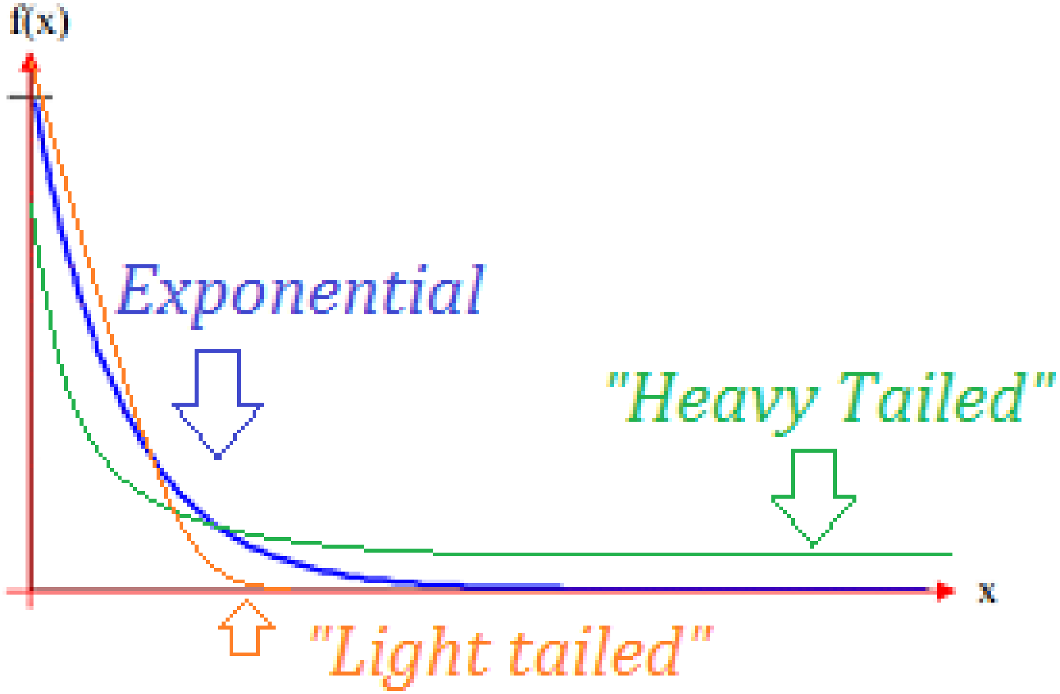


Power-laws, heavy tails, and rich-gets richer (things often observed in large-scale systems such as the internet …)

