# Peer-to-Peer-Systeme

## Teil III: Zufallsgraphen, kleine Welten und skalenfreie Netze

Björn Scheuermann

Humboldt-Universität zu Berlin
Wintersemester 2015/16

# Milgrams Small-World-Experiment

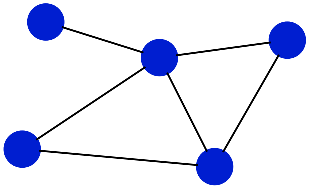
- 1967 untersuchte der Psychologe Stanley Milgram die Struktur sozialer Netzwerke
- Er führte folgendes Experiment durch:
  - zufällig ausgewählte Personen erhielten einen Brief, den sie einer ihnen nicht persönlich bekannten Zielperson zukommen lassen sollten
  - über die Zielperson waren einige Hintergrundinformationen verfügbar (Name, Beruf, Wohnort,. . . )
  - der Brief durfte nur über persönlich bekannte Kontakte weitergegeben werden, mit dem Ziel, der Zielperson „näher“ zu kommen
- Viele Briefe erreichten ihr Ziel in nicht mehr als sechs Schritten (und das 1967!)
- Wie lässt sich diese überraschende Eigenschaft sozialer Netzwerke verstehen?

# Netzwerke als Graphen

- Netzwerke (im allgemeinsten Sinne!) lassen sich als Graphen beschreiben
- Ein *Graph* besteht aus einer Menge von *Knoten V* und einer Menge von *Kanten E*

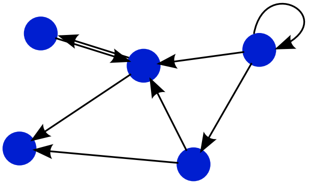
*Ungerichteter Graph*

$E \subseteq \{\{v_1, v_2\} \mid v_1, v_2 \in V, v_1 \neq v_2\}$



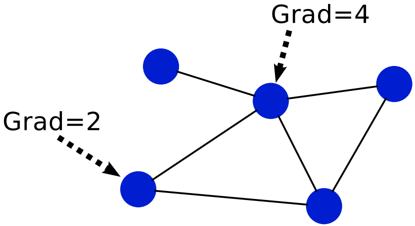
*Gerichteter Graph*

$E \subseteq \{(v_1, v_2) \mid v_1, v_2 \in V\}$



# Graphen: Knotengrad

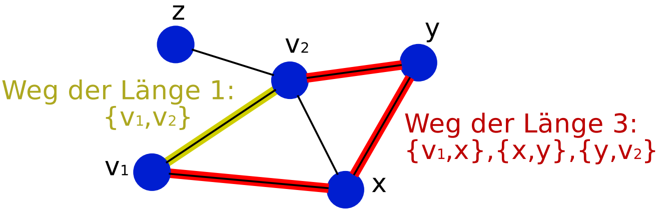
- Der *Grad* eines Knotens ist die Zahl der Kanten, an denen er beteiligt ist




- Bei gerichteten Graphen: *Eingrad* (Zahl der eingehenden Kanten) und *Ausgrad* (Zahl der ausgehenden Kanten)

# Graphen: Wege und Distanzen

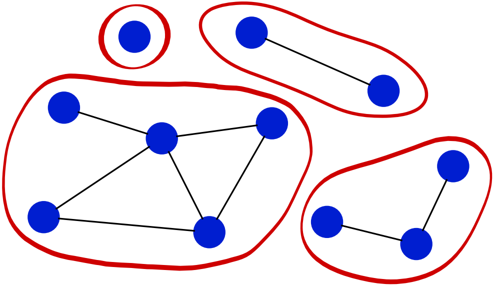
- Ein *Weg* oder *Pfad* von $v_1$ nach $v_2$ in einem Graphen ist eine Folge von Kanten, die von $v_1$ nach $v_2$ führt
- Die *Länge* eines Weges ist die Zahl der Kanten auf dem Weg




- Die *Distanz* zwischen $v_1$ und $v_2$ in einem Graphen ist die Länge des kürzesten Weges von $v_1$ nach $v_2$

