Netzbasierte Informationssysteme
Struktur und Erschließung des Web

Prof. Dr.-Ing. Robert Tolksdorf
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme
mailto: tolk@inf.fu-berlin.de
http://www.robert-tolksdorf.de

Größe des Web

Nach: Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener. Graph structure in the Web. Proc. 9th International World Wide Web Conference, 2000.

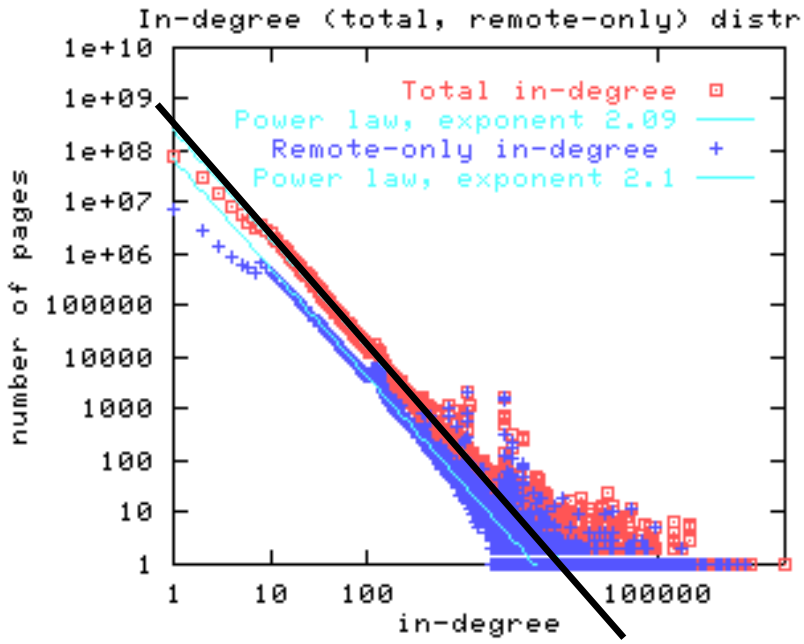
Grundlage

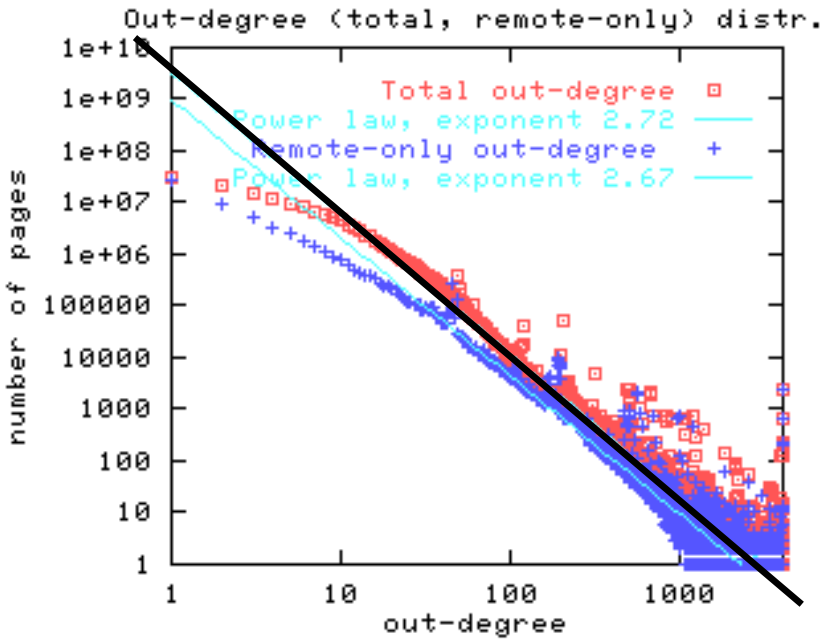
- Analyse der Struktur des Web
- Grundlagen
  - Daten von AltaVista
  - Repräsentation des Web-Graphen als Datenbank von URLs und Links
  -

	Datum	URLs	Links
Crawl1	Mai 99	203m	1466m
Crawl2	Oct 99	271m	2130m

Messung in- und out-Degree

- Web: Gerichteter Graph (V,E), Knoten V und Kanten E, Kante ist Paar (u,v) als Verbindung von u nach v
- in-degree: $|\{(u,v_1)\ldots(u,v_k)\}|$, out-degree: $|\{(v_1,u)\ldots(v_k,u)\}|$
- Anteil der Seiten mit in-degree i proportional zu $\frac{1}{i^{2,1}}$
- Anteil der Seiten mit out-degree i proportional zu $\frac{1}{i^{2,72}}$

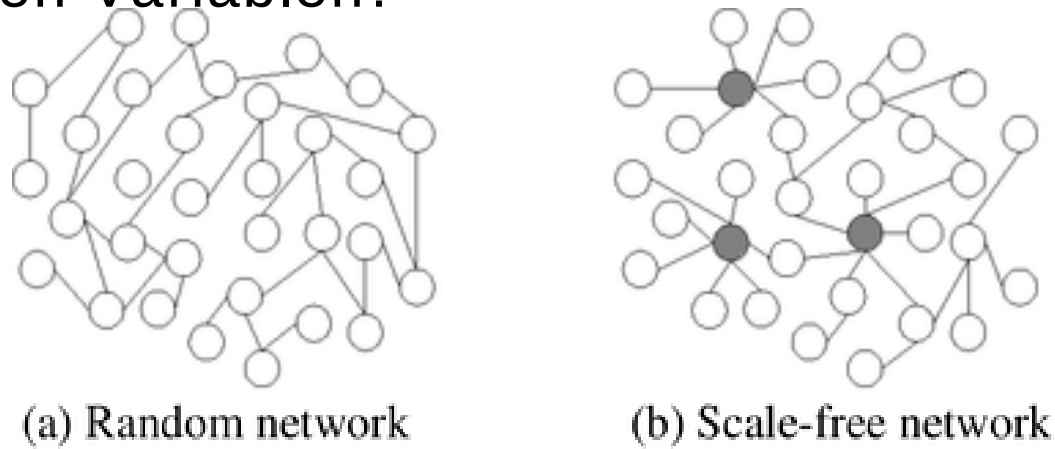




Power Laws

- Power Laws / Potenzgesetze beschreiben in verschiedenen Gebieten Verhältnisse zwischen Variablen:
  - Ökonomie (Pareto 1897)
  - Literaturanalyse (Yule 1944)
  - Soziologie (Zipf 1949)
  - Natur: Lawinenstärke
  - Web Charakteristiken

[Abb. wikipedia.org]

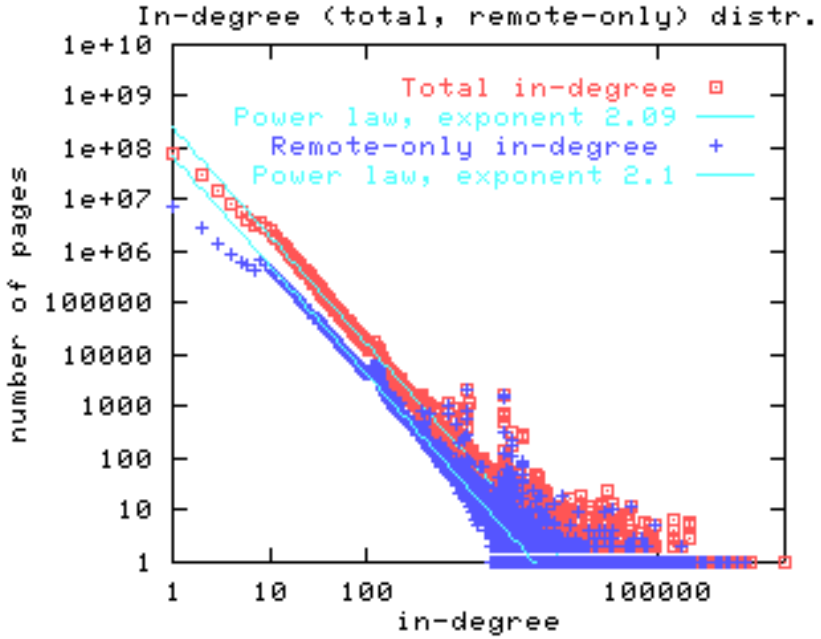


(a) Random network (b) Scale-free network

- Im Zufallsgraphen existiert zwischen zwei Knoten eine Kante oder eben nicht
  - Der Grad der Knoten (Anzahl der ein-/ausgehenden Kanten) ist Poisson-verteilt
- In „echten“ Graphen existiert eine andere Verteilung
  - Wenige Knoten haben einen hohen Grad
    - Wenige wissenschaftliche Arbeiten werden viel zititiert
  - Viele Knoten haben geringen Grad (long tail)
    - Sehr viele wissenschaftliche Arbeiten werden sehr wenig zitiert

Power Laws

- Auf logarithmischer Skala notiert:
- Form: $y \propto x^a$ für festes a>1
- a ist charakteristisch für Netzwerk