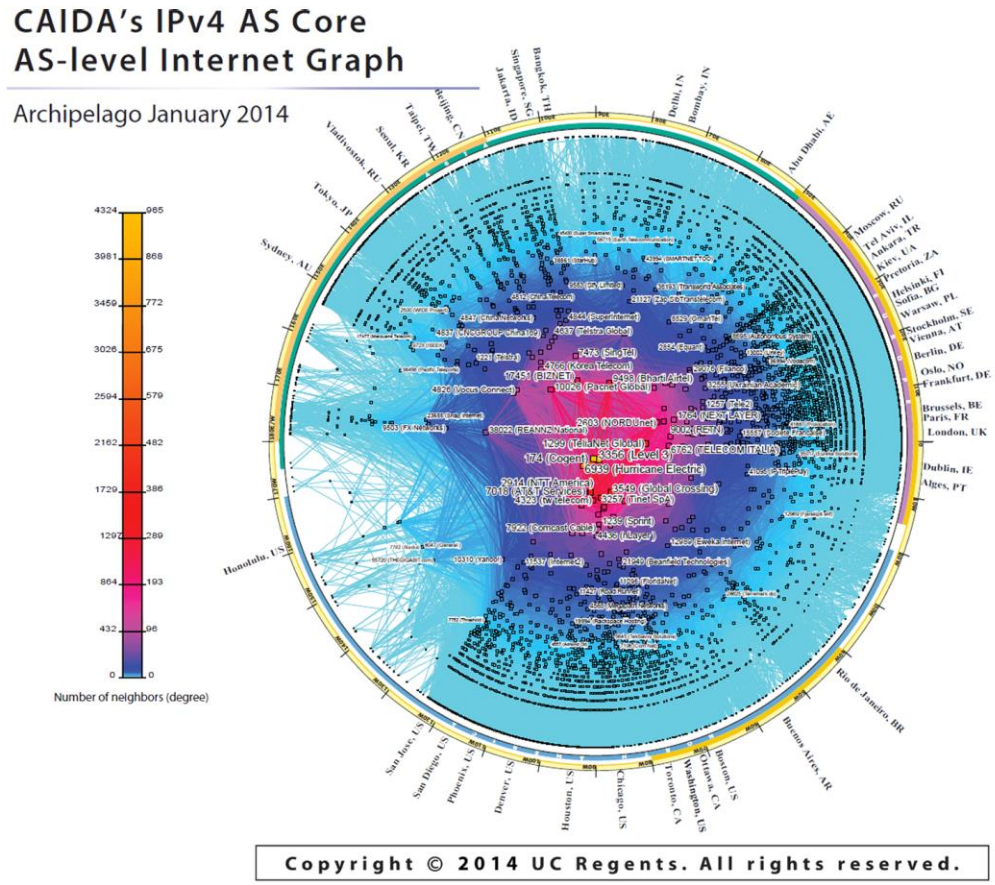
Slides by Niklas Carlsson

…

Things we often see in LARGE systems

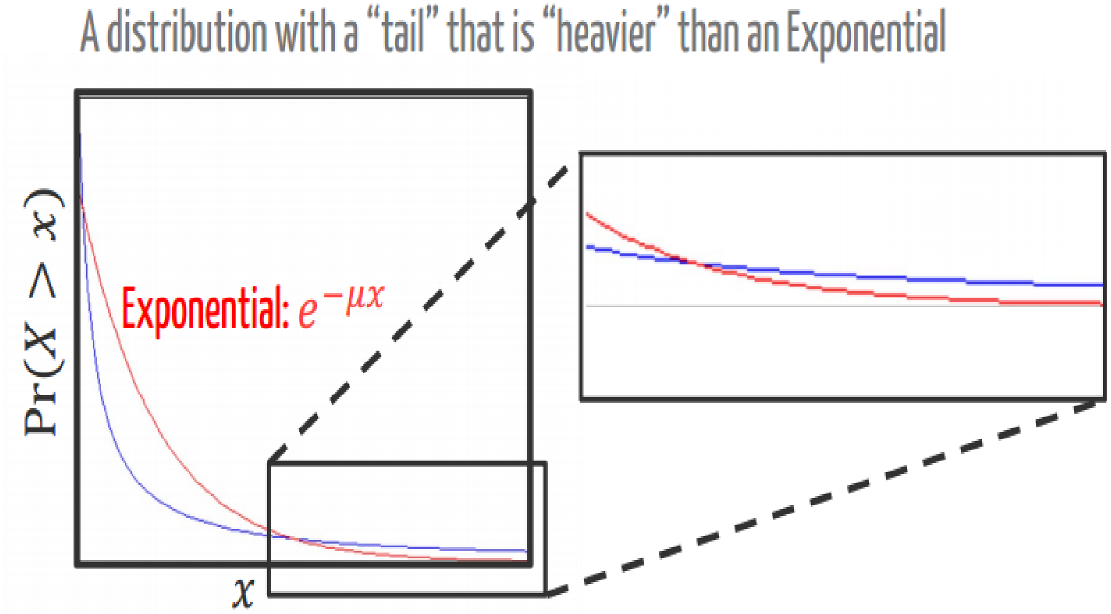
- Power laws, heavy tails, and skewed distributions in general
- Preferential attachment ("Rich gets richer")

First, example from last lecture



- Examples questions one may ask:
  - What degree distribution does this graph have?
  - And what implications does that have?

Heavy-tail distributions …



- "A probability distribution is said to have a heavy tail if the tail is not exponentially bounded"
  - E.g., paper and references therein: "A Tale of the Tails: Power-laws in Internet Measurements", IEEE Network, Mahanti et al., 2013
- Power-law, Pareto, Zipf (in some sense the same)
- … and then there are many other "heavy tail" distributions, variations and generalizations, including distributions such as log-normal, various generalized Zipf/Pareto distributions, etc.

Examples of power laws

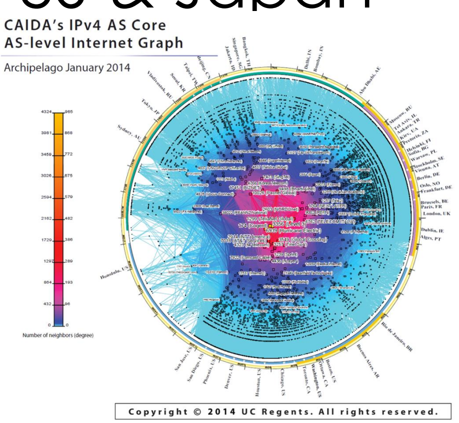
a. Word frequency: Estoup.
b. Citations of scientific papers: Price.
c. Web hits: Adamic and Huberman
d. Copies of books sold.
e. Diameter of moon craters: Neukum & Ivanov.
f. Intensity of solar flares: Lu and Hamilton.
g. Intensity of wars: Small and Singer.
h. Wealth of the richest people.
i. Frequencies of family names: e.g. US & Japan not Korea.
j. Populations of cities.

… AND many many more …

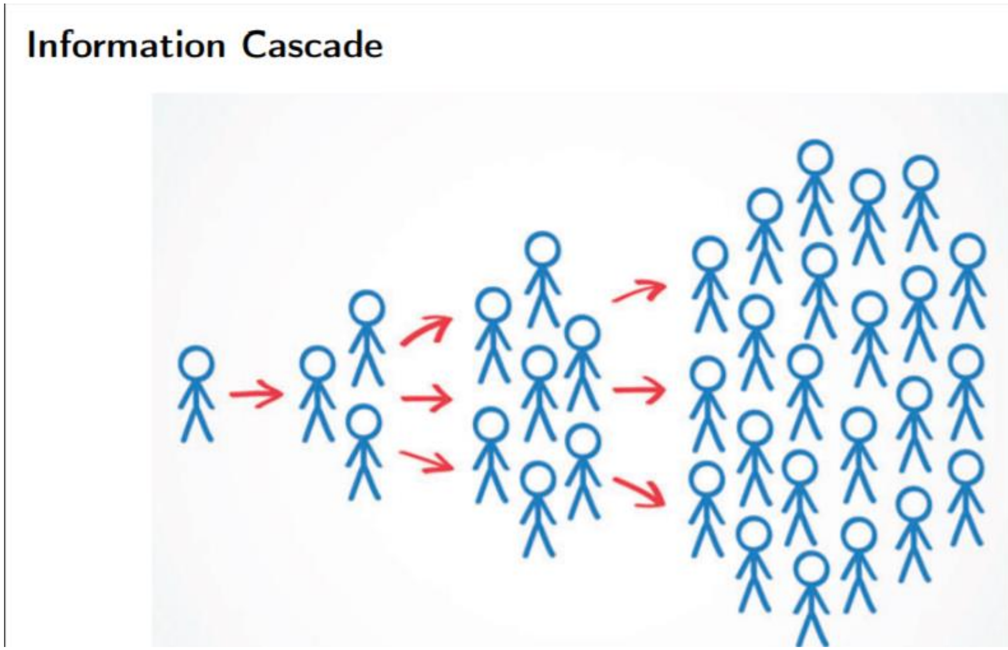
Examples of power laws

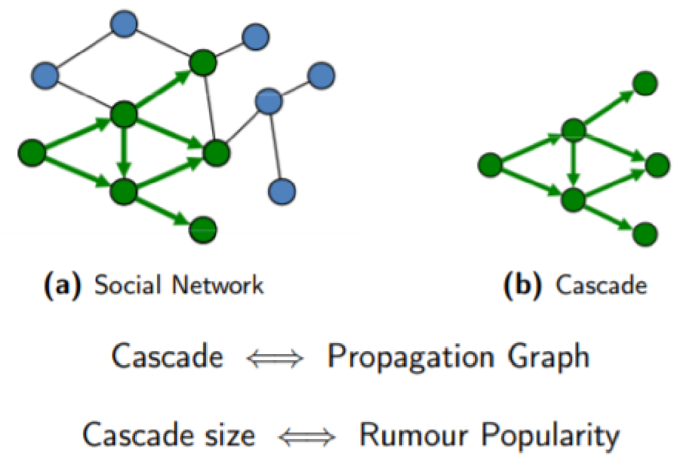
a. Word frequency: Estoup.
b. Citations of scientific papers: Price.
c. Web hits: Adamic and Huberman
d. Copies of books sold.
e. Diameter of moon craters: Neukum & Ivanov.
f. Intensity of solar flares: Lu and Hamilton.
g. Intensity of wars: Small and Singer.
h. Wealth of the richest people.
i. Frequencies of family names: e.g. US & Japan not Korea.
j. Populations of cities.