# Graphen: Zusammenhang

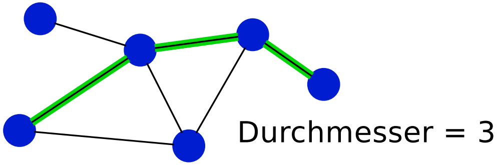
- Eine *Zusammenhangskomponente* eines Graphen ist eine Teilmenge $C \subseteq V$, so dass
  - von jedem Knoten in $C$ zu jedem anderen Knoten in $C$ ein Weg existiert und
  - keine andere Teilmenge existiert, für die das ebenfalls gilt und die $C$ vollständig enthält



- Ein Graph ist *zusammenhängend*, wenn seine Knotenmenge $V$ eine Zusammenhangskomponente ist

# Graphen: Durchmesser und Durchschnitts-Pfadlänge

- Der *Durchmesser* eines Graphen ist die längste Distanz zwischen zwei Knoten des Graphen




- Die *durchschnittliche Pfadlänge* ist die durchschnittliche Distanz zwischen zwei zufällig gewählten Knoten eines Graphen

# Zufallsgraphenmodelle

Es gibt zwei gängige Modelle für Zufallsgraphen:

1. Erdős und Rényi:

   $\mathcal{G}_{n,m}$ sei die Menge aller (ungerichteten) Graphen mit $n$ Knoten und $m$ Kanten

2. Gilbert:

   $\mathcal{G}_{n,p}$ sei die Menge aller (ungerichteten) Graphen mit $n$ Knoten, in denen jede mögliche Kante unabhängig mit Wahrscheinlichkeit $p$ vorhanden ist

Für $m \sim p \cdot N$ (wobei $N = \binom{n}{2} = \frac{n \cdot (n-1)}{2}$ die Zahl der möglichen Kanten in einem ungerichteten Graphen mit $n$ Knoten ist) sind die Modelle praktisch äquivalent.

Generiere mithilfe einer Münze als „Zufallsgenerator“ einen Gilbert-Zufallsgraphen aus $\mathcal{G}_{6,\,0.5}$

Wieviele Zusammenhangskomponenten hat dein Graph? Was ist der Durchmesser der größten Zusammenhangskomponente?

# Eigenschaften von Zufallsgraphen

- Wir können keine (sinnvollen) Aussagen über *alle* Graphen in $\mathcal{G}_{n,m}$ oder $\mathcal{G}_{n,p}$ machen, sondern nur über *erwartete* Eigenschaften eines zufällig gewählten Graphen $G_{n,m} \in \mathcal{G}_{n,m}$ bzw. $G_{n,p} \in \mathcal{G}_{n,p}$
- Die Graphen aus $\mathcal{G}_{n,p}$ weisen eine Eigenschaft *mit hoher Wahrscheinlichkeit* (m. h. W.) auf, wenn

  $$\lim_{n \to \infty} P(G_{n,p} \in \mathcal{G}_{n,p} \text{ hat die geforderte Eigenschaft}) = 1$$

  (analog auch für andere Graphenklassen)
- Für Erdős-Rényi-Zufallsgraphen aus $\mathcal{G}_{n,m}$ ist das i. d. R. nur sinnvoll, wenn $m$ eine Funktion von $n$ ist

# Zusammenhang von Zufallsgraphen

Sei
$$m_\gamma = \frac{n}{2}(\log n + \gamma),$$
wobei $\gamma = \gamma(n)$ eine Funktion von $n$ ist. Dann gilt

- wenn $\lim_{n\to\infty} \gamma = -\infty$, dann ist ein typisches $G_{n,m_\gamma}$ nicht zusammenhängend
- wenn $\lim_{n\to\infty} \gamma = \infty$, dann ein typisches $G_{n,m_\gamma}$ zusammenhängend

Typische Interpretation: Wenn der durchschnittliche Knotengrad in einem Zufallsgraphen $\Omega(\log n)$ ist, dann ist der Graph mit hoher Wahrscheinlichkeit zusammenhängend.

