 2002

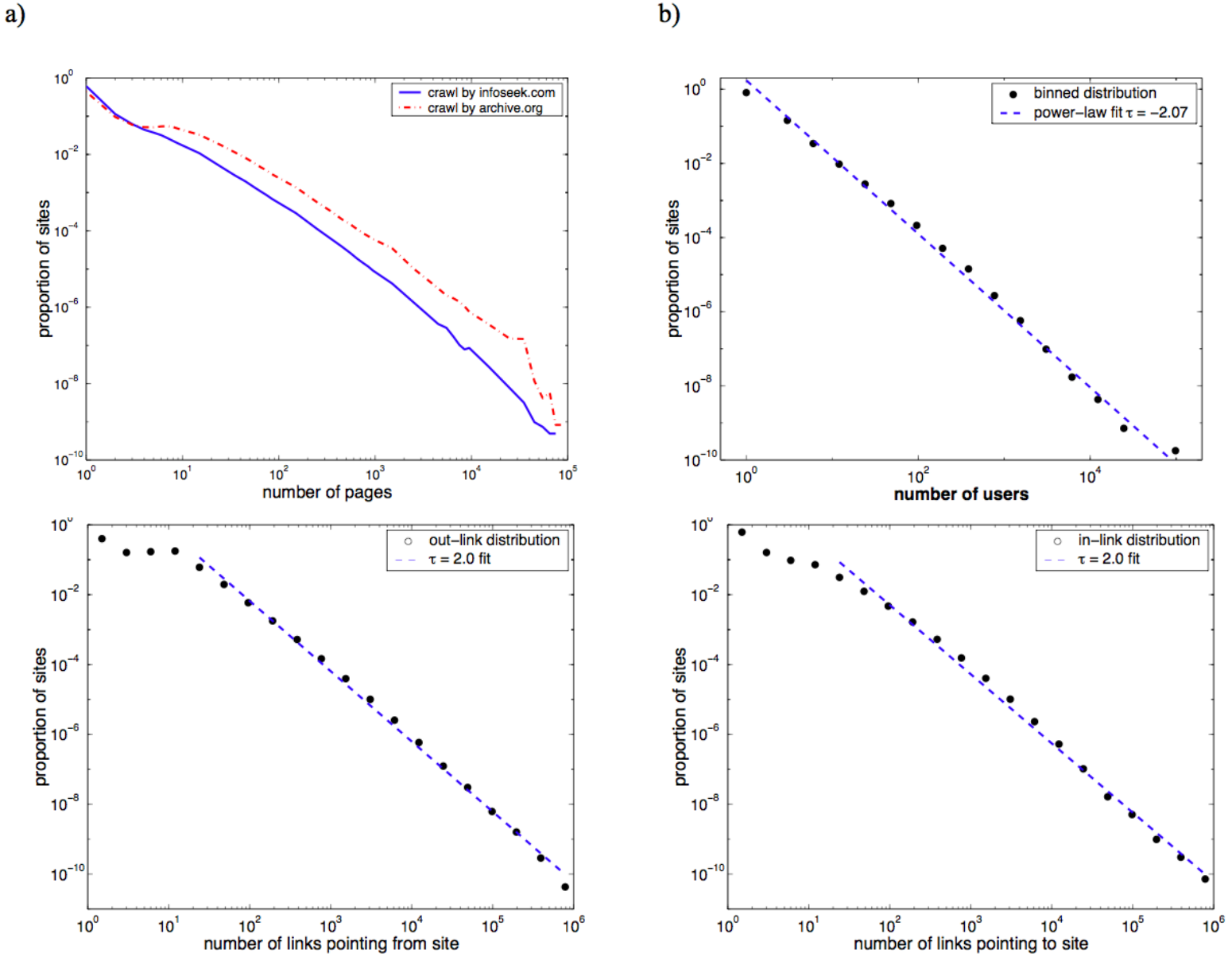


Figure 1. Fitted power law distributions of the number of site a) pages, b) visitors, c) out links, and d) in links, measured in 1997.