… AND many many more …



Or an even more timely example





A.Pacuk et al., Why Do Cascade Sizes Follow a Power-Law?, Proc. WWW 2017.

- The size of information cascaded, spread of fake news, and virus reach for that matter ...

File popularity distribution and "heavy" tails

- Example slides with YouTube popularity
  - but web object popularity, file size distributions, number of friends in social networks, etc. often see similar "heavy tail" distributions ...
  - This list can be made very very long, and include things such as the frequency words are used, the size of cities, the size of earthquakes, the size of bacteria cultures … and the list will go on ... and on … and on ...

Motivation

- Video dissemination (e.g., YouTube) can have wide-spread impacts on opinions, thoughts, and cultures

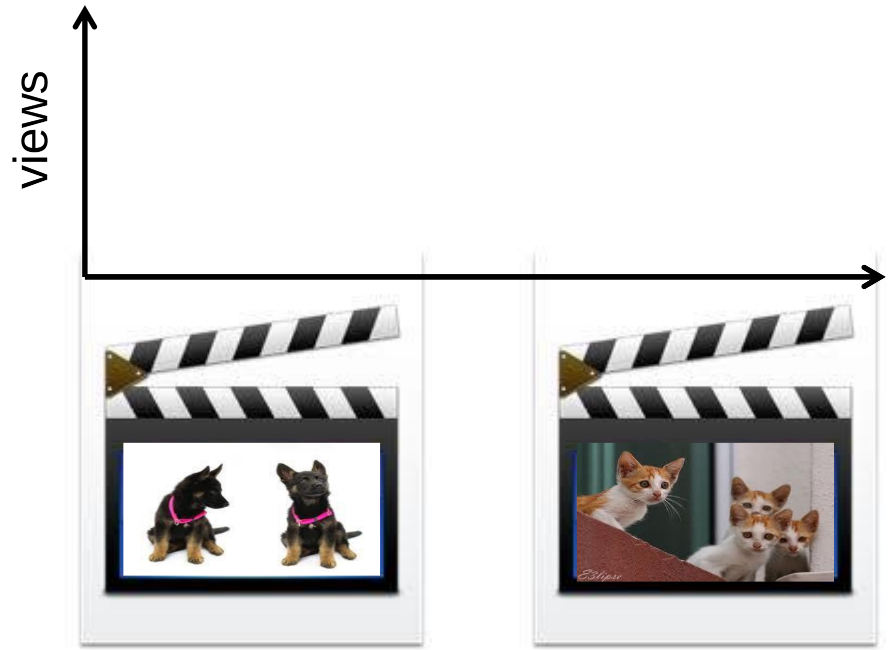
E.g., ACM KDD ’12, IFIP Performance ‘11, ACM TWEB

Motivation

- Not all videos will reach the same popularity and have the same impact

E.g., ACM KDD '12, IFIP Performance '11, ACM TWEB

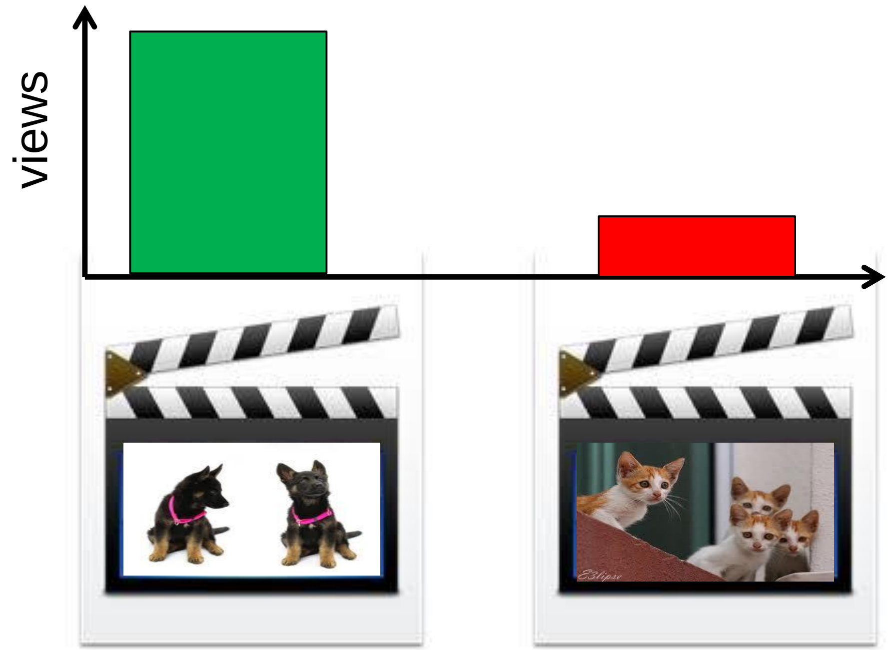
Motivation



- Not all videos will reach the same popularity and have the same impact

E.g., ACM KDD ’12, IFIP Performance ‘11, ACM TWEB

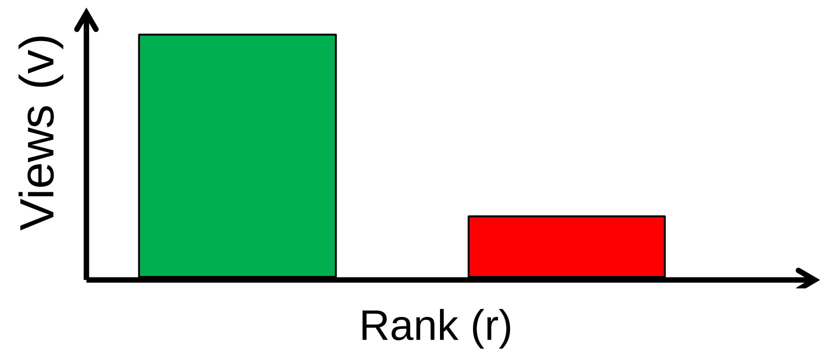
Motivation



- Not all videos will reach the same popularity and have the same impact

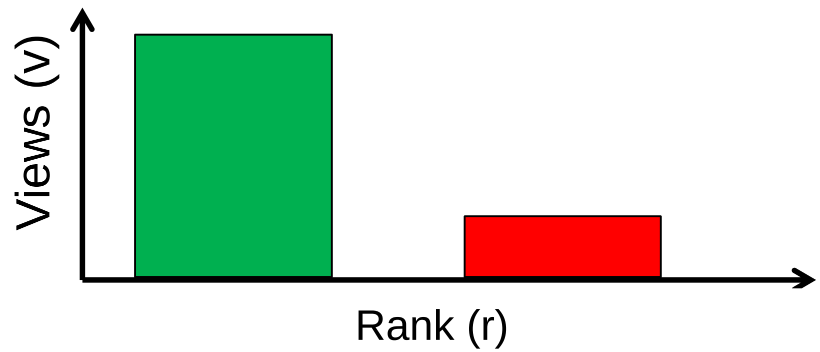
E.g., ACM KDD '12, IFIP Performance '11, ACM TWEB