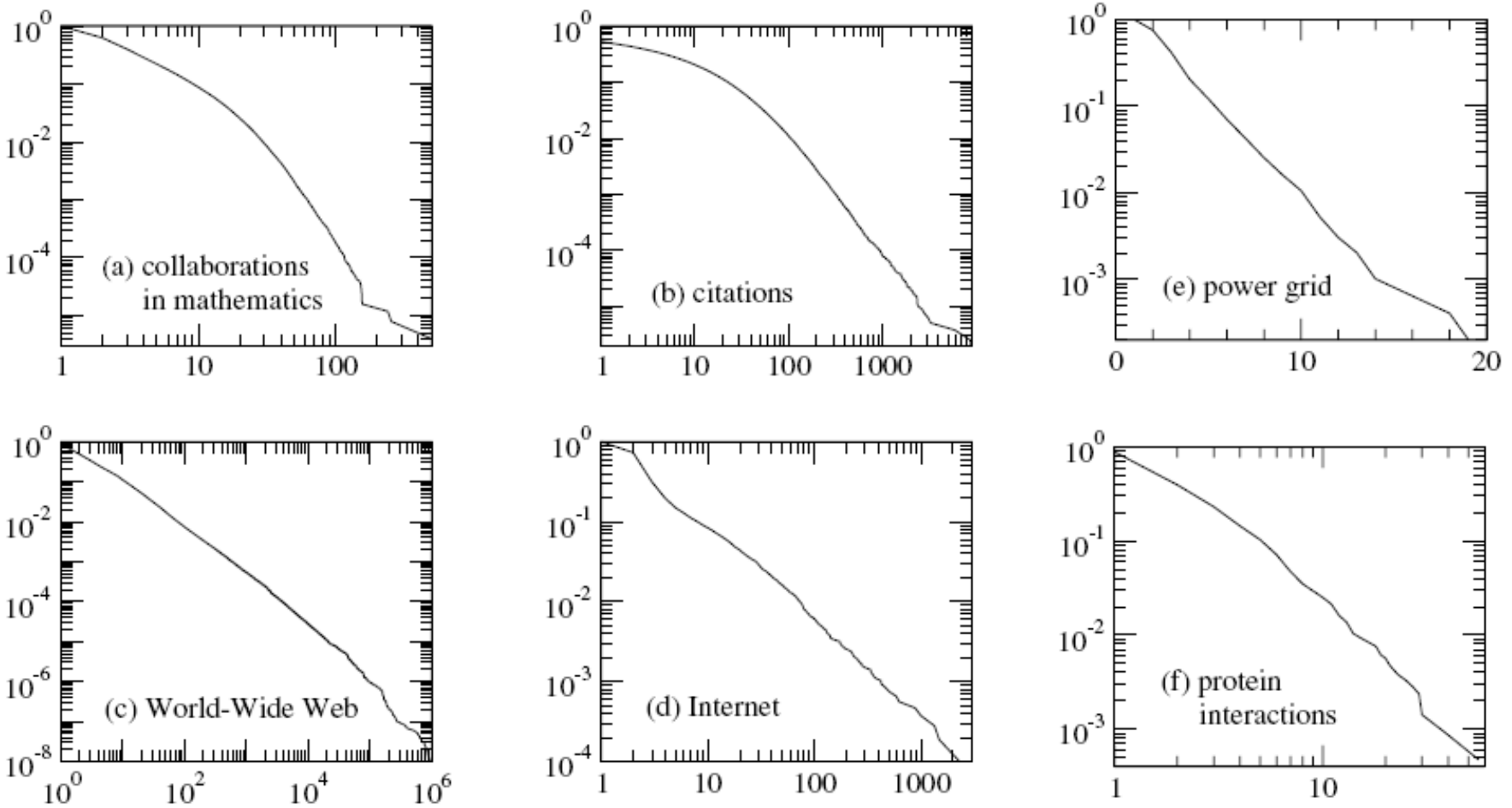
- Tritt als Phänomen an verschiedenen Stellen bei Web-Maßen auf (Topologie, Nutzerverhalten etc) auf
- Monotone strukturlose Verteilung
- Verhältnis ändert sich nicht entlang der Größenskalen -> Skalenfreiheit, komplette Verteilung ist durch a beschrieben

Power Laws

	Network	Type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
Social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	[20, 415]
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	[105, 322]
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	[107, 181]
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	[310, 312]
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	[310, 312]
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				[8, 9]
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		[136]
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	[320]
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	[45]
	sexual contacts	undirected	2 810				3.2				[264, 265]
Information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	[14, 34]
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				[74]
	citation network	directed	783 339	6 716 198	8.57		3.0/–				[350]
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	[243]
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		[119, 157]
Technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	[86, 148]
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	[415]
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	[365]
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	[317]
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	[394]
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	[155]
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	[6, 353]
Biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	[213]
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	[211]
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	[203]
	freshwater food web	directed	92	997	10.84	1.90	–	0.40	0.48	−0.326	[271]
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	[415, 420]

[M. E. J. Newman. The Structure and Function of Complex Networks. SIAM REVIEW Vol. 45,No . 2,167–256]

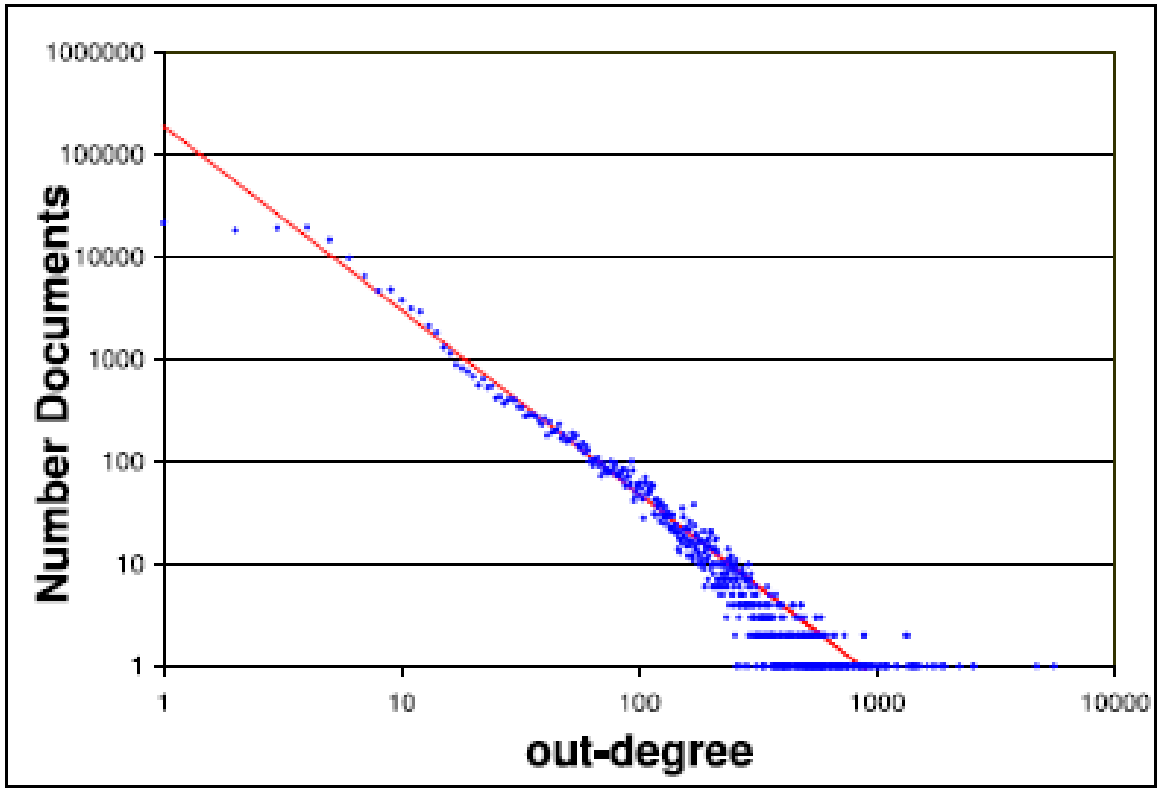
Power Laws



[M. E. J. Newman. The Structure and Function of Complex Networks. SIAM REVIEW Vol. 45,No . 2,167–256]