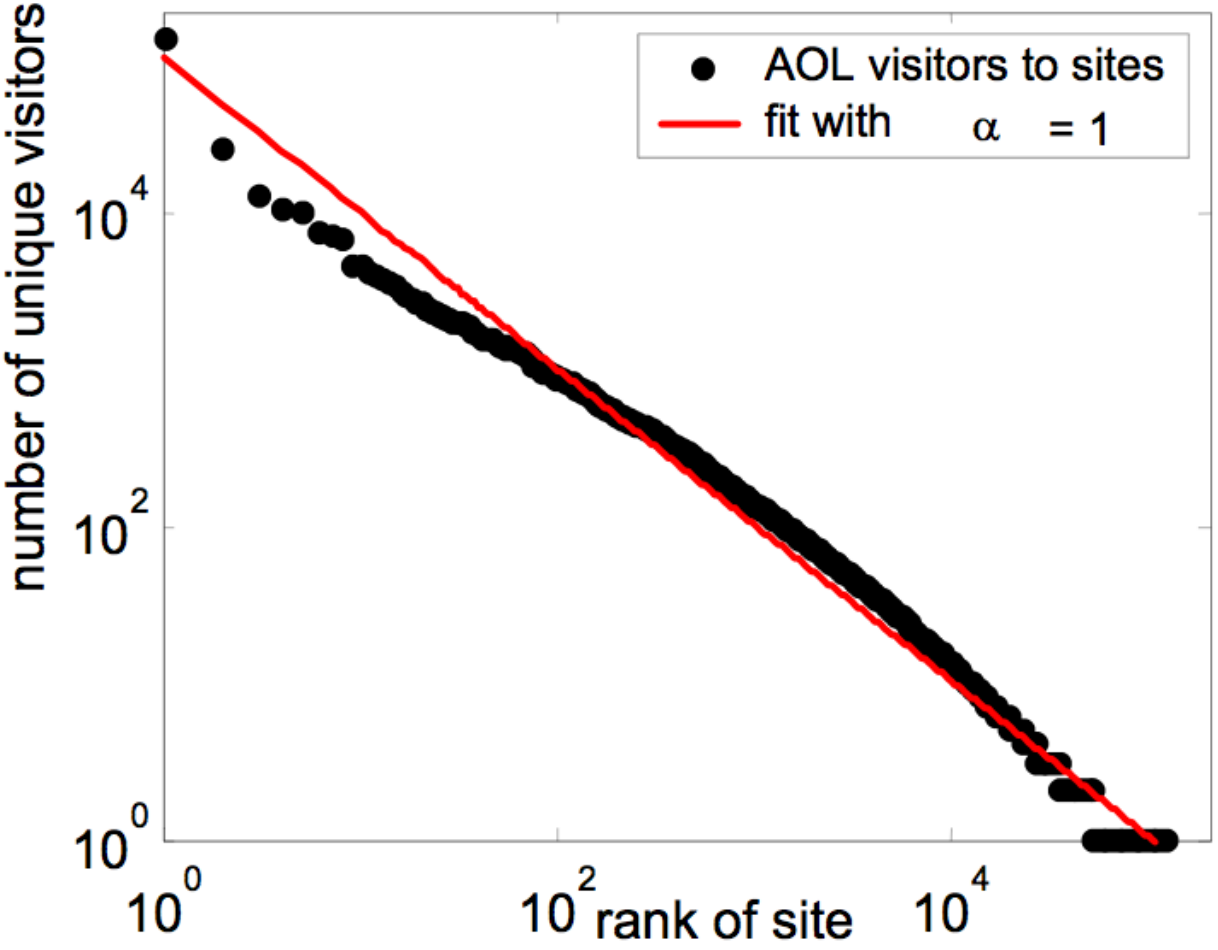
Pareto Verteilung

UNI FREIBURG

CoNe Freiburg

Zipf's Law and the Internet

Lada A. Adamic, Bernardo A. Huberman, 2002



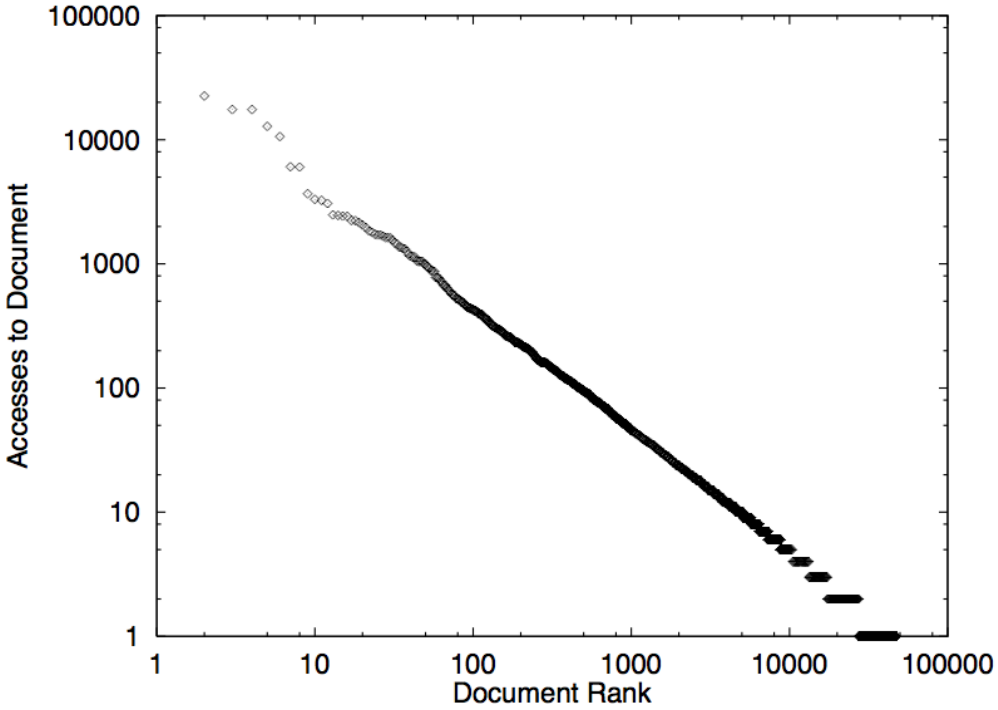
Zipf Verteilung

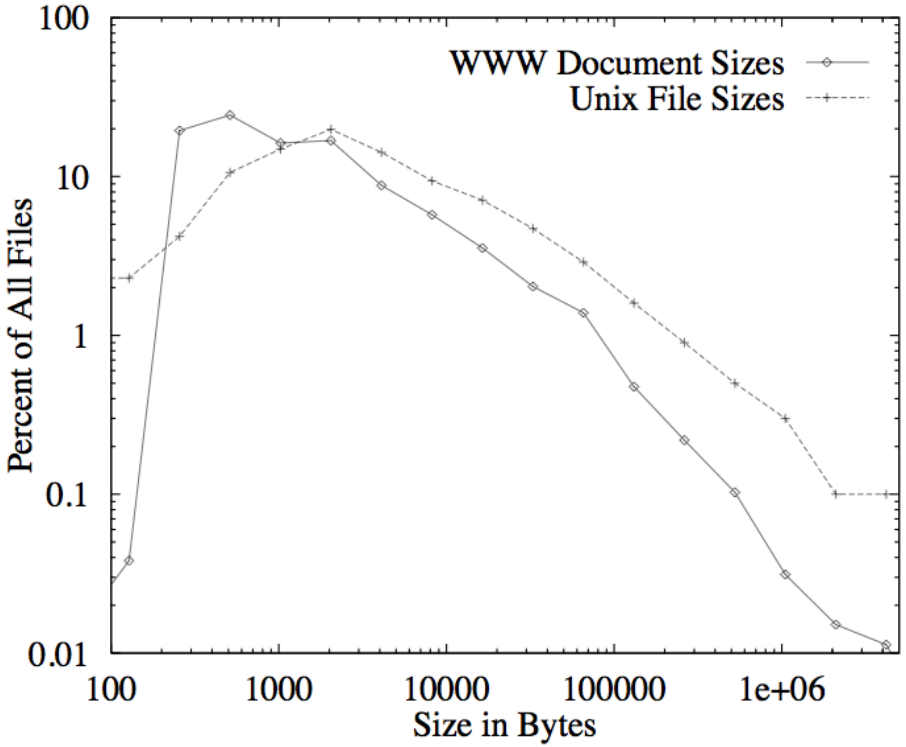
Figure 2. Sites ranked by the number of unique AOL visitors they received Dec. 1, 1997. AOL (America Online) is the largest Internet service provider in the United States. The fit is a Zipf distribution $n_r \sim r^{-1}$