Popularity distribution



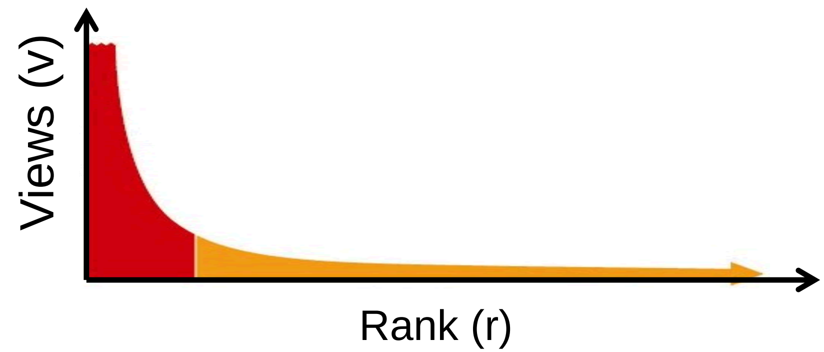
E.g., ACM KDD ‘12

Popularity distribution

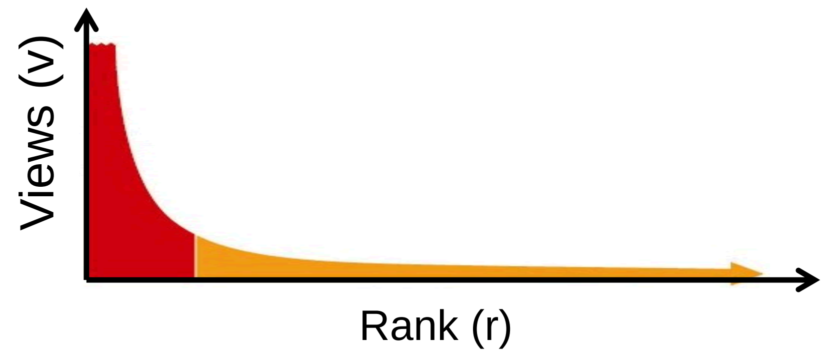


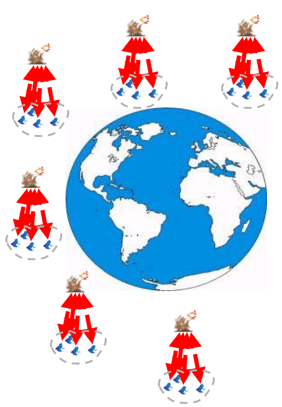
E.g., ACM KDD ‘12

Popularity distribution

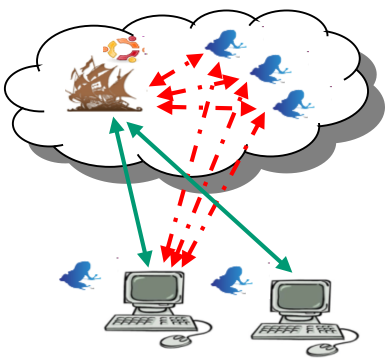


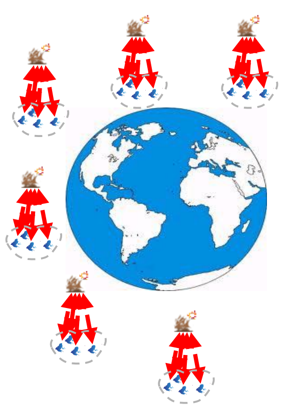
Popularity distribution



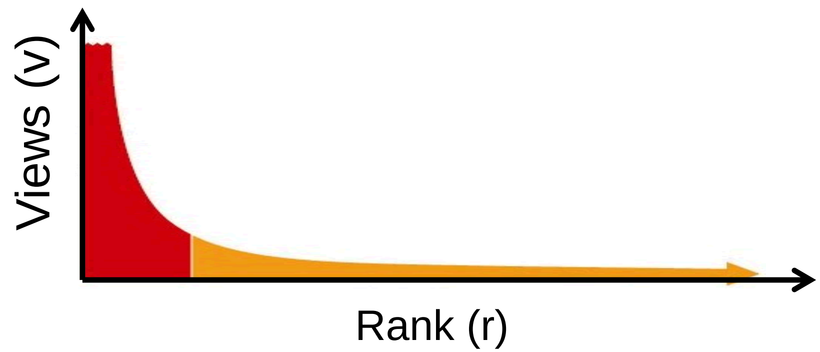


E.g., ACM KDD '12, PAM '12



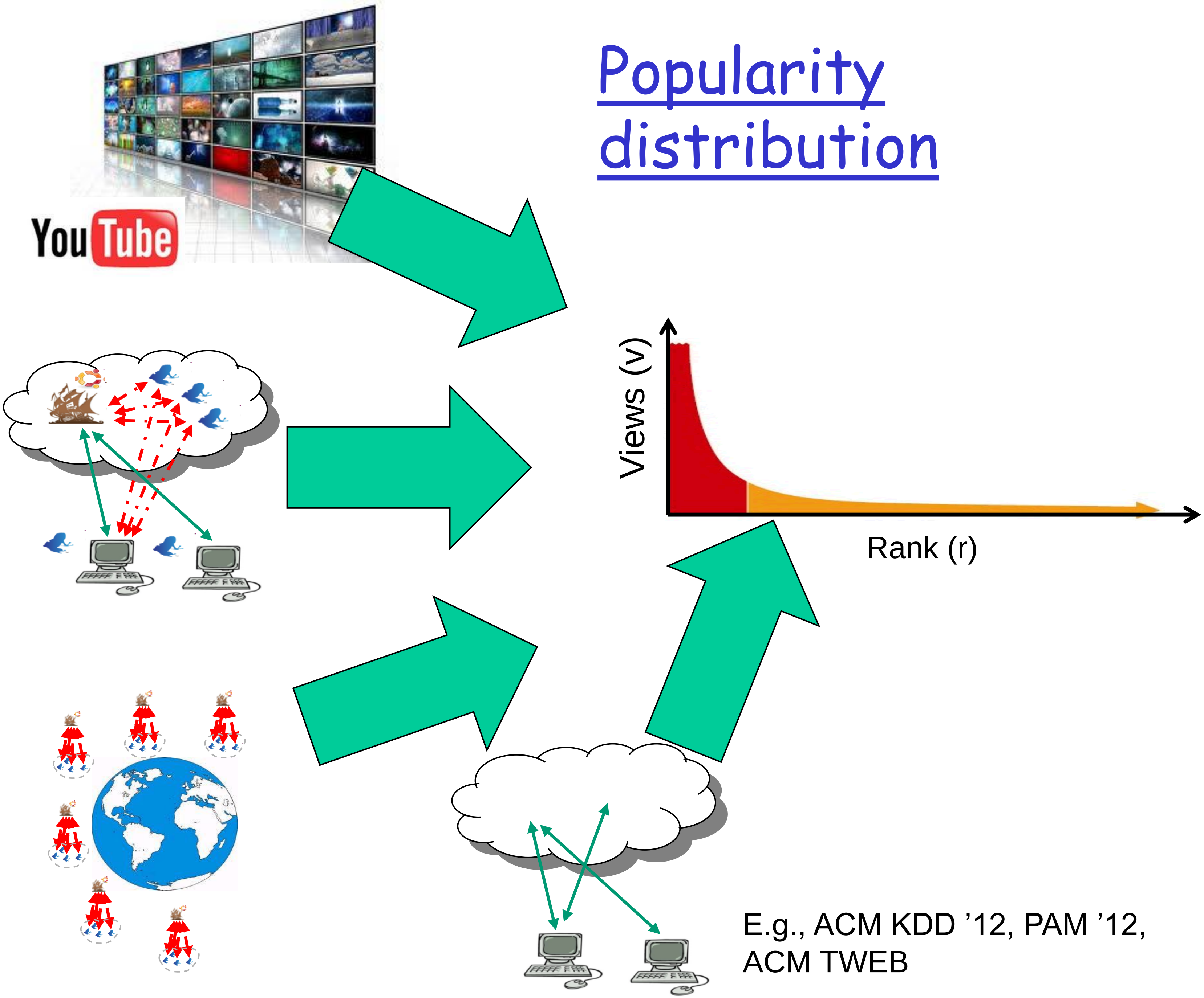


Popularity distribution



E.g., ACM KDD ’12, PAM ’12

Popularity distribution