Der XML Web Graph

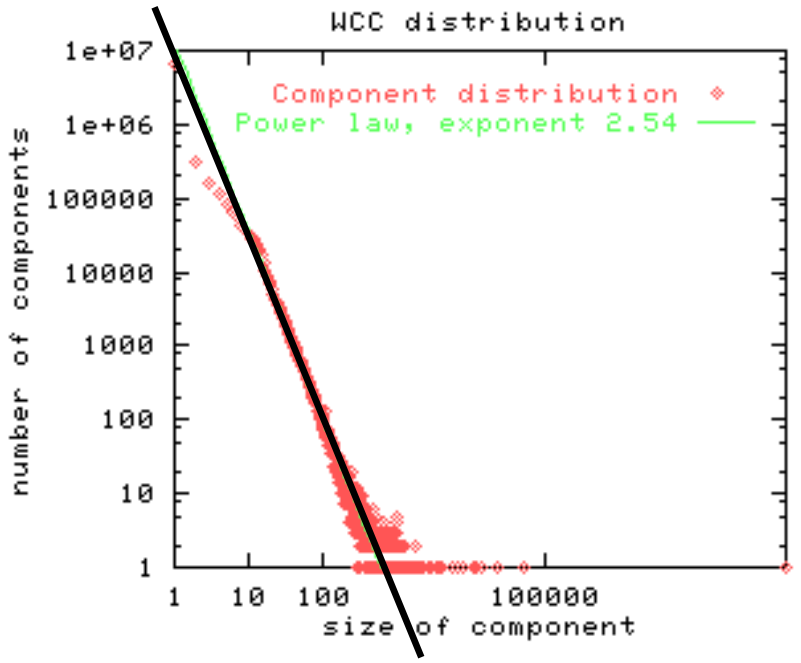
- [Barbosa, D., Mignet, L., and Veltri, P. 2005. Studying the XML Web: Gathering Statistics from an XML Sample. *World Wide Web* 8, 4 (Dec. 2005), 413-438. http://www.ucalgary.ca/~denilson/docs/WWWJ.pdf]
- Der durch href, xmlhref und xlink:href gebildete Graph aus XML Dokumenten:
- a=1,8



Komponenten im ungerichteten Graphen

- Ungerichteter Graph (V,E) mit Kanten als {u,v}
- Pfad: (u,u_1), (u_1,u_2), … (u_k,v), {u,v}⇒(u,v),(v,u)
- Komponente: Menge von Knoten, so dass für Knoten u und v im Graphen ein Pfad von u nach v existiert
- Eine große Komponenten mit 186m Knoten (91%)
- Verteilung der Größen der Komponenten folgt Potenzgesetz mit

$$\frac{1}{n^{2,54}}$$



Komponenten im ungerichteten Graphen

Freie Universität Berlin

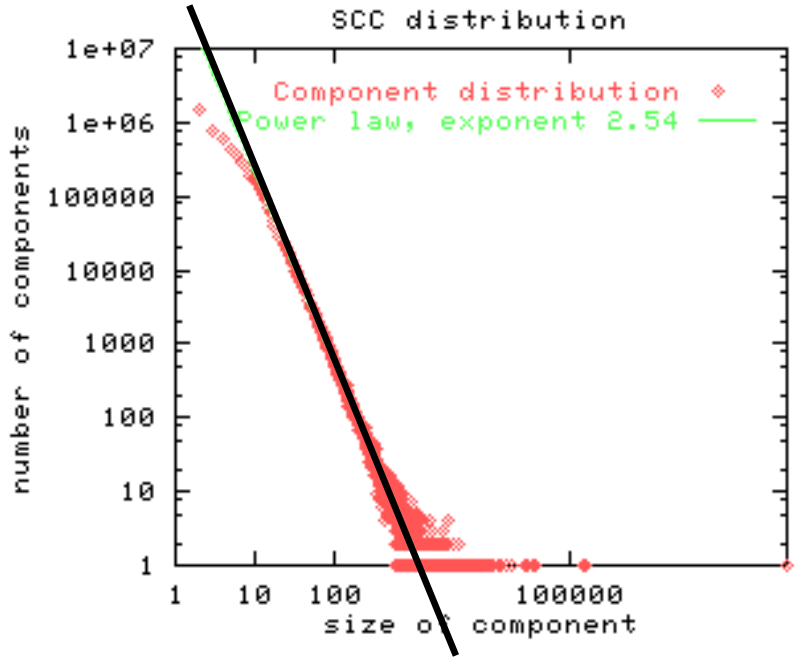
- *Autoritäten*:
  Seiten, auf die viele verweisen (hoher in-degree)
  Beispiel: www.w3c.org
- *Hubs*:
  Seiten, die auf viele verweisen (hoher out-degree)
  Beispiel: www.dmoz.org
- Sind Hubs und Autoritäten für die großen Komponenten verantwortlich?
- Links auf Seiten mit hohem in-degree entfernen (>5):
  Große Komponente mit Größe 59m Seiten

- Fazit:
  Das Web ist auch ohne Hubs und Autoritäten gut verknüpft

Komponenten im gerichteten Graphen

- Stark verbundene Komponente (SCC): Knotenmenge, so dass für alle u, v ein Pfad von u nach v existiert
- Eine große Komponente mit 56m Knoten (28%)
- Andere Komponenten deutlich kleiner
- Powerlaw für Größen der Komponenten mit

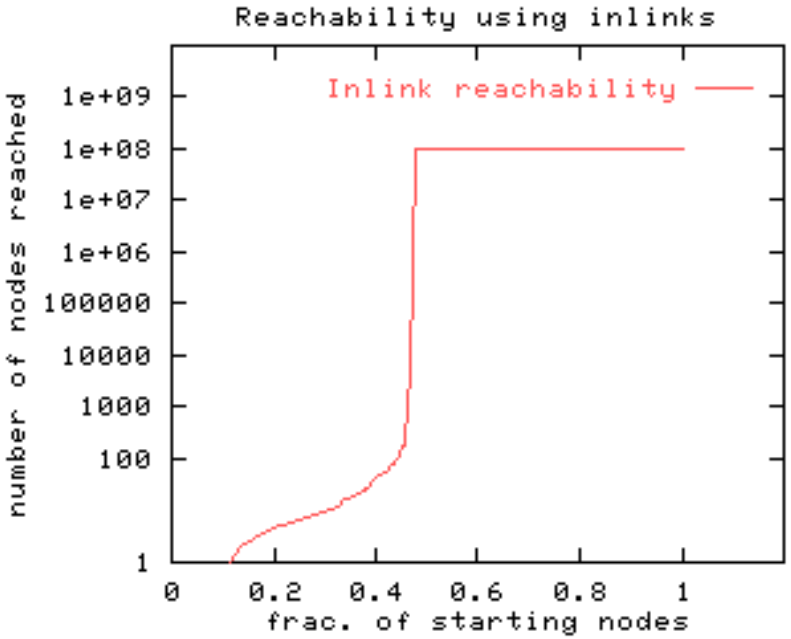
$$\frac{1}{n^{2,54}}$$

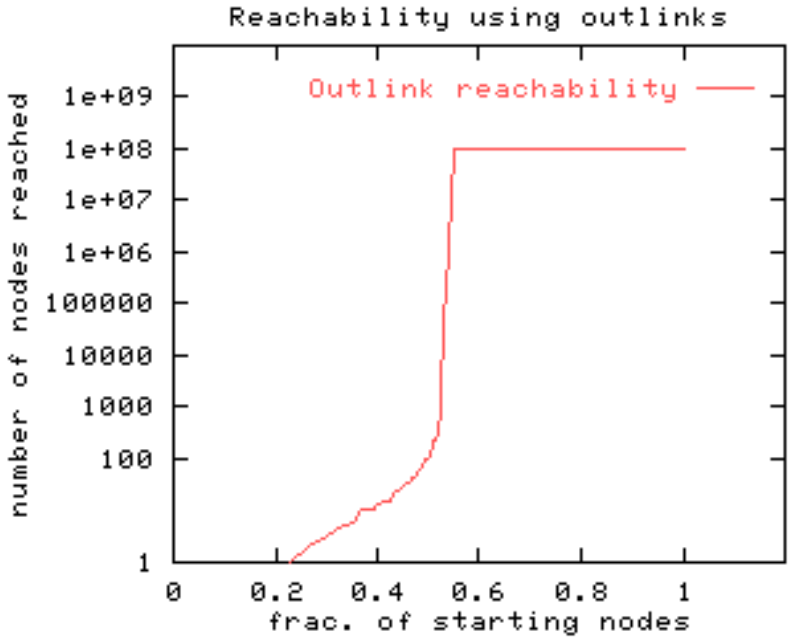


- Wo sind die restlichen 72% der Seiten?