UNI FREIBURG

CoNe Freiburg

Heavy-Tailed Probability Distributions in the World Wide Web
Mark Crovella, Murad, Taqqu, Azer Bestavros, 1996

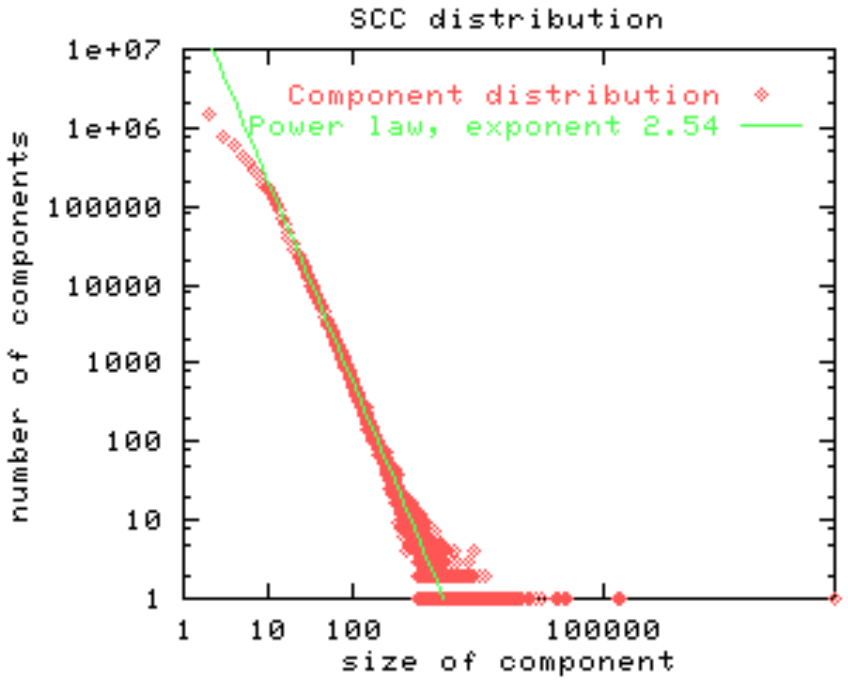


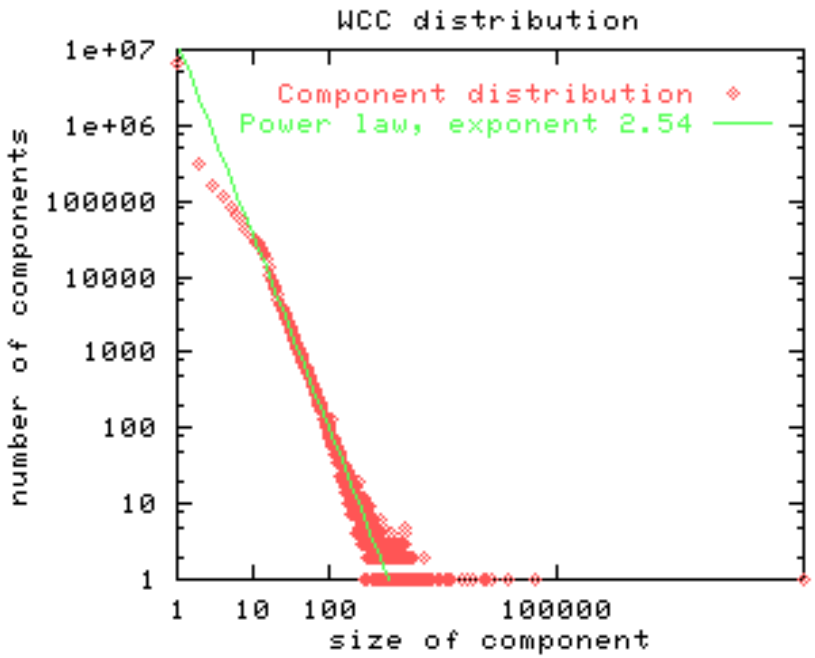


CoNe Freiburg

Größe der Zusammenhangskomponenten

- Starke und schwache Zusammenhangskomponenten unterliegen einer Pareto-Verteilung