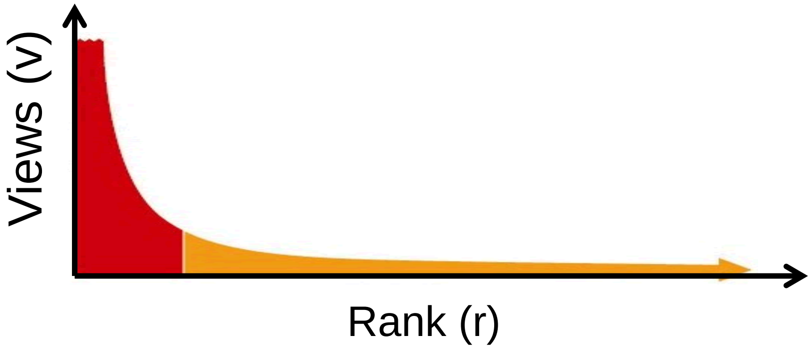


E.g., ACM KDD ’12, PAM ’12, ACM TWEB

Let's look at an example …

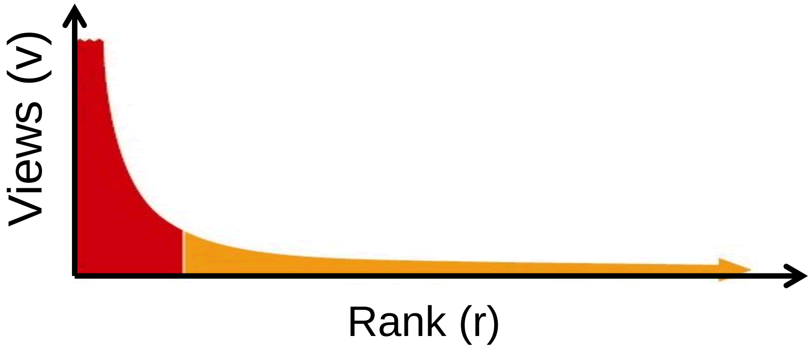
- Example 2

Zipf popularity…
… and long tails



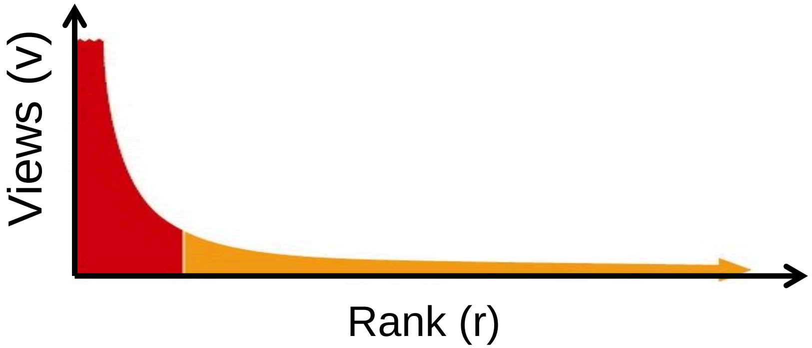
$$v_r \propto r^{-\alpha}$$

Zipf popularity…
… and long tails



$$v_r \propto r^{-\alpha}$$

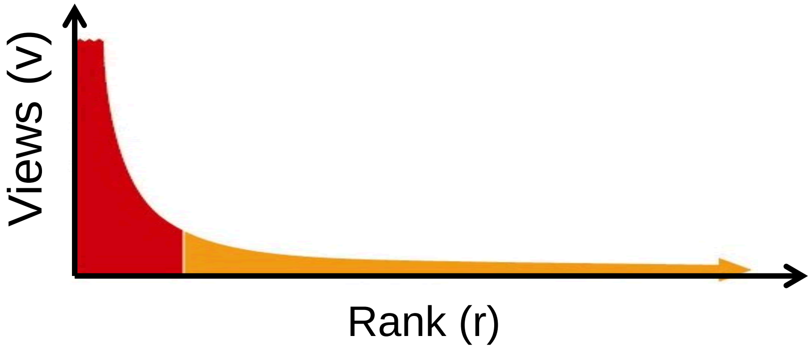
Zipf popularity…
… and long tails

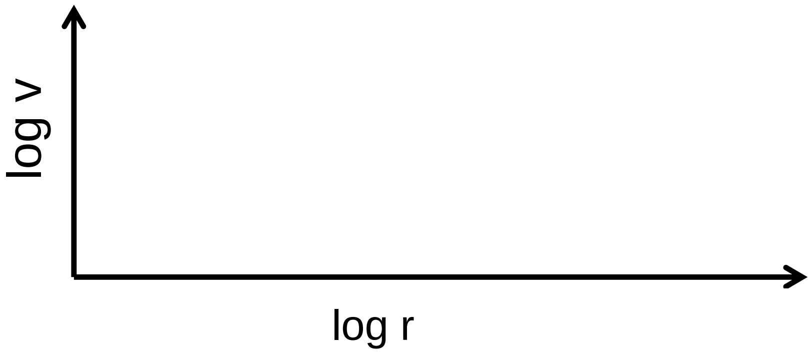


$$v_r \propto r^{-\alpha}$$

$$\log v_r = \log v_1 - \alpha \log r$$