Traversierungsmessung

- Breadth-first search (BFS): Von einem Knoten aus alle erreichbaren Knoten in Schichten nach Pfadlänge ordnen. Pfadlänge ∞ bei nicht erreichbaren Knoten
- BFS mit zufälligem Startknoten in beiden Richtungen:
  - Entweder: Ende des Algorithmus nach wenigen Knoten (<90 Knoten in 90% der Fälle)
  - Oder: Explosion zu einer Abdeckung von ca. 100m Knoten

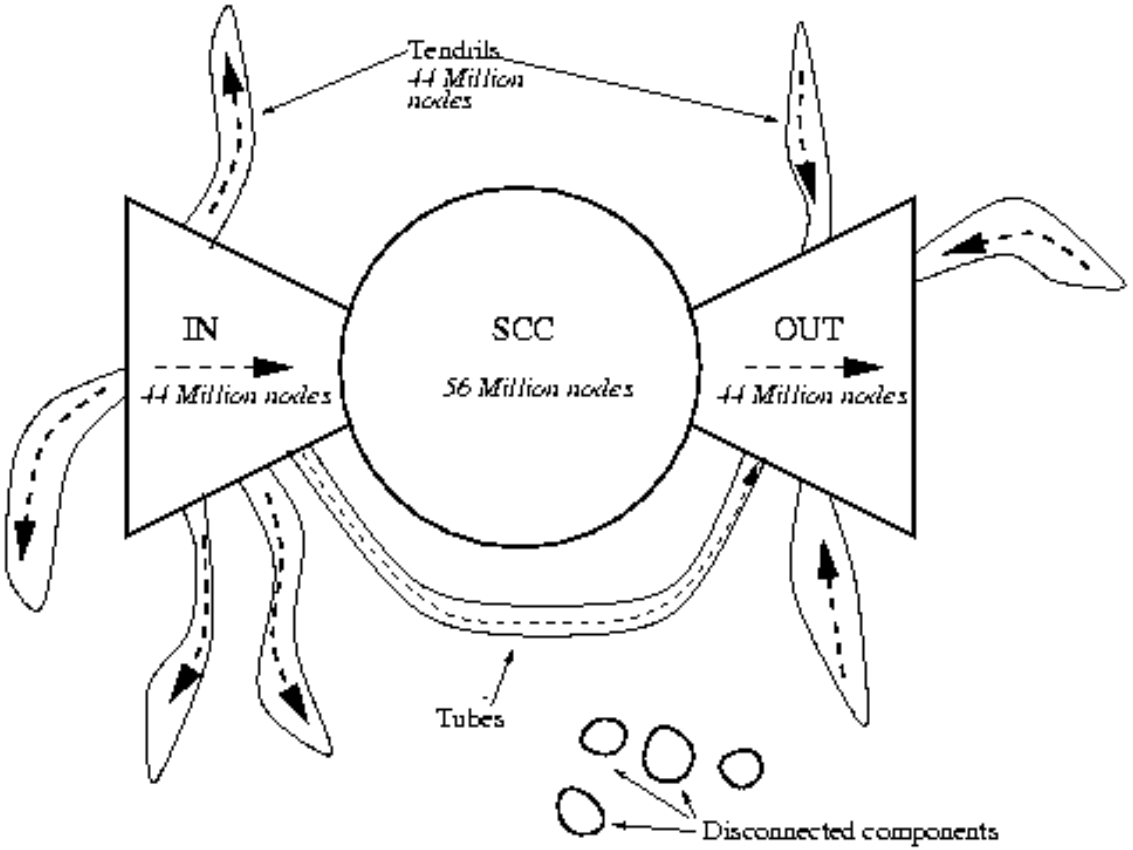




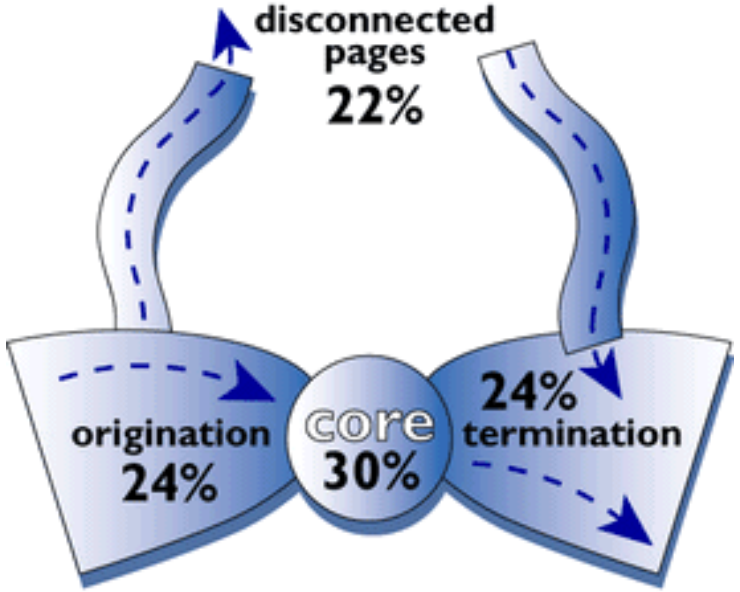
Ermittelte Struktur

- Startpunkte für BFS, die „vorwärts“ explodieren sind entweder in SCC oder in einer Menge IN
- IN: Es existiert für jeden Knoten ein Pfad nach SCC
- Startpunkte für BFS, die „rückwärts“ explodieren sind entweder in SCC oder in einer Menge OUT
- OUT: Es existiert für jeden Knoten ein Pfad von SCC
- Zusätzlich:
  - TENDRILS aus IN ohne SCC zu erreichen
  - TENDRILS nach OUT ohne aus SCC zu kommen
  - TUBES von IN nach OUT
  - DISCONNECTED ohne Verbindung

Struktur des Web



"Bow tie":



Region	SCC	IN	OUT	Tendrils	Disc.	Total
Grösse	56463993	43343168	43166185	43797944	16777756	203549046
Anteil	28%	21%	21%	22%	8%	100%

Weitere Maße

- Erreichbarkeit:
  - zwischen zwei zufällig gewählten Knoten existiert nur mit einer Wahrscheinlichkeit von 25% ein Pfad
- Durchmesser:
  - Durchmesser eines Graphen: Maximum aller kürzesten Pfade über alle Paare (u,v)
  - Durchmesser von SCC > 28
- Entfernungen:
  - Entfernung zwischen zwei Knoten ohne Berücksichtigung der Richtung von Links: 6,83
  - „Vorwärts“, entlang Out-links: 16,18
  - „Rückwärts“, entlang In-links: 16,12
  - Beides nur falls ein Pfad existiert (75% der Fälle nicht)

Crawling

Information Discovery

- Lynch, C. (1995). Networked Information Resource Discovery: An Overview of Current Issues (Invited paper). IEEE Journal on Selected Areas of Communications, 13(8):1505–1522:

  "information discovery is a complex collection of activities that can range from simply *locating a well-specified digital object on the network* through lengthy iterative research activities which involve the *identification of a set of potentially relevant networked information resources*, the *organization and ranking resources in this candidate set*, and the *repeated expansion or restriction of this set* based on characteristics of the identified resources and exploration of specific resources."

Web Information Discovery

- Das Web ist
  - Verteilt
  - Dezentral organisiert
  - Dynamisch

- Resource Discovery Problem:
  Wo sind Informationsquellen von Interesse

- Lösungsidee für das Web:
  - Automatisches Navigieren über Seiten
  - Indexierung der gefundenen Seiten
  - *Crawler* (auch *Spider*, *Robot*, *Worm* etc.)