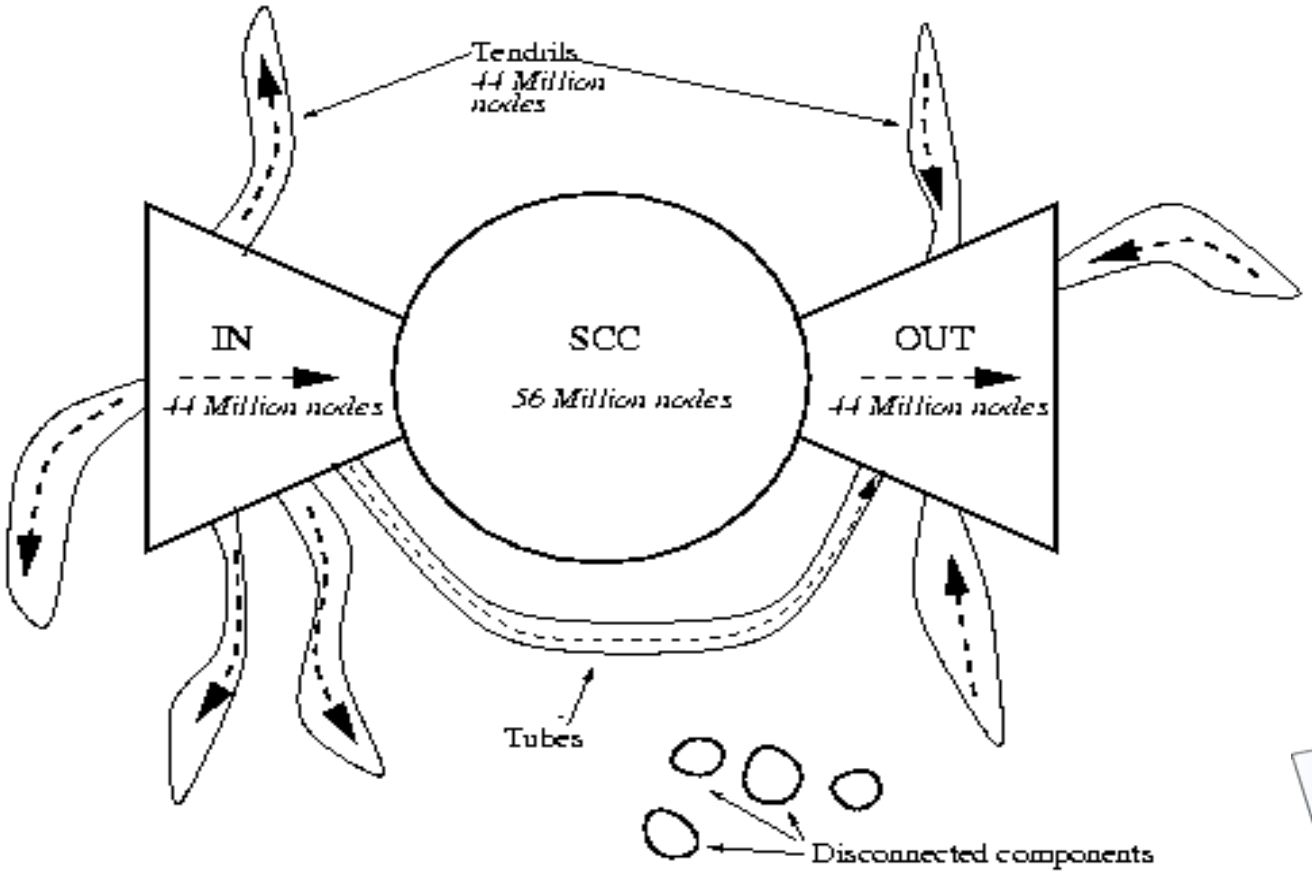


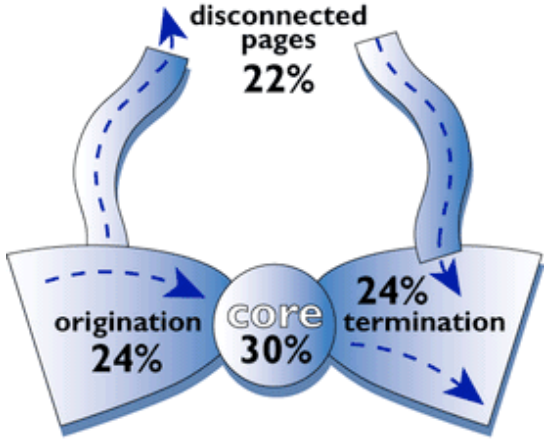


- A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. “Graph Structure in the Web: Experiments and Models.” In Proc. of the 9th World Wide Web Conference, pp. 309—320. Amsterdam: Elsevier Science, 2000.
  - Größte schwache Zusammenhangskomponente hat 91% aller Web-Seiten
  - Größte starke Zusammenhangskomponente hat 28% aller Webseiten
  - Durchmesser ist ≥ 28

CoNe
Freiburg

Der Web-Graph (1999)





CoNe Freiburg

Power-Laws im Internet !?!

- Erste Arbeit zu Power-Laws im Internet sagte JA
  - C. Faloutsos, P. Faloutsos, and M. Faloutsos, “On Power-Law Relationships of the Internet Topology,” in Proceedings of the ACM SIGCOMM, Sept. 1999.
- Diese Ergebnisse wurden 2002 angezweifelt:
  - Chen, Chang, Govindan, Jamin, Shenker, Willinger,The Origin of Power Laws in Internet Topologies Revisited, Infocom 2002
  - Gründe: Datenmenge zu gering, veraltet und spezielle Teilmengen ausgewählt
- Mittlerweile
  - Die Struktur des Internets gehorcht einem Power-Law, wenn man die Struktur berücksichtigt