Zipf popularity…
… and long tails

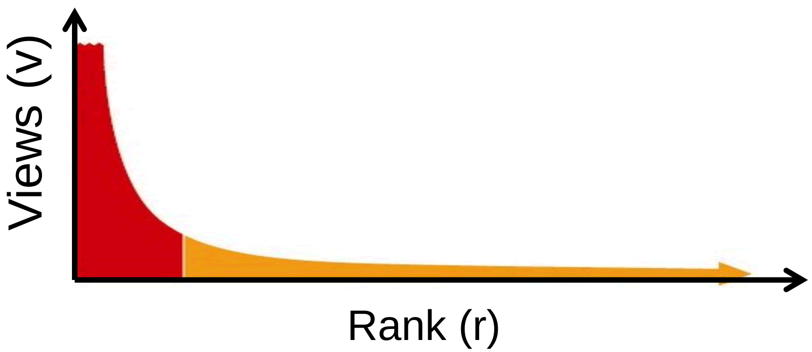




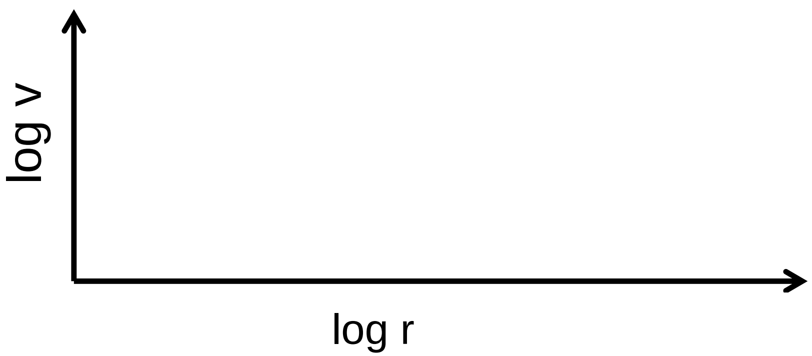
$$v_r \propto r^{-\alpha}$$

$$\log v_r = \log v_1 - \alpha \log r$$

Zipf popularity…
… and long tails



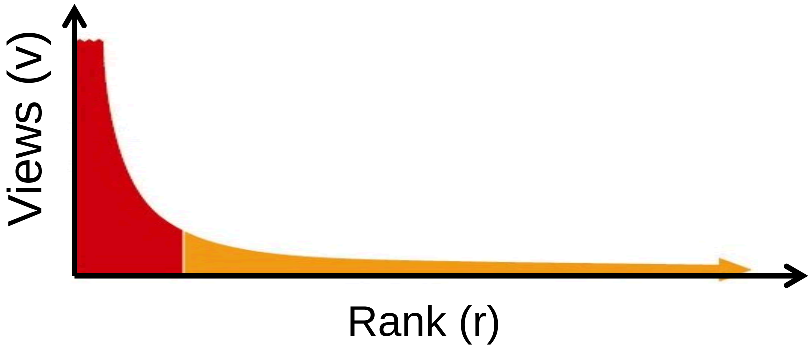
$$v_r \propto r^{-\alpha}$$

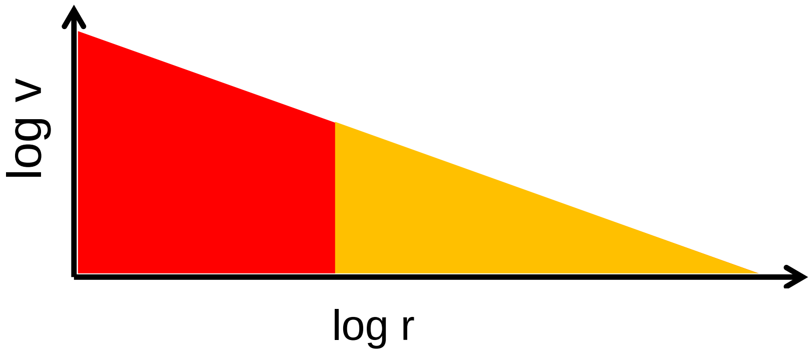


$$\log v_r = \log v_1 - \alpha \log r$$

$y(x) \quad = \quad x_0 \quad -- \quad \alpha \quad x$

Zipf popularity…
… and long tails



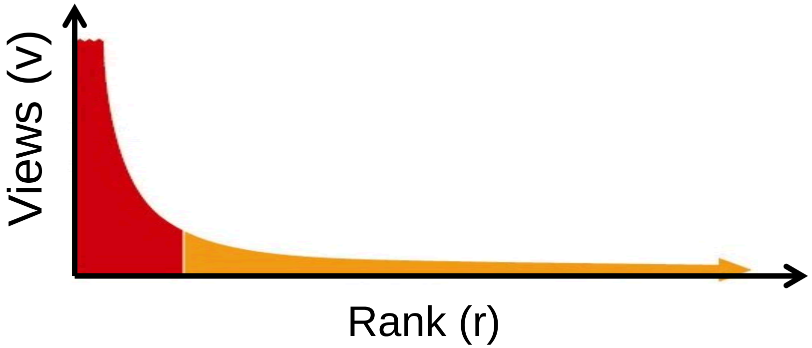