# Giant Component

Sei $c > 0$, $p = \dfrac{c}{n}$

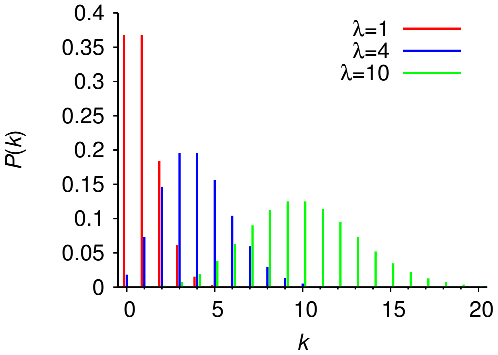
- Für $c < 1$ hat m. h. W. jede Zusammenhangskomponente von $G_{n,p}$ die Größenordnung $O(\log n)$.
- Für $c > 1$ gibt es m. h. W. eine Zusammenhangskomponente mit Größe $\Theta(n)$ („Giant Component"), andere Komponenten haben die Größe $O(\log n)$.

Beobachtung: Die Giant Component entsteht m. h. W., wenn der durchschnittliche Grad der Knoten *eins* übersteigt!

# Gradverteilung in Zufallsgraphen

In Zufallsgraphen folgt die Gradverteilung der Knoten (asymptotisch) einer Poisson-Verteilung:

$$P(\text{Grad } k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



# Eigenschaften realer Netze: Watts + Strogatz

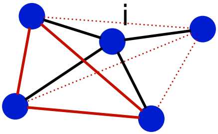
- Zufallsgraphen sind analytisch gut zugänglich und wurden lange zur Modellierung vieler realer Netzwerke eingesetzt
- Aber wie akkurat geben sie die Eigenschaften realer Netze wieder?
- Watts und Strogatz haben 1998 die Eigenschaften einer Reihe von realen Netzwerken untersucht und mit Zufallsgraphen mit gleicher Knoten- und Kantenzahl verglichen:
  - Zusammenarbeit von Filmschauspielern (Kante = gemeinsamer Film)
  - Neuronales Netz des Fadenwurms C. elegans
  - Stromnetz der westlichen USA
- Vergleich anhand von:
  - durchschnittlicher Pfadlänge
  - Clustering-Koeffizient

# Clustering-Koeffizient

Sei $i$ ein Knoten in einem ungerichteten Graphen, $d(i)$ der Grad von $i$ und $E(i)$ die Zahl der Kanten zwischen Nachbarn von $i$.

Der *Clustering-Koeffizient* $C(i)$ von $i$ ist

$$C(i) = \frac{E(i)}{\binom{d(i)}{2}} = \frac{\text{Zahl der Kanten zwischen Nachbarn von } i}{\text{Zahl der } \textit{möglichen} \text{ Kanten zwischen Nachbarn von } i}$$




— Kante von i

— existierende Kante zwischen Nachbarn von i

⋯ mögliche, aber nicht vorhandene Kante

C(i) = 3/6 = 1/2

(Lässt sich analog für gerichtete Graphen definieren.)

Suche in deine zuvor erzeugten Zufallsgraphen den Knoten mit dem höchsten Grad; bestimme den Clustering-Koeffizienten dieses Knotens.

Entspricht der Clustering-Koeffizient dem Wert, den du für einen Graphen aus $\mathcal{G}_{6,\,0.5}$ erwarten würdest?

# Clustering-Koeffizient

- Der Clustering-Koeffizient ist die Wahrscheinlichkeit, dass zwei Nachbarn von $i$ wiederum Nachbarn sind
- In $\mathcal{G}_{n,p}$ ist deshalb ein Clustering-Koeffizient von $p$ zu erwarten
- Der Clustering-Koeffizient eines Graphen ist der durchschnittliche Clustering-Koeffizient seiner Knoten
- Vorsicht: Der Clustering-Koeffizient für einen Knoten mit $d(i) < 2$ ist nicht definiert!
  - wird dann manchmal $= 0$, manchmal $= 1$ gesetzt, oder solche Knoten werden in der Auswertung ignoriert

# Eigenschaften realer Netze: Watts + Strogatz

Wir vergleichen jetzt die von Watts und Strogatz untersuchten Netze mit Zufallsgraphen mit gleicher Knoten- und Kantenzahl

Welches Ergebnis würdest du beim Vergleich der Clustering-Koeffizienten etwarten?