CoNe Freiburg

Zipf's Law and the Internet

Lada A. Adamic, Bernardo A. Huberman, 2002

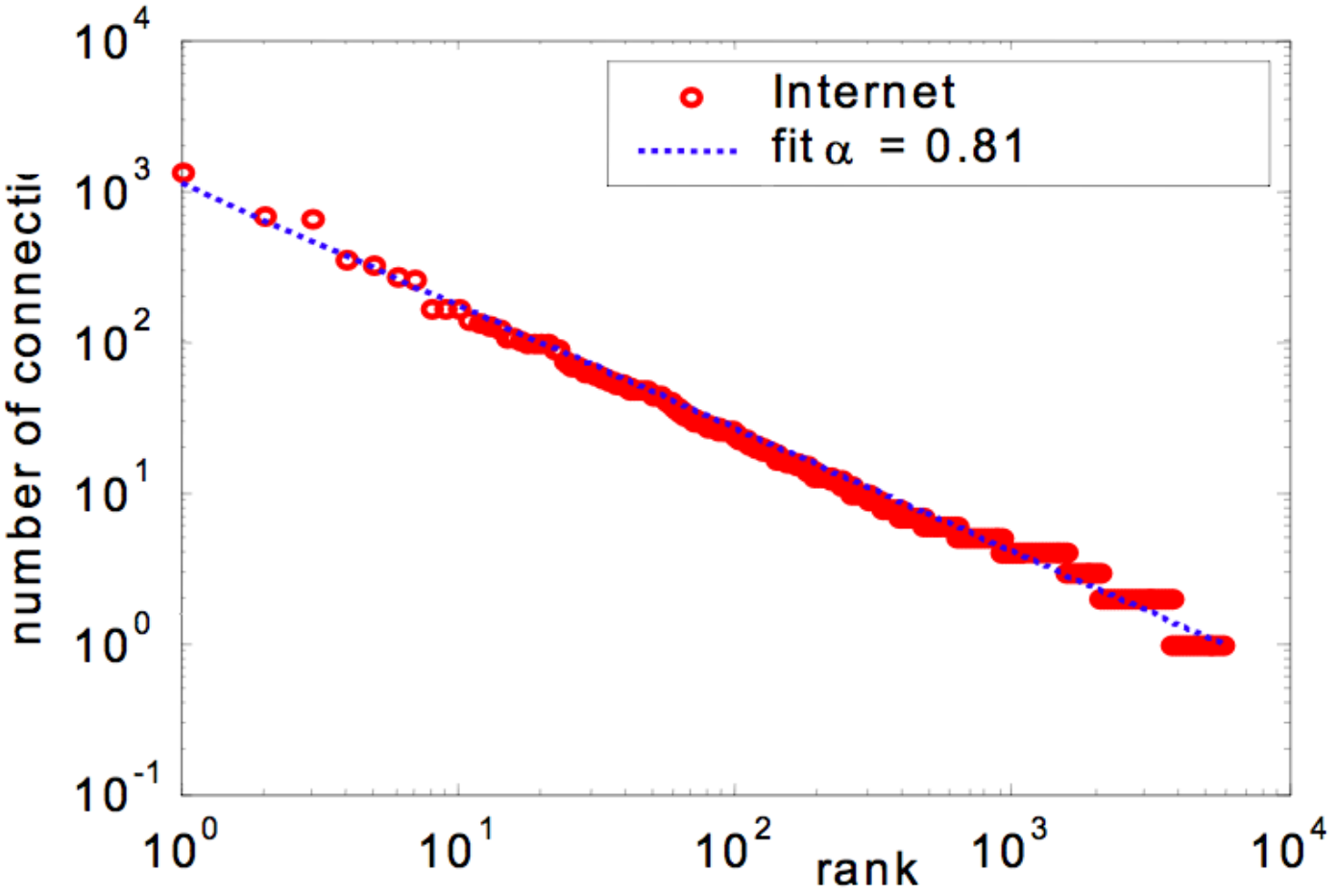


Figure 3. The connectivity of the internet backbone at the autonomous system (AS level). Each AS is itself a network corresponding to a single ISP, business entity or educational institution.

CoNe Freiburg

Die Struktur des Internets

- Aus
  - “Comparing the structure of power-law graphs and the Internet AS graph”, Sharad Jaiswal, Arnold L. Rosenberg, Don Towsley, INCP 2004
- Stub AS
  - verbinden zwei Hosts; Sackgassen
  - typischerweise Universitäten oder große Unternehmen
  - vertrauen auf Transit ASes für Anbindung ans Internet.
- Transit ASes
  - Service providers
  - typisch regionaler oder nationaler Internet Service Provider oder Backbone Network
- Tier (oder Ebene/Level)
  - bezeichnet Verbindungshierarchie

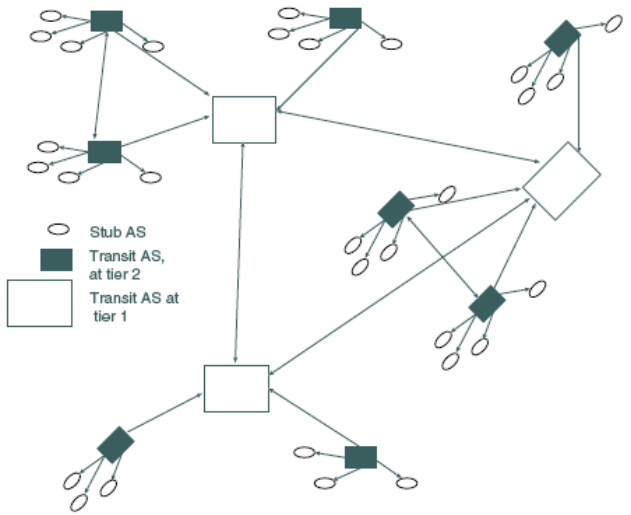


Fig. 1. Structure of the AS graph

UNI FREIBURG

CoNe Freiburg

Identifikation der Ebenen

- Die Identifikation der Ebenen ist nicht trivial und mehrdeutig
- Alternative Methoden zur Bestimmung der Ebenen
  - Nach Grad der Knoten
  - Mit Hilfe von PageRank-Algorithmus
    - Bestimmt “wichtigsten” Knoten
  - Anzahl der kürzesten Wege durch einen Knoten
    - Bestimmt Flaschenhälse
- All diese Methoden führen zu vergleichbaren Ergebnissen

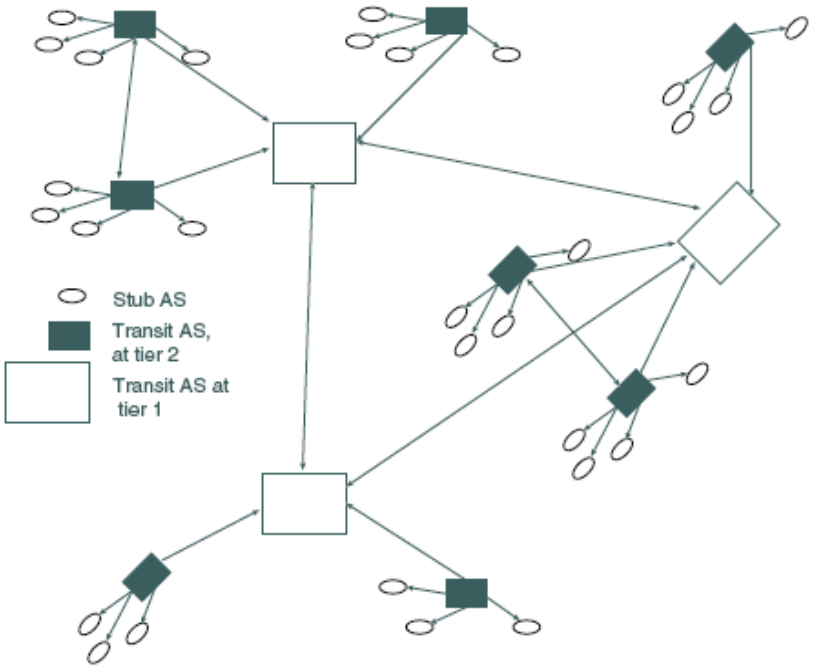


Fig. 1. Structure of the AS graph

“Comparing the structure of power-law graphs and the Internet AS graph”, Sharad Jaiswal, Arnold L. Rosenberg, Don Towsley, INCP 2004