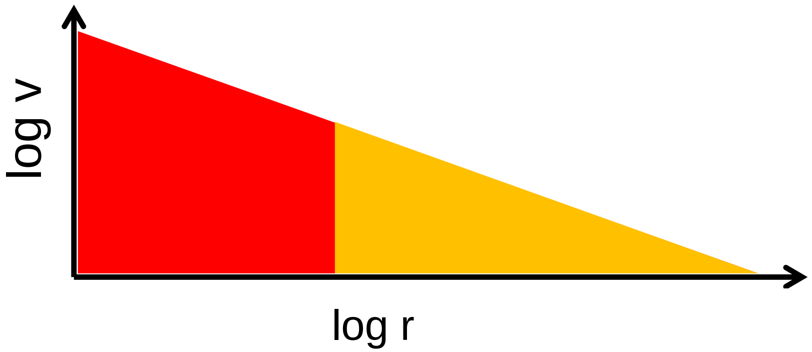
$$v_r \propto r^{-\alpha}$$

$$\log v_r = \log v_1 - \alpha \log r$$

y(x) = x_0 -- α x

Zipf popularity…
… and long tails

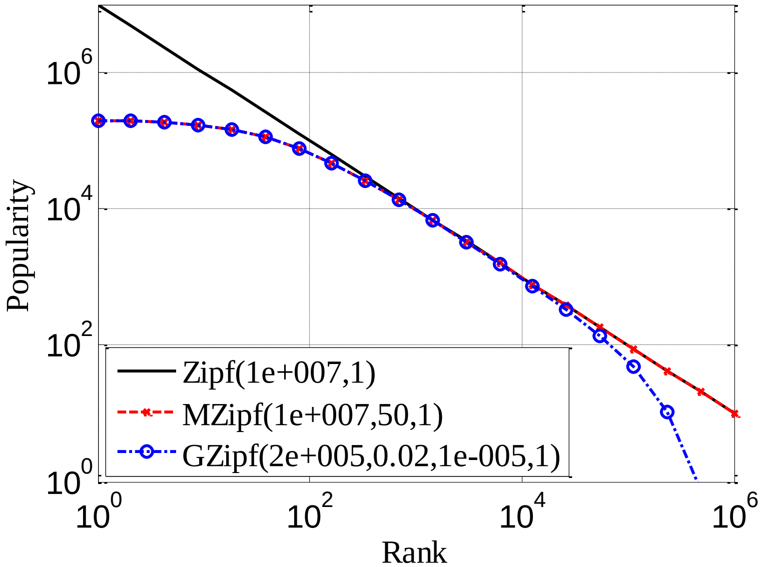


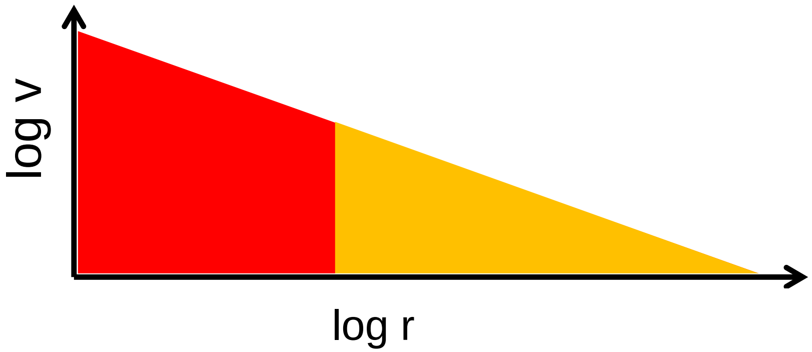


$$v_r \propto r^{-\alpha}$$

$$\log v_r = \log v_1 - \alpha \log r$$

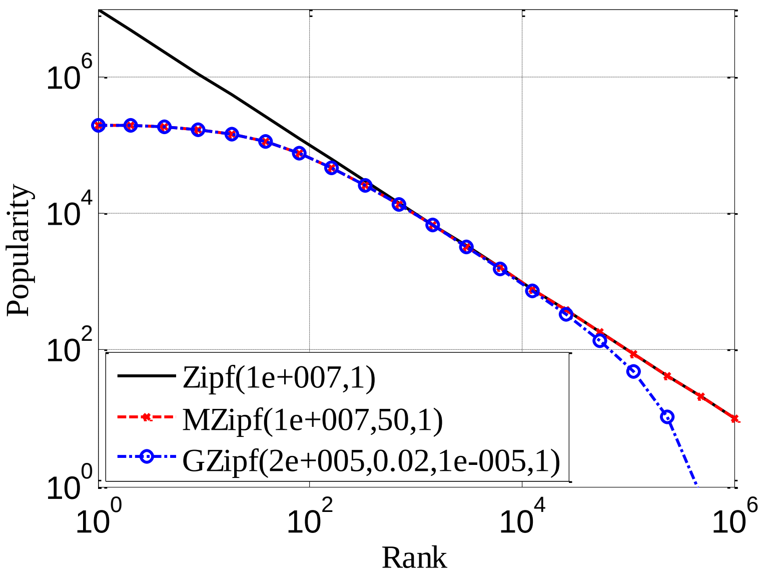
Zipf popularity…
… and long tails





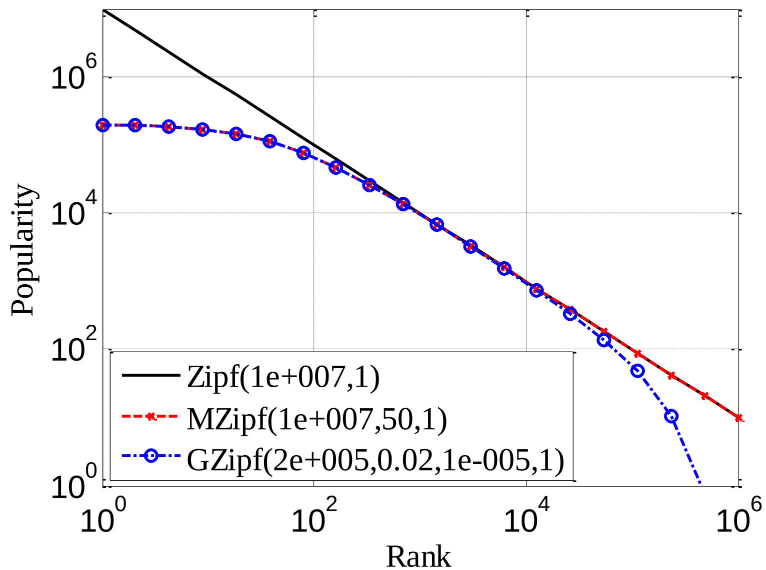
$$\log v_r = \log v_1 - \alpha \log r$$

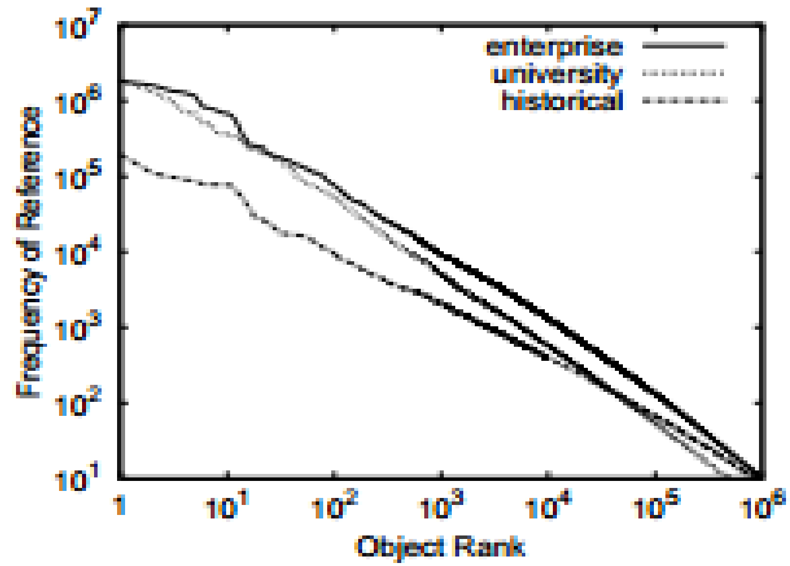
Zipf popularity…
… and long tails