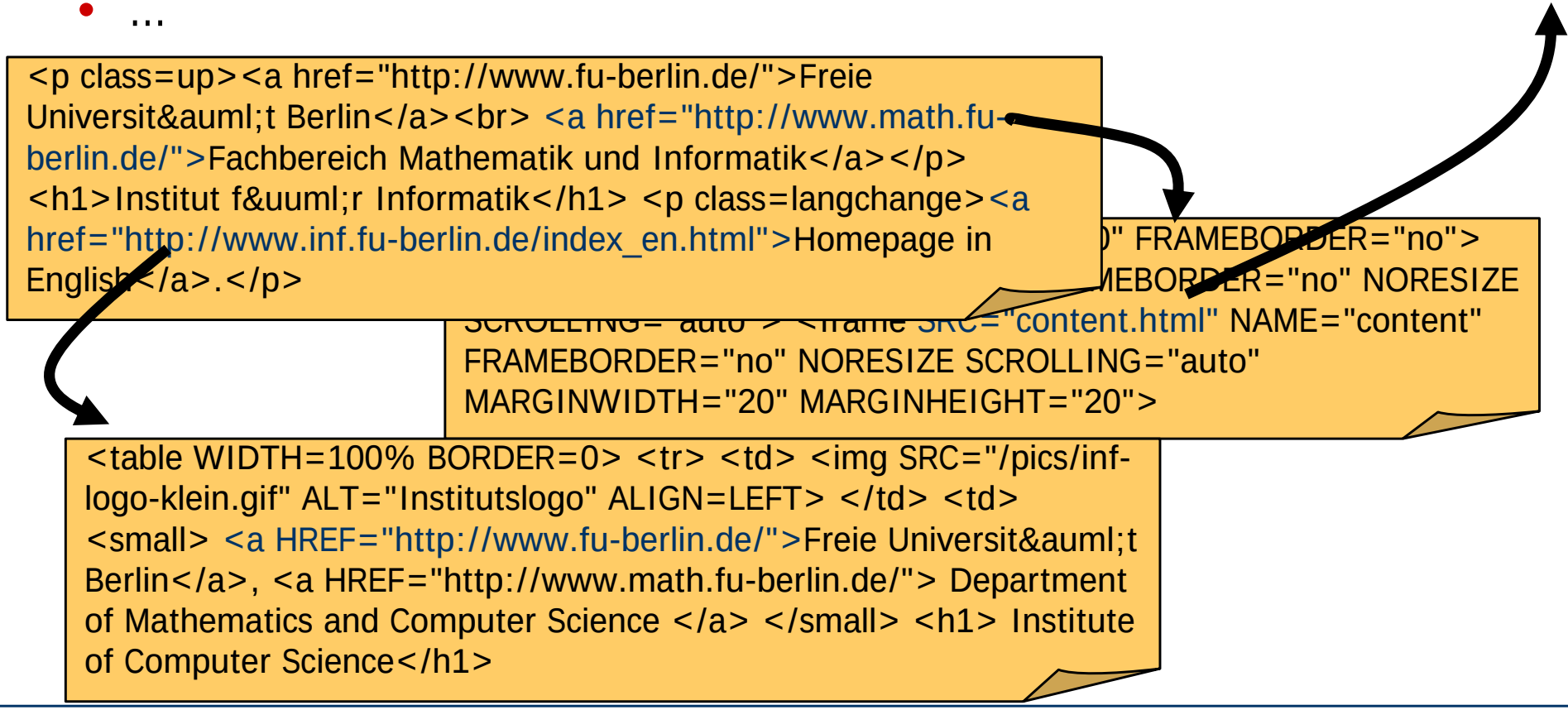
WebCrawler

- Eines der ersten Systeme: *WebCrawler* [Pinkerton94]
- Zwei Funktionen
  - Indexierung des Web
  - Automatische Navigation nach Bedarf
- WebCrawler in 94:
  - 50000 Dokumente von 9000 Quellen indexiert
  - 6000 Anfragen täglich
  - Updates wöchentlich
- Suchmaschinen 11/04: [Searchenginewatch.com]
- Google geschätzt 9/05: 24 Milliarden Seiten

Search Engine	**Reported Size**	**Page Depth**
Google	8.1 billion	101K
MSN	5.0 billion	150K
Yahoo	4.2 billion (estimate)	500K
Ask Jeeves	2.5 billion	101K+

Crawling Algorithmus

- Das Web als traversierbarer Graph von Seiten die über Links als Kanten verbunden sind
  - <a>, <link>, <meta>, <img>, <object>, <frameset>
  - FTP-Server, Adressen in nicht-HTML Dokumenten
  - ...

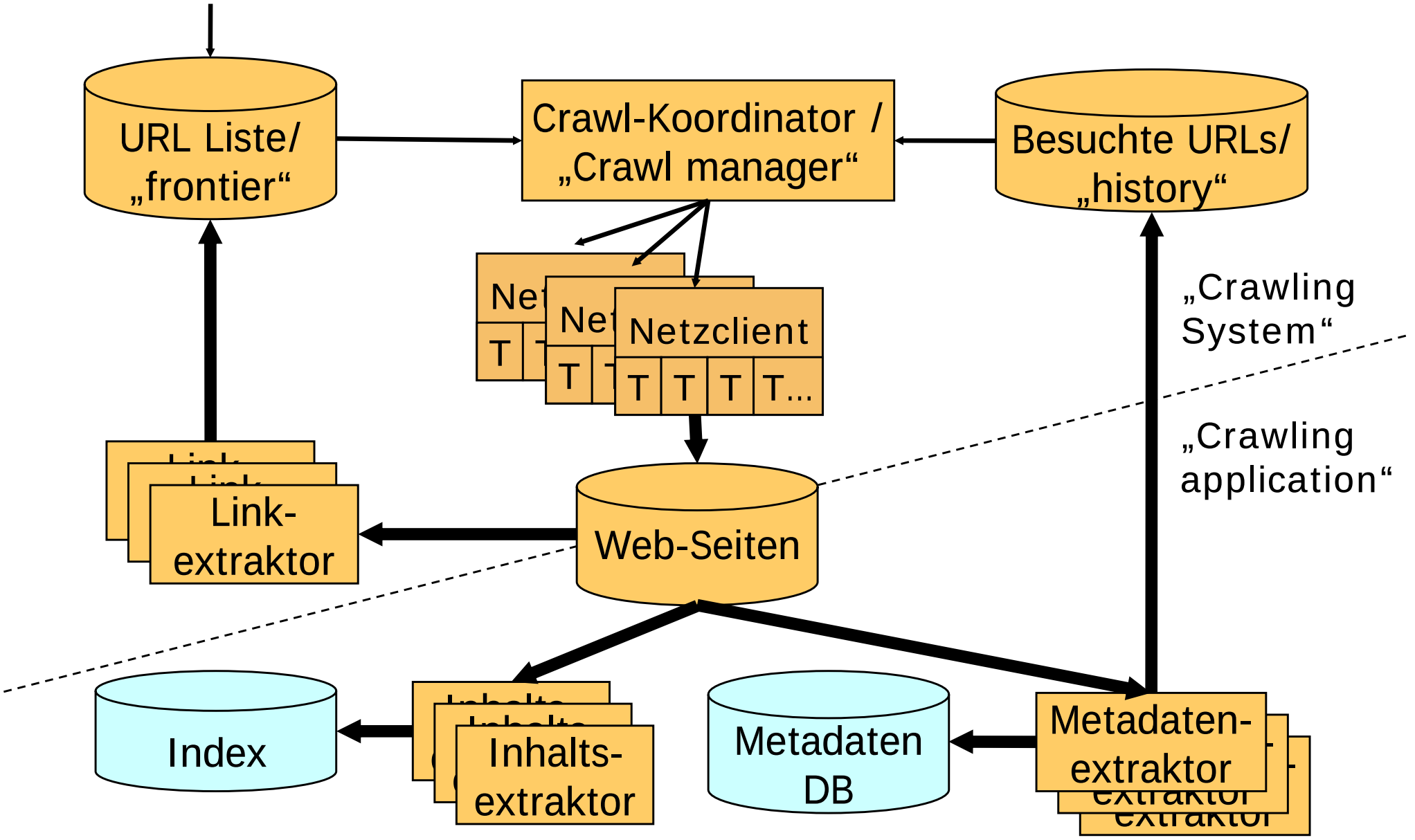


Crawling Algorithmus

1. URL-Liste mit unbesuchten URLs initial füllen
2. Nehme URL aus Liste und teste
   - schon besucht?
   - passender Medientyp (html/ps/pdf/gif/...)?
   - andere Kriterien (Ort/...)?
3. hole Seite
4. extrahiere URLs und
   schreibe sie in URL-Liste
5. extrahiere und indexiere Seiteninhalt
6. extrahiere und speichere Metadaten
7. gehe nach 2

„Crawling loop“

Einfache Architektur



Design Optionen / URL Liste

- URL-Liste / Frontier
  - Größe: Annahme: 7 Links pro Seite ->
    - Frontier wächst schnell
    - Frontier wird groß
  - Duplikate: Keine URLs doppelt
    - Serielle Suche teuer
    - Hash-Table mit URL als Schlüssel auch teuer

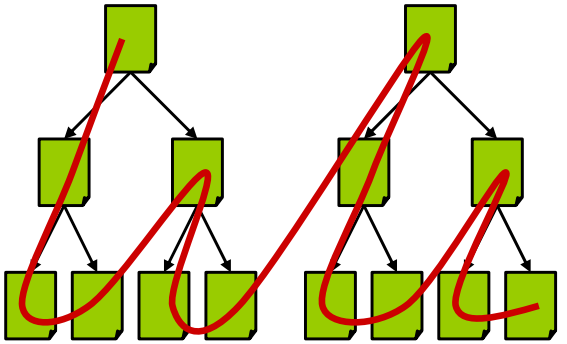
Design Optionen / Link Extraktion

Freie Universität Berlin

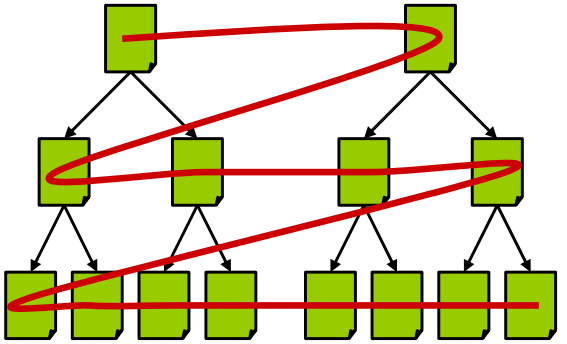
- Welche Links verfolgen?
  - <a>, <link>, <meta>, <img>, <object>, <frameset> etc.?
- Im Web notierte URLs sind gar keine eindeutigen Schlüssel -> URL Normalisierung notwendig
  - HTTP://www.UIOWA.edu -> http://www.uiowa.edu.
  - http://myspiders.biz.uiowa.edu/faq.html# -> http://myspiders.biz.uiowa.edu/faq.html
  - http://dollar.biz.uiowa.edu/%7Epant/ -> http://dollar.biz.uiowa.edu/~pant/
  - http://dollar.biz.uiowa.edu -> http://dollar.biz.uiowa.edu/
  - http://www.foo.com/index.html -> http://www.foo.com/
  - http://dollar.biz.uiowa.edu/~pant/BizIntel/Seeds/../Seeds.dat -> http://dollar.biz.uiowa.edu/~pant/BizIntel/Seeds.dat.
  - http://www.foo.com:80/ -> http://www.foo.com/
- Viele weitere möglich, Heuristiken auch andersherum gültig