UNI FREIBURG

CoNe Freiburg

Pareto-Verteilung des Grades von ASen im Internet

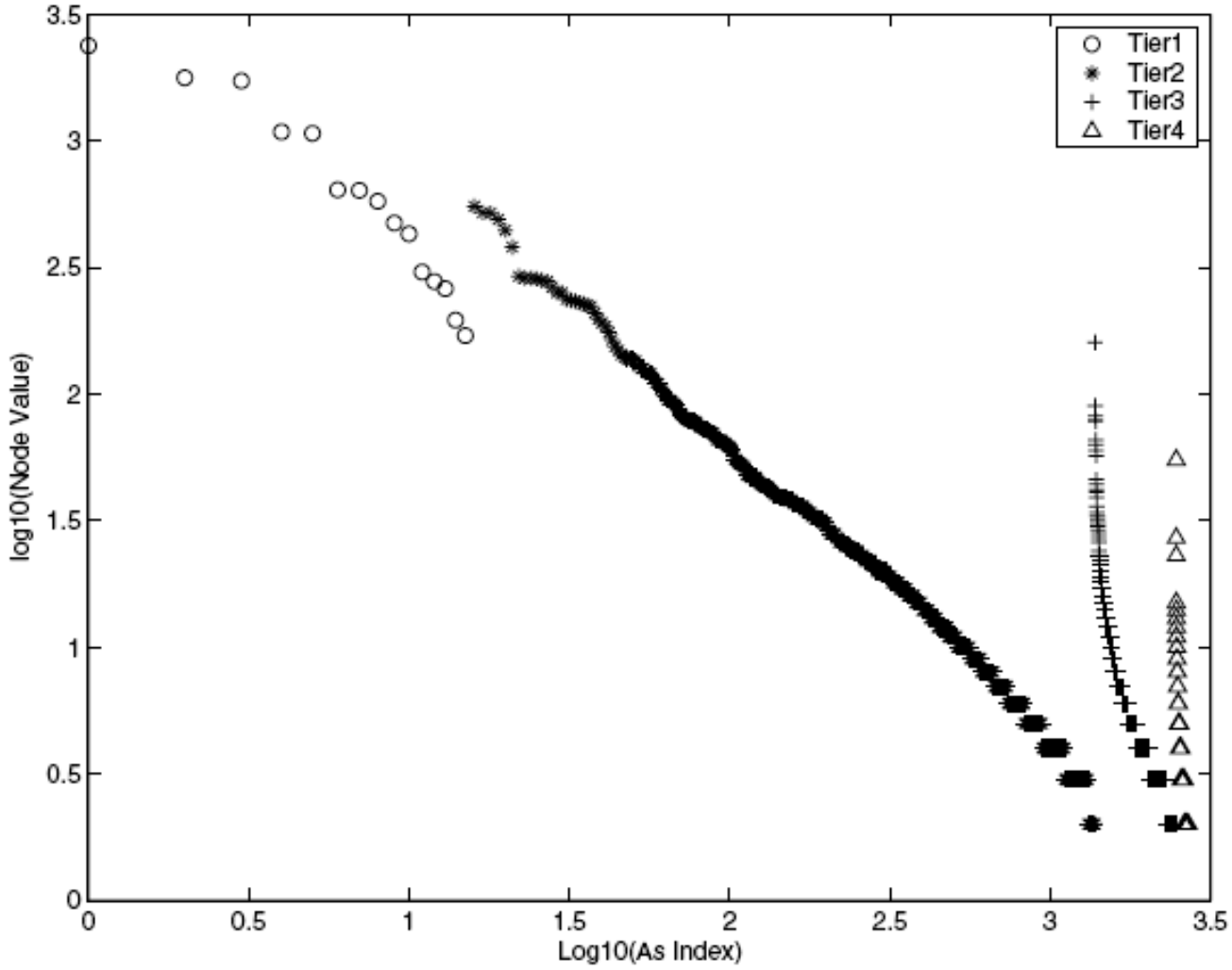


Fig. 2. Degree of ASes in different tiers

"Comparing the structure of power-law graphs and the Internet AS graph", Sharad Jaiswal, Arnold L. Rosenberg, Don Towsley, INCP 2004

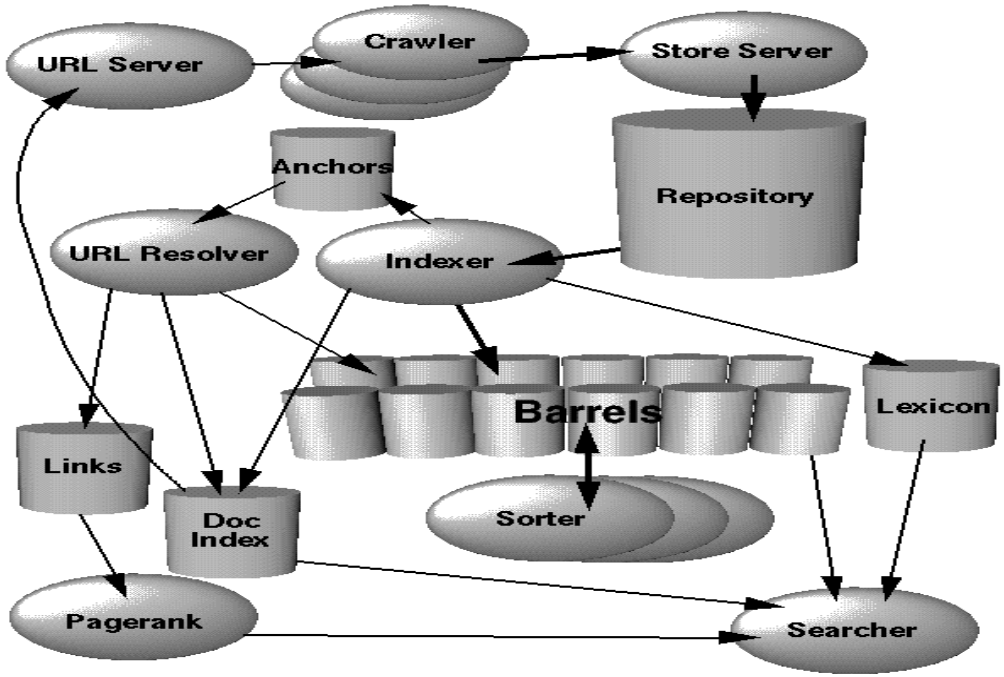
CoNe Freiburg

Webseitensuche

- PageRank [Brin&Page 98]
  - Vergibt jeder Web-Seite einen absoluten Rang (rank)/Autorität
  - Rang berücksichtigt Eingrad und Autorität des Eingrads
  - Idee: Seiten sind wichtig, wenn wichtige Seite auf sie zeigen

Die Anatomie einer Web Search Maschine

- "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Sergey Brin and Lawrence Page, Computer Networks and ISDN Systems, Vol. 30, 1-6, S. 107-117, 1998
- Design des Prototyps von Google
  - Stanford University 1998
- Hauptkomponenten
  - Web Crawler
  - Indexer
  - Pagerank
  - Searcher
- Hauptunterschied zwischen Google und anderen Suchmaschinen (1998)
  - Der Pagerank Algorithmus