Was erwartest du hinsichtlich der durchschnittlichen Pfadlänge?

# Eigenschaften realer Netze: Watts + Strogatz

| Netzwerk | $n$ | ∅-Grad | ∅-Pfadlänge real | ∅-Pfadlänge Zufall | $C$ real | $C$ Zufall |
|---|---|---|---|---|---|---|
| Schauspieler | 225 226 | 61 | 3.65 | 2.99 | 0.79 | 0.00027 |
| Stromnetz | 4 941 | 2.67 | 18.7 | 12.4 | 0.08 | 0.005 |
| C. elegans | 282 | 14 | 2.65 | 2.25 | 0.28 | 0.05 |

Beobachtung: Pfadlänge passt gut, aber reale Netze sind lokal sehr viel dichter (= hoher Clustering-Koeffizient)!

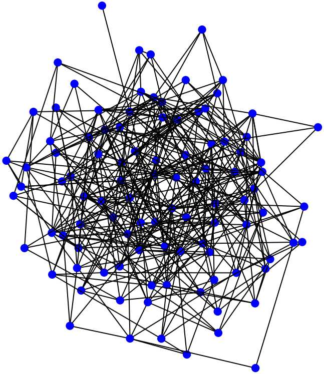
Ein hoher Clustering-Koeffizient lässt sich auch in sozialen Netzwerken beobachten: Unsere eigene „nahe Umgebung“ ist auch sehr viel besser „vernetzt“ als das in einem Zufallsgraphen der Fall wäre!

[Watts, Strogatz: Collective dynamics of ‘small-world’ networks, Nature, 1998]

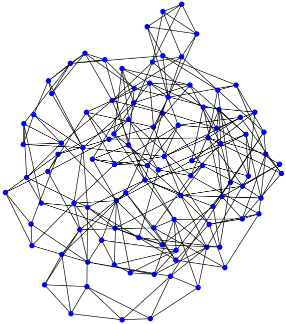
# Small Worlds

Eine *Small World* ist ein Netzwerk mit

1. kleiner durchschnittlicher Pfadlänge (wie in Zufallsgraphen) und
2. hohem Clustering-Koeffizienten (anders als in Zufallsgraphen!)



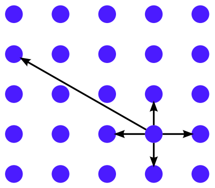
Zufallsgraph



Small World

# Kleinbergs Small-World-Graphen

- Konstruktionsprinzip von Kleinberg für Small-World-Graphen mit effizientem Routing:
  - ordne die Knoten in einem Gitter an
  - jeder Knoten ist mit seinen Nachbarn verbunden
  - jeder Knoten hat außerdem eine zufällige „Fernkante“
  - Wahrscheinlichkeit für Wahl des Ziels der Fernkante fällt wie $1/d^2$, wobei $d$ der Abstand in „Gitterschritten“ ist



- Erlaubt (bei richtiger Parameterwahl) Greedy-Routing zu Zielkoordinaten in $O(\log^2 n)$ Schritten
- Auch mehrdimensional möglich

[Kleinberg: The small-world phenomenon: An algorithmic perspective, STOC 2000]

# Kleinberg-Graphen für ein P2P-System?

- Kleinberg-Graphen erlauben effizientes Routing mit rein lokalem Wissen
- Aber für die *Konstruktion* ist globales Wissen notwendig:
  - woher bekommt ein Knoten seine (eindeutige) Position im Gitter?
  - wie findet er alle seine Nachbarn?
  - um eine Fernkante zufällig zu wählen, müssen *alle* entfernten Knoten bekannt sein
  - ...
- Deshalb taugen Kleinberg-Graphen nur bedingt als Basis für ein Peer-to-Peer-Overlay
- Es gibt aber verschiedene Ansätze, Small-World-Overlays mit effizienten Routing-Algorithmen gezielt zu erzeugen und in P2P-Overlays zu verwenden (hier nicht weiter besprochen)