Designoptionen / Entnahme/Erweiterung der URL-Liste

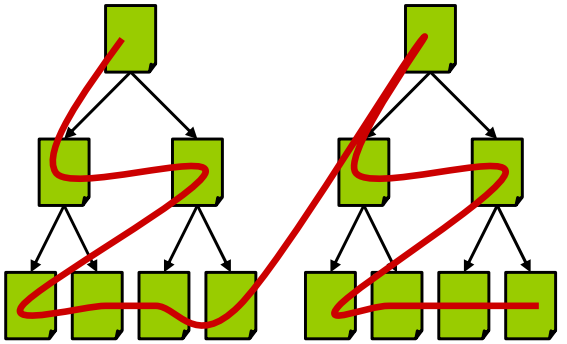
- Durch Ordnung der Frontier wird die Crawl-Strategie bestimmt
  - Depth-First
    "Enge" Suche in die Tiefe einzelner Sites



  - Breadth-First
    "Breite" Suche über viele Sites, übliches Vorgehen



  - Breadth-First pro Site,
    Nicht mehr beliebig, aber "breit" genug



Designoptionen / Entnahme/Erweiterung der URL-Liste

Freie Universität Berlin

- Best-first: Crawler versucht in „gute Richtung“ zu crawlen
  - Es gibt eine Vorgabe in Form einer Anfrage
  - Repräsentiert als Vektor von Termen
  - Crawler repräsentiert Seite als Vektor von Termen
  - Crawler ermittelt Ähnlichkeit der Vektoren
  - Alle auf der Seite gefundenen URLs erhalten Ähnlichkeit als Priorität
  - Frontier ist priorisierte Schlange
  - Crawl wird bei der nächsten „guten“ URL fortgesetzt
  - Weitere Prioritätsanhaltspunkte:
    - Entfernung von /
    - Angenommener Medientyp
    - Ankertext?

Designoptionen / Crawl-Koordinator

Freie Universität Berlin

- Crawl-Koordinator
  - Schon gesehen?
  - Eigenschaften der URL
    - aus .de?
  - Verarbeitbarer Filetyp?
    - HTML
    - PDF, Postscript, Word
    - Excel?
    - MP3?
  - Serverzugriff zurückstellen?
    - Kurz vorher schon zugegriffen?
    - Schon zu viel von Server geholt?
  - Koordination mit weiteren Crawlern bei
    - Nebenläufigkeit
    - Verteilung

Designoptionen / Crawl-Koordinator

Freie Universität Berlin

- Netzzugriffe
  - Wieviele Zugriffe parallel?
  - Welche Timeouts?
  - Umgang mit Fehlern
  - Verteilte Zugriffe?

- Erste Google-Versionen ca. 1997/8 (http://google.stanford.edu):
  - 3 Netzclients
  - je ca. 300 Verbindungen
  - mit 4 Clients ca. 100 Web Seiten/Minute crawlbar (144000/Tag, 6944 Tage für 1 Milliarde Seiten = 19 Jahre)
  - ca. 600Kb / Sekunde Netzlast

Designoptionen / Inhaltsextraktion

- Inhaltsextraktion
  - Welche Teile des Inhalts indexieren?
    - Überschriften
    - Nur Ankertexte
    - Titel
    - Gesamtdokument oder Teile davon?

Search Engine	Reported Size	Page Depth
Google	8.1 billion	101K
MSN	5.0 billion	150K
Yahoo	4.2 billion (estimate)	500K
Ask Jeeves	2.5 billion	101K+

Designoptionen / Metadaten

- Metadaten ermitteln
  - Welche Metadaten speichern?
    - Titel
    - Besucht
    - <meta> Tag
    - Klassifikation?
    - Wann besucht
    - Quersumme?

Diverse Probleme

- Framesets
- Unterschiedliche URLs für dieselbe Seite
  Sitzungs-IDs, dynamisch erzeugte Pfade
- Errechnete Links ("Next year" auf einem Kalender)
- Dynamische Seiteninhalte (Javascript etc.)
- Fehlerhafte Seiten
- Transportprobleme durch Netz
- Transportprobleme durch Größe

Crawling aus Server-Sicht

Crawler Last

- Crawler erzeugen Last beim Server
  - Verarbeitung der Anfragen
  - Auslieferung der Ergebnisse

- “Freundliche” Crawler versuchen das zu vermeiden
  - Keine fortlaufenden Anfragen zum Indexieren einer gesamten Site auf einen Schlag
  - Beachtung des Robot Exclusion Protokolls
  - Beachtung der <meta>-Tags zum Steuern von Robotern

Robots Exclusion Protokol

- Definiert einen Mechanismus mit dem ein Server festlegt, ob er von einem Crawler besucht werden will
- Daten /robots.txt auf Server
- http://www.inf.fu-berlin.de/robots.txt:

```
# robots.txt for http://www.inf.fu-berlin.de/
User-agent: *
Disallow:   /tec/net/
Disallow:   /tec/rechner/
Disallow:   /tec/software/packages/
Disallow:   /cgi-bin/
User-agent: MOMspider/1.00
Disallow:   /cgi-bin/
Disallow:   /tec/software/packages/
```

robots.txt

- `User-agent:` bezeichnet den Roboter, für die die folgenden Regeln gelten sollen
  - Namen wie (s. http://www.robotstxt.org/wc/active.html)
    - `Googlebot`
    - `Grapnel/0.01 Experiment`
    - `InfoSeek Robot 1.0`
  - Platzhalter * für alle Roboter
- Bezeichnet jeweils einen Teil der Dokumentenraums, der nicht besucht werden soll
  - Eintrag
    ```
    Disallow:   /tec/net/
    ```
  - `http://www.inf.fu-berlin.de/tec/net` soll nicht besucht werden

robots.txt

- Alle Roboter ausschließen:
```
User-agent: *
Disallow: /
```
- Einzelne Roboter ausschließen:
```
User-agent: Roverdog
Disallow: /
```
- Einzelne Seiten schützen:
```
User-agent: googlebot
Disallow: cheese.htm
```
- Nur einen Crawler zulassen:
```
User-agent: WebCrawler
Disallow:
User-agent: *
Disallow: /
```

<meta>-Element

- Das HTML <meta>-Tag kann ebenfalls zur Roboter-Steuerung genutzt werden

```
<html>
 <head>
  <meta name="robots"
        content="noindex,nofollow">
  <title>...</title>
 </head>
```

- Verbreitung bei Robots unklar

<meta>-Element

- `index`: Diese Seite soll indexiert werden
- `noindex`: Diese Seite soll nicht indexiert werden
- `follow`: Die Links dieser Seite weiterverfolgen
- `nofollow`: Die Links dieser Seite nicht weiterverfolgen
- `all = index,follow`
- `none = noindex,nofollow`

- Keine Möglichkeit, Verhalten für bestimmte Crawler zu bestimmen
- Kein Zugriff auf `robots.txt` notwendig

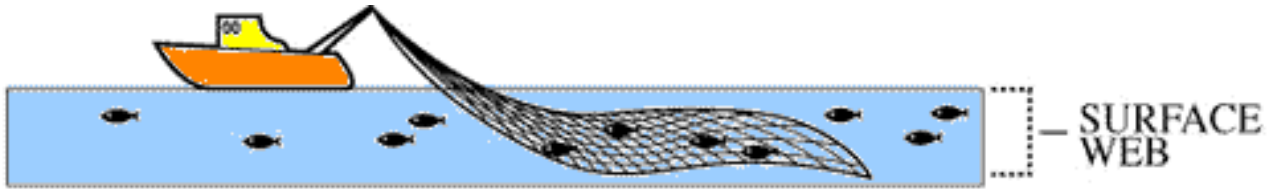
Das "Deep Web"

Michael K. Bergman. The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing August, 2001. Volume 7, Issue 1 und http://www.brightplanet.com/resources/details/deepweb.html