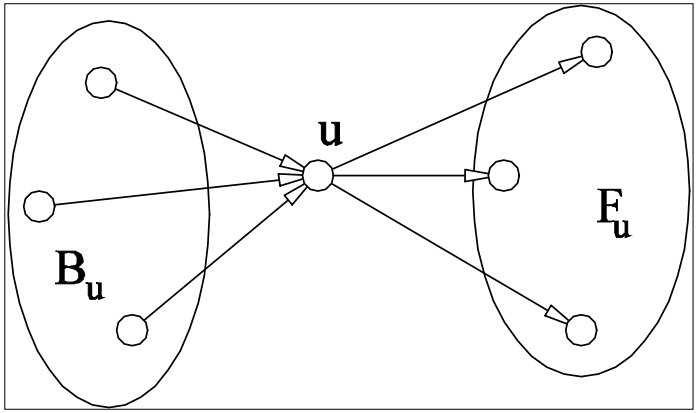


Der vereinfachte PageRank-Algorithmus

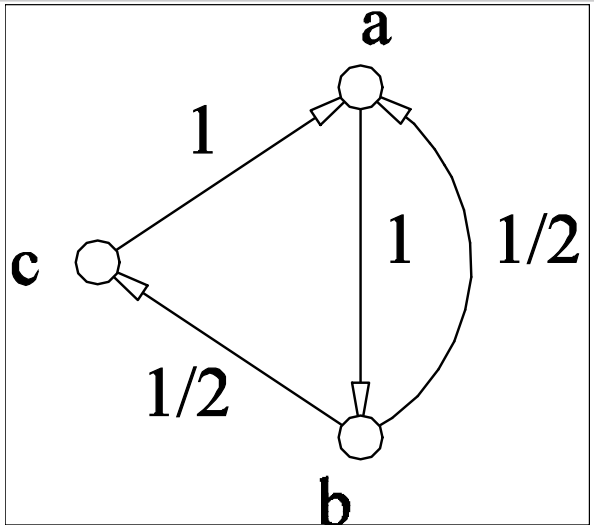
- Vereinfachter PageRank-Algorithmus
  - Rang einer Web-seite R(u) ∈ [0,1]
  - Wichtige Seiten übergeben ihre Gewicht an verlinkte Seite

$$R(u) \leftarrow c \sum_{v \in B_u} \frac{R(v)}{|F_v|}$$

  - c ist Normalisierungsfaktor so dass $||R(u)||_1 = 1$, d.h.
    - die Summe der Pageranks ist 1
  - Vorgänger-Knoten B_u
  - Nachfolger-Knoten F_u



Vereinfachter Pagerank-Algorithmus mit Beispiel



$$
\begin{aligned}
x &\leftarrow 0 \cdot x + \tfrac{1}{2} \cdot y + 1 \cdot z \\
y &\leftarrow 1 \cdot x + 0 \cdot y + 0 \cdot z \\
z &\leftarrow 0 \cdot x + \tfrac{1}{2} \cdot y + 0 \cdot z
\end{aligned}
$$

$$
\begin{pmatrix} x \\ y \\ z \end{pmatrix} \leftarrow \begin{pmatrix} 0 & \frac{1}{2} & 1 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}
$$

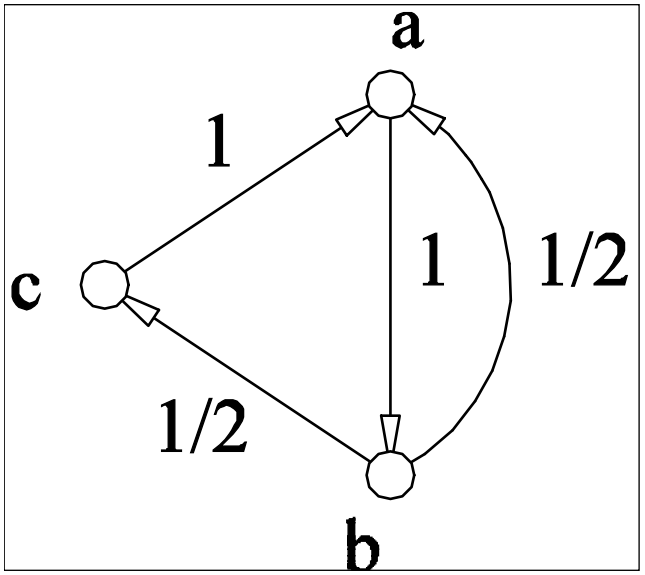
Runde	a	b	c
0	1	1	1
1	3/2	1	1/2
2	1	3/2	1/2
3	5/4	1/2	3/4
⋮	⋮	⋮	⋮
10	1,19..	1,22..	0,59..
⋮	⋮	⋮	⋮
20	1,20..	1,20..	0,60..

CoNe Freiburg

Matrixdarstellung

- $R \leftarrow c\,M\,R$,

  - wobei R ein Vektor (R(1),R(2),… R(n)) ist und
  - M die folgende n × n – Matrix bezeichnet



$$M_{ij} := \begin{cases} \dfrac{1}{|F_j|}\,, & \text{if } i \in F_j \\ 0\,, & \text{else.} \end{cases}$$

$$M = \begin{pmatrix} 0 & \frac{1}{2} & 1 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}$$

CoNe Freiburg

Eigenvektor als Fixpunkt der Rekursion Stochastische Matrix

- Die L1-Norm eines Vektors ist gegeben als $||x||_1 := \sum_{i=1}^{n} |x|$

Definition 1 *Eine $n \times n$-Matrix M heißt stochastisch, genau dann wenn $M_{ij} \geq 0$ und für alle $j \in [n]$ gilt:*

$$\sum_{i=1}^{n} M_{ij} = 1 ,$$

d.h. die Spaltensummen sind 1.