# Gnutella ist eine Small World

- Unabhängige Untersuchungen aus mehreren Jahren zeigen, dass das Gnutella-Overlay Small-World-Eigenschaften hat
- Gilt für das ursprüngliche Gnutella (∼2001) ebenso wie für das spätere Ultrapeer-Overlay (∼2005)
- Typische Pfadlänge in Gnutella 2008: 4–5 Hops (!)
- Mögliche Ursachen für das beobachtete Clustering:
  - Bootstrapping-Mechanismen
  - Suche nach Peers für weitere Verbindungen
  - Dynamik des Overlays (Peers mit langer Uptime „sammeln“ Verbindungen zu anderen Peers mit ähnlichen Eigenschaften)
  - endgültig ist das nicht geklärt. . .

[Stutzbach, Rejaie, Sen: Characterizing Unstructured Overlay Topologies in Modern P2P File-Sharing Systems, Transactions on Networking, 2008]

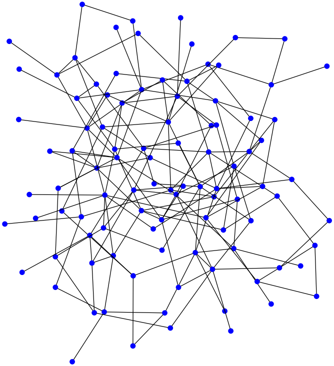
# Skalenfreie Netze

- Wie sieht die Gradverteilung in realen Netzwerken aus?
- In der AS-Topologie des Internet ist die Häufigkeit des Knotengrades $k$ proportional zu $k^{-\alpha}$ mit einer Konstanten $\alpha > 0$ („power law“)
- Auch der Link-Grad von Webseiten, das Stromnetz der USA und das Schauspieler-Zusammenarbeits-Netzwerk verhalten sich so
- Solche Netzwerke heißen *skalenfreie Netzwerke*

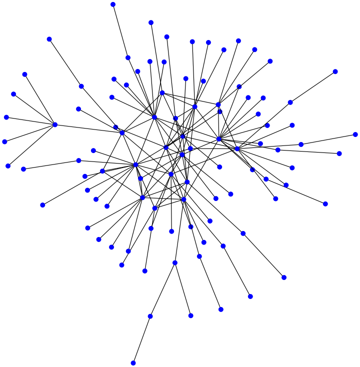
[Faloutsos, Faloutsos, Faloutsos: On Power-law Relationships of the Internet Topology, SIGCOMM 1999]

[Barabási, Albert: Emergence of Scaling in Random Networks, Science, 1999]

# Skalenfreie Netze



Zufallsgraph



Skalenfreies Netz

- Die meisten Knoten haben sehr niedrige Grade
- Es gibt wenige zentrale Knoten mit hohem Grad

# Pareto-Verteilung

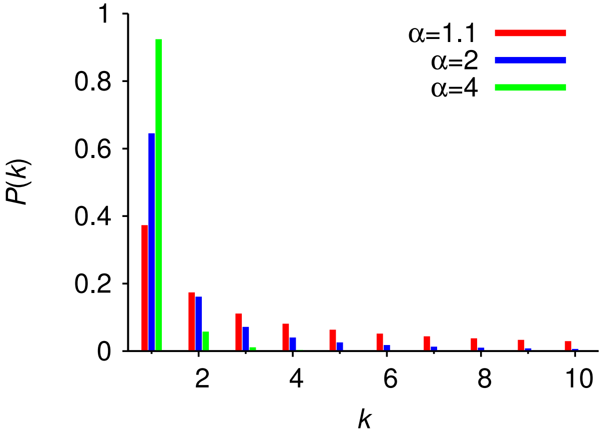
- Die entsprechende Wahrscheinlichkeitsverteilung ist die *diskrete Pareto-Verteilung* (für $k \geq 1$):

$$P(\text{Grad } k) = \frac{1}{\zeta(\alpha) k^{\alpha}} \qquad \zeta(\alpha) = \sum_{i=1}^{\infty} \frac{1}{i^{\alpha}}$$

- Heavy-Tail-Eigenschaft: Im Vergleich zur Poisson-Verteilung treten große Knotengrade mit hoher Wahrscheinlichkeit auf
- (Definition passt nicht ganz für $\alpha \leq 1$, da dann $\zeta(\alpha) = \infty$.
  Für einen Graphen mit endlicher Zahl von Knoten kann man den Normalisierungsfaktor $\zeta$ aber einfach entsprechend anpassen.)