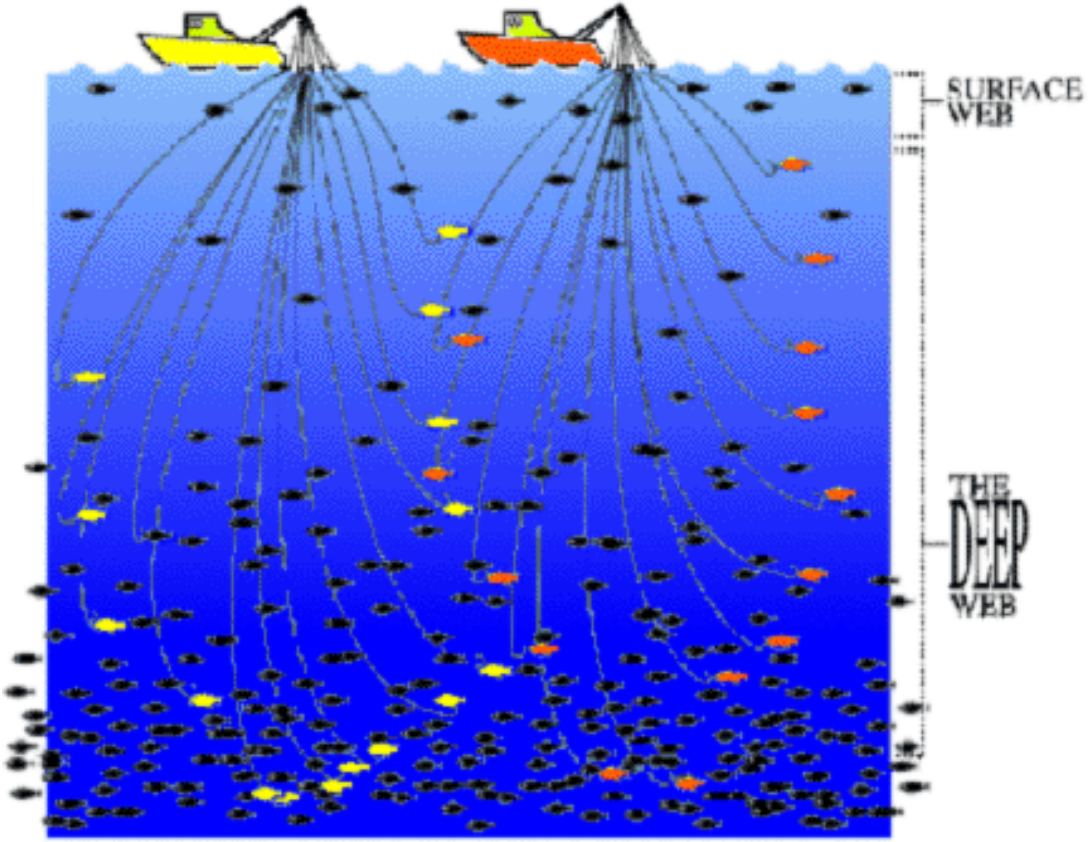
He, B., Patel, M., Zhang, Z., and Chang, K. C. 2007. Accessing the deep web. *Commun. ACM* 50, 5 (May. 2007), 94-101. DOI= http://doi.acm.org/10.1145/1230819.1241670

"Deep Web"-Argumentation

- Traversierung des Web über Links führt nur zu einem Bruchteil der Informationen



- "Deep Web" wird von Datenbankinhalten gebildet
- Umfang 400-500 mal größer als "normales" Web
- 500 Mrd Dokumente vs. 1 Mrd Dokumente
- Zugriff aber nur durch Datenbank-anfragen möglich



Deep Web Studie

- 100 Sites analysiert
  - Schätzung der enthaltenen Datensätze oder Dokumente
  - Abfrage von Stichprobe von 10 Dokumenten zu Größenabschätzung durch Mittelwertbildung
  - Indexierung und Klassifizierung des Suchformulars
- Größenschätzung
  - Nachfrage bei Betreibern
  - Aussagen auf Site
  - Aussagen über Site in anderen Berichte
  - Zahlen bei Suchantworten, z.B. Treffer für "NOT sfgjsljffjd"
  - Ausschluss aus Untersuchung
- Schätzung: Durchschnittlich 74,4 MB pro Site

Größenschätzung Sites des Deep Web

Freie Universität Berlin

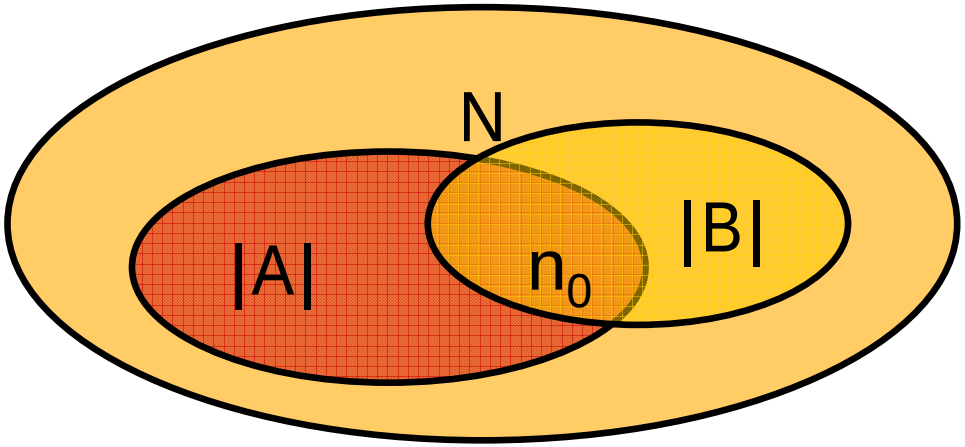
Name	Type	Web Size (GBs)
National Climatic Data Center (NOAA)	Public	366,000
NASA EOSDIS	Public	219,600
National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	32,940
Alexa	Public (partial)	15,860
...	...	...
Subtotal Public and Mixed Sources		**673,035**
DBT Online	Fee	30,500
Lexis-Nexis	Fee	12,200
Dialog	Fee	10,980
Genealogy - ancestry.com	Fee	6,500
ProQuest Direct (incl. Digital Vault)	Fee	3,172
...	...	...
Subtotal Fee-Based Sources		**75.469**
Total		**748,504**

Anzahl von Sites des Deep Web

Freie Universität Berlin

- Manuell und teilweise automatisch unterstützt:
  - 53220 URL-Hinweise aus anderen Sites
  - 45732 ohne Duplikate
  - 43348 noch zugängige
  - 17579 anscheinend suchbare
  - 13,6% davon nicht suchbar

Overlap analysis: Gesucht N - Größe des Deep Web



- n_A, n_B Abdeckung durch je eine Suchmaschine / ein Verzeichnis
- n_0 Überlappung
- |A|, |B|: Größe von A, B
- p(A): Wahrscheinlichkeit, Seite von A gefunden wird
- $p(A \cap B) = p(A) * p(B)$
- $|A| = N * p(A)$, $|B| = N * p(B)$, $|A \cap B| = N * p(A \cap B)$

- $N = |A| * |B| / |A \cap B|$

- Da Verzeichnisse nicht zufällig: Untere Grenze

Schätzung Anzahl der Sites

					DB A			Tot Est Deep Web
DB A	**A no dups**	**DB B**	**B no dups**	**A+ B**	**Uniq.**	**DB Fract.**	**DB Size**	**Sites**
Lycos	5,081	Internets	3,449	256	4,825	0.074	5,081	68,455
Lycos	5,081	Infomine	2,969	156	4,925	0.053	5,081	96,702
Internets	3,449	Infomine	2,969	234	3,215	0.079	3,449	43,761