Lemma 1 *Für jede stochastische Matrix und Vektor x gilt:*

$$||M\ x||_1 \leq ||x||_1$$
$$||M\ x||_1 = ||x||_1 \qquad \textit{falls } x \geq 0 \textit{ oder } x \leq 0$$

⇒ Eigenwerte von M $|\lambda_i| \leq 1$

Theorem 4 *Für jede stochastische Matrix M gibt es einen Eigenvektor x mit Eigenwert 1, wobei $x \geq 0$ und $||x||_1 = 1$.*

CoNe Freiburg

Nachteile des vereinfachten Pagerank-Algorithmus

- Der Web-graph hat Senken, d.h. Seiten ohne Links
  - M ist keine stochastische Matrix
- Der Web-Graph ist periodisch
  - Daher ist die Konvergenz unsicher
- Der Web-Graph ist nicht stark zusammenhängend
  - Er ist noch nicht einmal schwach zusammenhängend
  - Unterschiedliche Konvergenzvektoren
- Rank-Senken
  - Starke Zusammenhangskomponenten saugen alle Gewichte der Vorgänger auf
  - Alle Vorgänger dieser Web-Seiten verlieren ihr Gewicht

CoNe Freiburg

Der Pagerank-Algorithmus

- Füge einer Senke Links auf alle Web-Seiten hinzu

- Wähle gleichwahrscheinlich eine Web-Seite
  - Mit einer gewissen Wahrscheinlichkeit q < 1 führe Schritt des vereinfachten Pagerank-Algorithmus durch
  - Mit Wahrscheinlichkeit 1-q wähle gleichwahrscheinlich eine Web-Siete

$$R(u) \leftarrow \frac{1-q}{n} + q \cdot \left(\sum_{v \in B_u} \frac{R(v)}{|F_v|} + \frac{|\{v \mid F_u = \emptyset\}|}{n} \right)$$

- Beachte: M ist stochastisch

$$M_{ij} := \begin{cases} \frac{1-q}{n} + \frac{q}{|F_j|}, & \text{falls } i \in F_j \\ \frac{1-q}{n}, & \text{falls } F_j \neq \emptyset \text{ und } i \notin F_j \\ \frac{1}{n}, & \text{falls } F_j = \emptyset \end{cases}$$

CoNe Freiburg

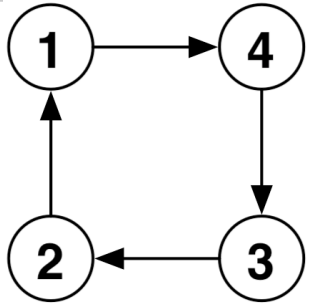
Eigenschaften des Pagerank-Algorithmus

- Der Graph der Matrix ist stark zusammenhängend
- Es gibt Rundwege der Länge 1
- Theorem
  - In nicht-periodischen Matrizen mit stark zusammenhängenden Graphen konvergiert der Markov-Prozess zu einem eindeutigen Eigenvektor mit Eigenwert 1
- ⇒ Pagerank konvergiert zu diesem eindeutigen Eigenvektor

Was passiert beim Matrix-Auffüllen

- Beispiel: q = 0.96, n = 4,

Graph:



- Startvektor $x = (1,0,0,0)^T$

$$\begin{pmatrix} 0{,}01 & 0{,}97 & 0{,}01 & 0{,}01 \\ 0{,}01 & 0{,}01 & 0{,}97 & 0{,}01 \\ 0{,}01 & 0{,}01 & 0{,}01 & 0{,}97 \\ 0{,}97 & 0{,}01 & 0{,}01 & 0{,}01 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0{,}01 \\ 0{,}01 \\ 0{,}01 \\ 0{,}97 \end{pmatrix}$$

- Beobachtung:
  - Alle Einträge von M sind mindestens (1-q)/n (hier: 0,01)
  - Alle Einträge von Mx sind mindestens $||x||_1$ (1-q)/n (hier: 0,01)
- Fakt: Für alle Vektoren $x \geq 0$: $(M\,x)_i \geq -||x||_1(1-q)/n$
- Fakt: Für alle Vektoren $x \leq 0$: $(M\,x)_i \leq ||x||_1(1-q)/n$

Was passiert mit gemischten Vektoren?

$$\begin{pmatrix} 0.01 & 0.97 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.97 \\ 0.97 & 0.01 & 0.01 & 0.01 \end{pmatrix} \begin{pmatrix} -1 \\ +1 \\ 0 \\ 0 \end{pmatrix} =$$

$\|z\|_1 = 2$

$\|p\|_1 = 1$

$\|m\|_1 = 1$

$$\begin{pmatrix} 0.01 & 0.97 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.97 \\ 0.97 & 0.01 & 0.01 & 0.01 \end{pmatrix} \begin{pmatrix} 0 \\ +1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.01 & 0.97 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.97 \\ 0.97 & 0.01 & 0.01 & 0.01 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix} =$$

$$\begin{pmatrix} 0.97 \\ 0.01 \\ 0.01 \\ 0.01 \end{pmatrix} + \begin{pmatrix} -0.01 \\ -0.01 \\ -0.01 \\ -0.97 \end{pmatrix} = \begin{pmatrix} 0.96 \\ 0 \\ 0 \\ -0.96 \end{pmatrix}$$

Für alle i: $\|(M\ p)_i\| \geq 0.01$

Für alle i: $\|(M\ m)_i\| \leq -0.01$

$\|M\ z\|_1 \leq 2 - 8 \cdot 0.01 = 1.92$

CoNe Freiburg

Eigenschaften des Pagerank-Algorithm

- Es gibt einen eindeutigen (reellen) Eigenvektor mit Eigenwert 1 für die Matrix des Pagerank-Algorithmus

- Pagerank konvergiert zu einer (1+ε)-Approximation des eindeutigen Eigenvektors in höchsten (-ln ε - ln n) / ln q Iterationen

CoNe Freiburg

Diskussion

- q = W'keit den vereinfachten Pagerank-Algorithmus zu benutzen
- Falls q klein ist
  - Pagerank konvergiert schneller
  - Kleinere Pfade sind relevanter
  - Weniger Strukturinformation wird benutzt
  - Die Pageranks sind ähnlicher
- Falls q groß ist
  - Pagerank konvergiert (möglicherweise) langsamer
  - Länge Wege spielen eine größere Rolle
  - Web-Senken sammeln mehr Gewicht auf
    - Daher löscht Google aus dem Web-Graph
- Problem:
  - Wie wählt man q
  - Macht es Sinn, dass jede Web-Seite unabhängig vom Suchwort gewählt wird?

CoNe Freiburg

IIF INSTITUT FÜR INFORMATIK FREIBURG

Systeme II

9. Die Struktur des Webs

Christian Schindelhauer
Technische Fakultät
Rechnernetze und Telematik
Albert-Ludwigs-Universität Freiburg

UNI FREIBURG