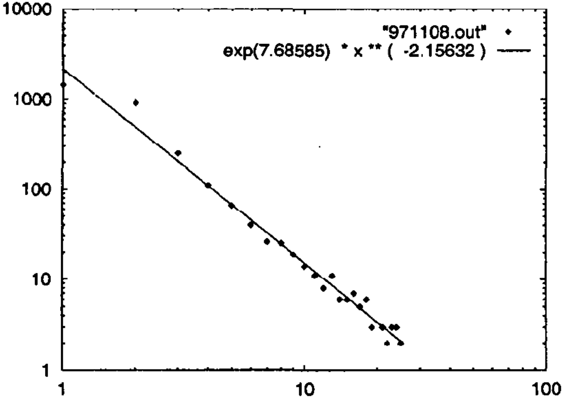
# Pareto-Verteilung

$$P(\text{Grad } k) = \frac{1}{\zeta(\alpha) k^{\alpha}} \qquad \zeta(\alpha) = \sum_{i=1}^{\infty} \frac{1}{i^{\alpha}}$$



# Gradverteilung im Internet

Inter-AS-Topologie des Internet von 1997, Grad vs. Häufigkeit;
„Power Law“ ergibt linear fallende Kurve im Log-Log-Diagramm:




(Abbildung: [Faloutsos et al. 1999])

# Eigenschaften skalenfreier Netzwerke

Welche Aussagen können wir über ein Netzwerk alleine auf Basis der Tatsache machen, dass es skalenfrei ist?

Für große Pareto-Graphen gilt m. h. W.:

- $\alpha < 1 \Rightarrow$ der Graph ist zusammenhängend
- $\alpha > 1 \Rightarrow$ der Graph ist nicht zusammenhängend
- $1 < \alpha < 2 \Rightarrow$ eine Giant Component ($\Theta(n)$), sonstige Komponenten $O(1)$
- $2 < \alpha < 3{,}4785\ldots \Rightarrow$ eine Giant Component ($\Theta(n)$), sonstige Komponenten $O(\log n)$
- $\alpha > 3{,}4785\ldots \Rightarrow$ keine Giant Component

[Aiello, Chung, Lu: A Random Graph Model for Power Law Graphs, Experimental Mathematics, 2001]

# Rich gets richer

- Skalenfreie Netzwerke entstehen, wenn neu hinzukommende Knoten sich bevorzugt mit existierenden Knoten mit hohem Grad verbinden

$\Rightarrow$ „Rich gets richer“

- Konkrete Bedingung: Wahrscheinlichkeit, eine Verbindung zu einem existierenden Knoten aufzubauen, ist proportional zur Anzahl der Verbindungen, die dieser Knoten bereits hat
- Tatsächlich ist dies die *einzige* Bedingung für das Entstehen von skalenfreien Netzwerken

[Albert, Barabási: Statistical Mechanics of Complex Networks, Reviews of Modern Physics, 2002]

# Robustheit skalenfreier Netze

- In skalenfreien Netzen haben nur wenige Knoten sehr hohe Grade
  - (…obwohl die Wahrscheinlichkeit für hohe Grade wie erwähnt asymptotisch höher ist als bei poissonverteilten Graden in Zufallsgraphen – dort gibt es praktisch gar keine Knoten mit hohem Grad!)
- Zufällige Ausfälle treffen wahrscheinlich Knoten mit geringem Grad, die für den Zusammenhang des Netzwerks nicht wichtig sind

$\Rightarrow$ Skalenfreie Netzwerke sind robust gegenüber Ausfällen

- 2,5 % Ausfälle ändern den Internet-Durchmesser kaum
- Zufallsgraphen zerfallen sehr viel schneller in einzelne Komponenten