- Schätzung: Ca. 100000 Deep Web Sites

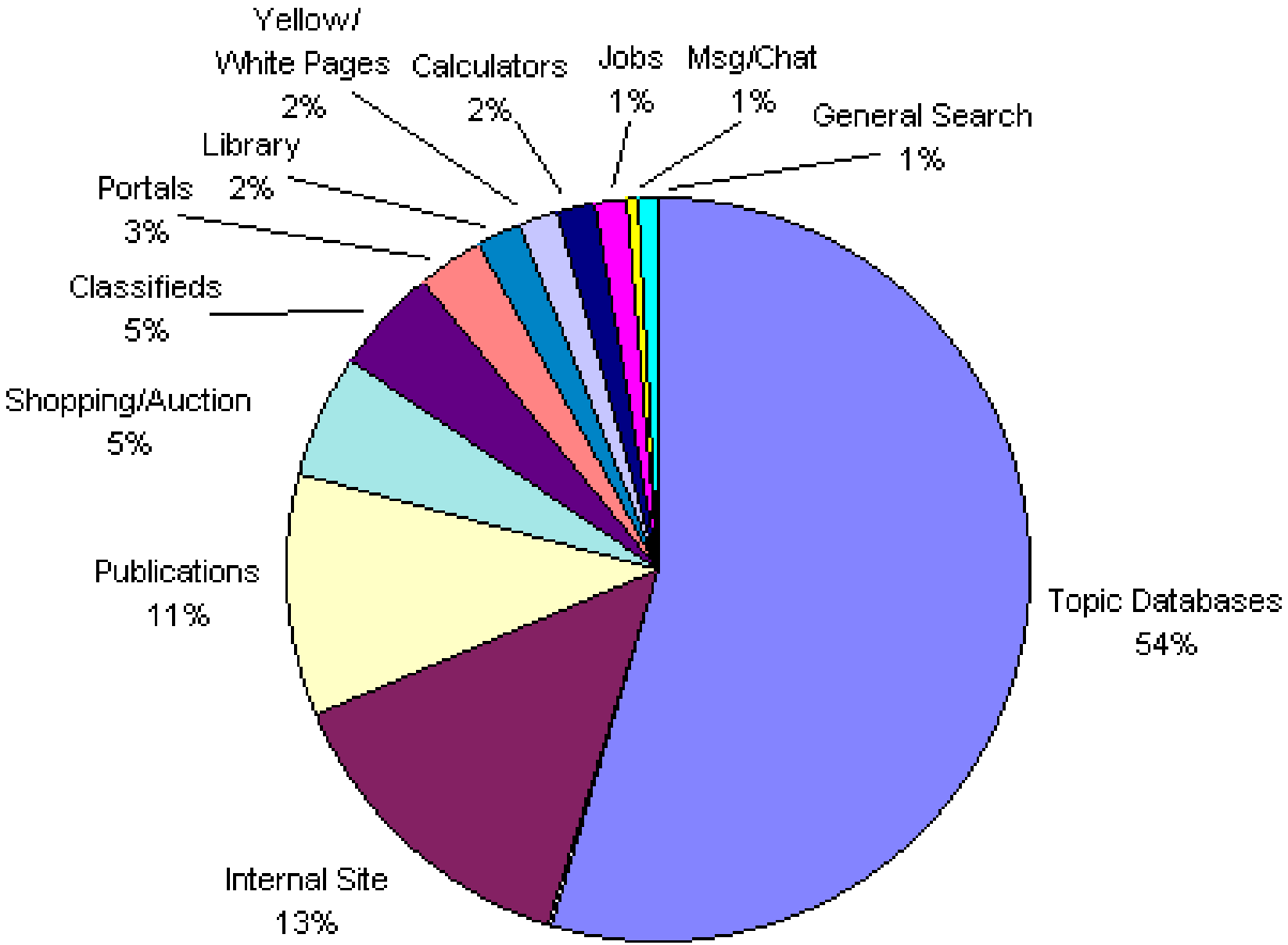
Inhaltsanalyse

Freie Universität Berlin

- Inhaltsüberprüfung durch Anfragen aus 20 Gebieten
- Typanalyse durch Handauswertung von 700 Sites

Agriculture	2.7%	Law/Politics	3.9%
Arts	6.6%	Lifestyles	4.0%
Business	5.9%	News, Media	12.2%
Computing/Web	6.9%	People, Companies	4.9%
Education	4.3%	Recreation, Sports	3.5%
Employment	4.1%	References	4.5%
Engineering	3.1%	Science, Math	4.0%
Government	3.9%	Travel	3.4%
Health	5.5%	Shopping	3.2%
Humanities	13.5%	Law/Politics	3.9%

Site-Klassifikation



Vergleiche

- Deep Web: 7500 Terabytes, Web: 19 Terabytes
- Deep Web: 550 Mrd Docs, Web: 1 Mrd Docs
- Mehr Traffic auf Deep Web Sites (50%)
- Mehr Wachstum im Deep Web
- Deep Web Sites mehr inhaltliche Tiefe und weniger inhaltliche Breite
- 95% des Deep Web frei zugänglich

- Probleme:
  - Intention der Deep Web Studie
  - Erschließung?

Erschließung des Deep Web

- He/Patel/Szang/Chang: Überlappungsanalyse geht von Unabhängigkeit zwischen Indizes der Suchmaschinen aus
  - Das ist aber nicht gegeben
  - -> Deep Web Größe ist unterschätzt

- Vorgehen
  - 1000000 IP-Nummer auswählen
  - Auf Web-Server testen
  - Suchfelder ermitteln
  - Def. Deep Web Server: Server der über ein Suchformular Datenbankinhalte herausgibt

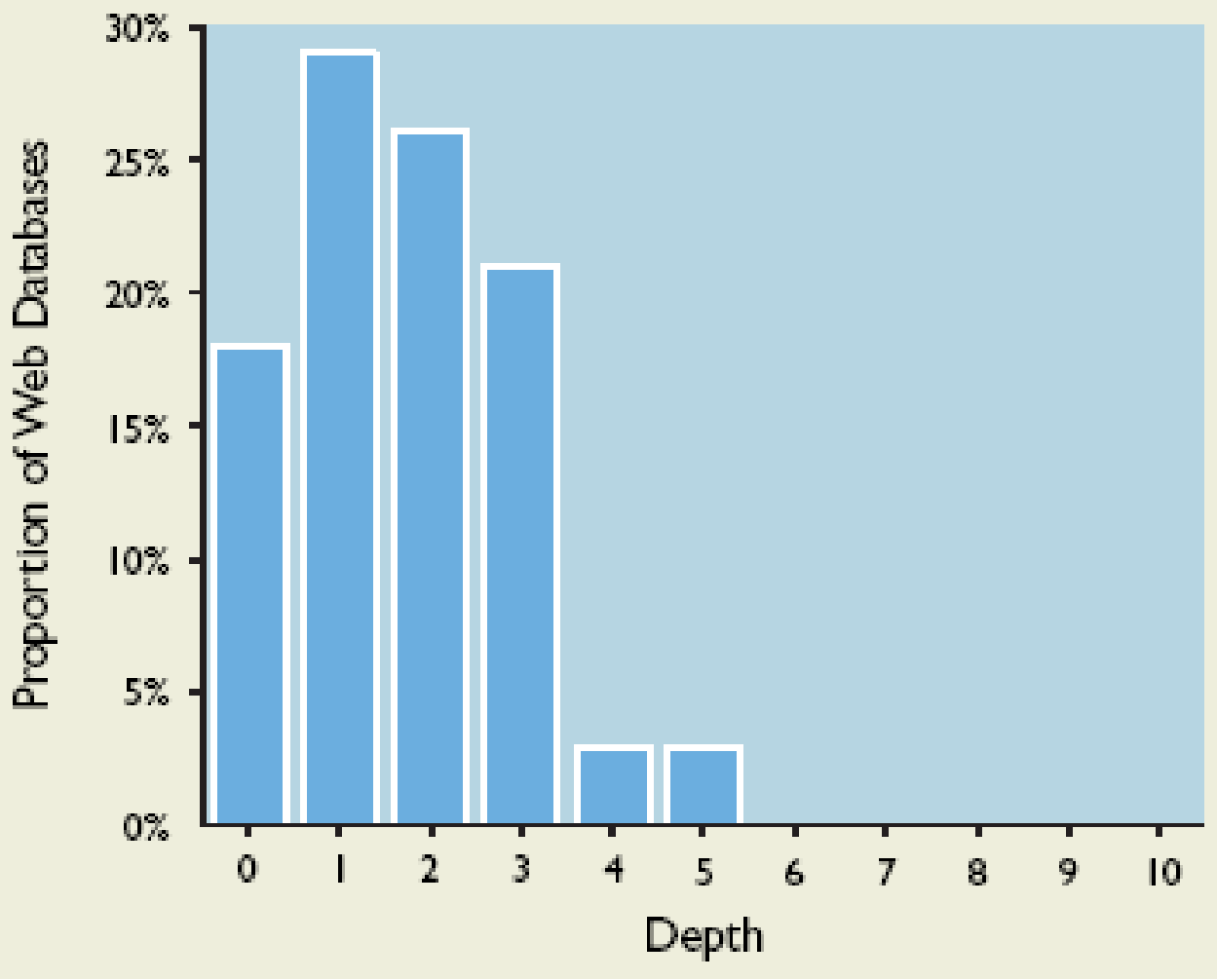
- #Suchformulare->#Datenbanken->#Deep Web Server

- Duplikate ausschließen
  - Suchfelder für „site search“, „login“ etc. herausnehmen
  - Formulare mit gleichem Ziel herausnehmen
  - Durch zufällige Anfragen gleiche Datenbanken ermitteln

[alle folgenden Abbildungen aus HePatelZhangChang2007]

Ergebnisse

- Wo befinden sich die Suchformulare des Deep Web?
  - 100000 IP Nummern in Tiefe untersucht



Ergebnisse

- Aus 1000000 IP Nummern 2256 Web Server ermittelt
- Davon 126 Deep Web Sites
- Mit 406 Suchformularen zu 190 Datenbanken

- Internet (IPv4) Adressraum = 2230124544 Nummern
- Hochrechnung aus Tiefenuntersuchung
  - 307000 Deep Web Sites
  - 450000 Datenbanken
  - 1258000 Suchformulare

- Vgl: 43000-96000 Deep Web Sites in Brightplanet Studie

Ergebnisse

- Abdeckung durch Suchmaschinen
  - Aus Datenbanken Ergebnisobjekte ermitteln
  - In Suchmaschinen anfragen
- Abdeckung durch Suchmaschinen ca. 1/3:
  - Google, Yahoo: 32%
  - MSN: 11%
  - Große Überlappung
- Abdeckung durch Deep Web Verzeichnisse: Gering

Verzeichnis	#	Abdeckung
completeplanet.com	70000	15,6%
lii.org	14000	3,1%
turbo10.com	2300	0,5%
invisible-web.net	1000	0,2%

Literatur

- Brian Pinkerton. Finding What People Want: Experiences with the WebCrawler. Second International World-Wide Web Conference: Mosaic and the Web, Chicago, IL, October 17--20 1994. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/pinkerton/WebCrawler.html
- G. Pant, P. Srinivasan, and F. Menczer. Crawling the Web. In M. Levene and A. Poulovassilis, editors, Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer-Verlag, 2004. http://citeseer.ist.psu.edu/579280.html
- www.searchenginewatch.com
- The Web Robots Pages. www.robotstxt.org