# Robustheit skalenfreier Netze

Größenordnung der größten Zusammenhangskomponente nach dem zufälligen Entfernen von Knoten aus einem 10 000-Knoten-Graphen:

| entfernte Knoten | Zufallsgraph | skalenfreies Netz |
|---:|---:|---:|
| 5 % | 9 000 | 9 500 |
| 28 % | 100 | 7 000 |
| 45 % | 1 | 5 000 |

Beim gezielten Entfernen der Knoten mit dem höchsten Grad:

| entfernte Knoten | Zufallsgraph | skalenfreies Netz |
|---:|---:|---:|
| 5 % | 9 000 | 8 500 |
| 20 % | 6 000 | 1 |
| 30 % | 1 | 1 |

[Albert, Jeong, Barabási: The Internet's Achilles' Heel: Error and attack tolerance of complex networks, Nature, 2000]

# Robustheit skalenfreier Netze

- Umgekehrt machen also die wenigen zentrale Knoten ein skalenfreies Netzwerk anfälliger für *gezielte* Angriffe!
- Man kann dem Netz großen Schaden zufügen, indem man gezielt Knoten mit hohem Grad ausfindig macht und lahmlegt

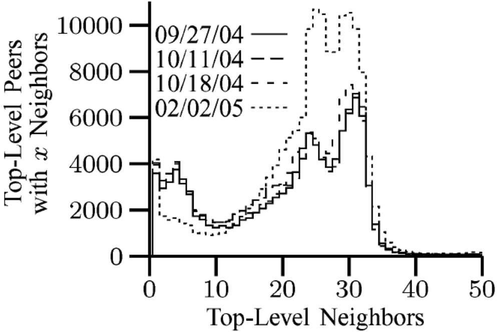
# Ist Gnutella skalenfrei?

- Untersuchungen des ursprünglichen Gnutella-Overlays legten Skalenfreiheit nahe
- Betrachte den *ursprünglichen* Mechanismus, mit dem neue Gnutella-Servents dem Overlay beitreten
  - „Gut verbundene“ Knoten haben eine höhere Wahrscheinlichkeit, beim Ping-Pong gefunden zu werden

$\Rightarrow$ Rich gets richer!

# Ist Gnutella skalenfrei?

Neuere Ergebnisse zeigen, dass die Grade im *späteren* Ultrapeer-Overlay *nicht* Pareto-verteilt sind:




(Abbildung: [Stutzbach et al. 2008])

# Ist Gnutella skalenfrei?

- Dass keine Pareto-Verteilung vorliegt ist eigentlich schon aufgrund der Strategie verbreiteter Ultrapeer-Implementationen (LimeWire, Bearshare) klar
- Die Ergebnisse in den älteren Studien *könnten* durch problematische Untersuchungsmethoden entstanden sein
- Erkenntnis: Schon das Sammeln der Rohdaten für die Untersuchung realer Systeme ist ein schwieriges Forschungsproblem!

[Stutzbach, Rejaie, Duffield, Sen, Willinger: On Unbiased Sampling for Unstructured Peer-to-Peer Networks, Transactions on Networking, 2009]

# Ist Gnutella skalenfrei?

- Die Robustheit des Gnutella-Ultrapeer-Overlays ist bedeutend besser als die eines skalenfreien Netzwerks
- Das Ultrapeer-Overlay widersteht sowohl zufälligen Ausfällen als auch gezielten Angriffen gut:
  - 85 % zufällige Ausfälle $\Rightarrow$ 90 % der verbleibenden Ultrapeers sind weiterhin verbunden
  - 50 % (!) der Peers mit hohem Grad entfernt $\Rightarrow$ 75 % der verbleibenden Ultrapeers sind weiterhin verbunden

[Stutzbach et al. 2008]

# Zusammenfassung

- In diesem Kapitel haben wir uns mit Graphenstrukturen und den sich daraus ergebenden Eigenschaften für entsprechende Netzwerke beschäftigt
- Wir haben zunächst Zufallsgraphen eingeführt und uns dann mit Eigenschaften realer Netzwerke beschäftigt, die sich wesentlich von denen von Zufallsgraphen unterscheiden
- Insbesondere haben wir *Small-World-Graphen* und *skalenfreie Netze* kennen gelernt und deren Eigenschaften am Beispiel des Gnutella-Overlays diskutiert