E.g., ACM TWEB, PAM '11
IFIP Performance '11, IPTPS '10

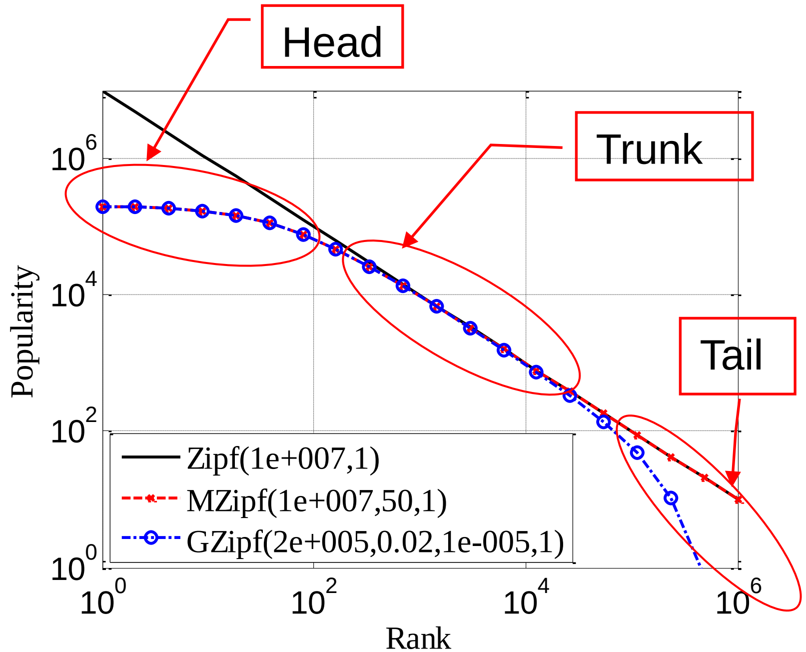
Zipf popularity…
… and long tails

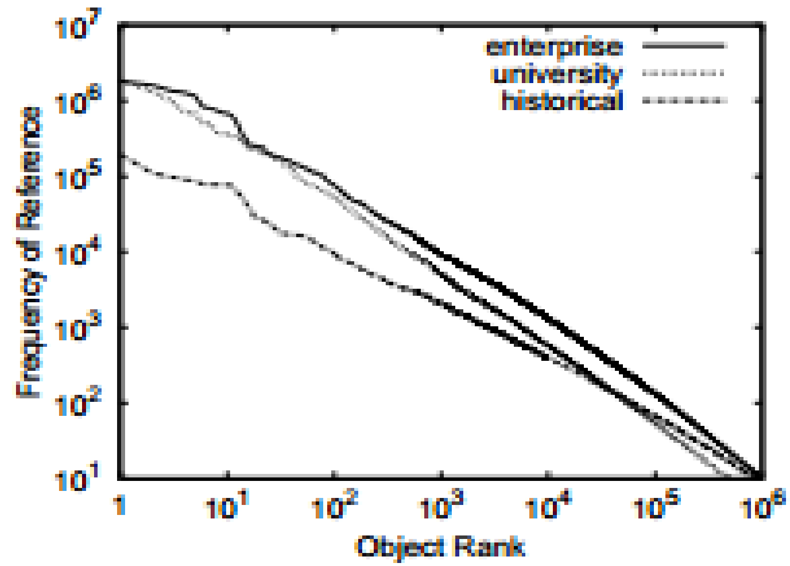




E.g., ACM TWEB, PAM ‘11
IFIP Performance ‘11, IPTPS ‘10

Zipf popularity…
… and long tails



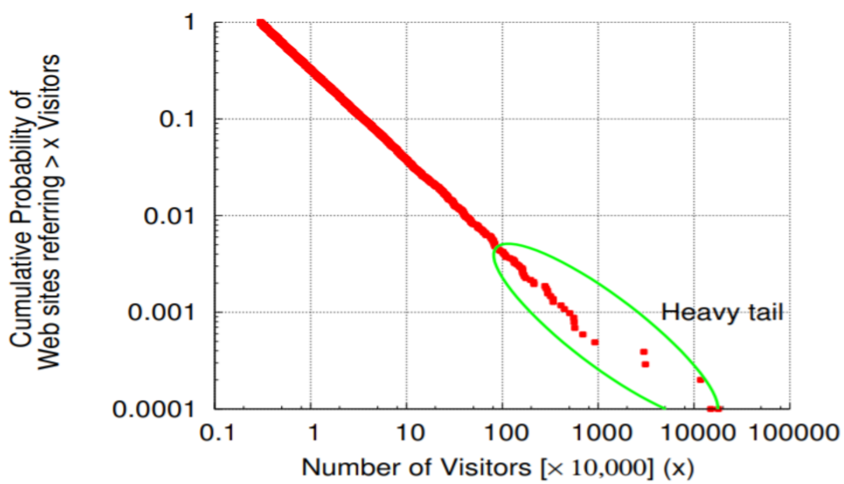


- Popularity distribution statistics
  - Across services (impact on system design)
  - Lifetime vs current
  - Over different time period (churn)
  - Different sampling methods
  - Different measurement location

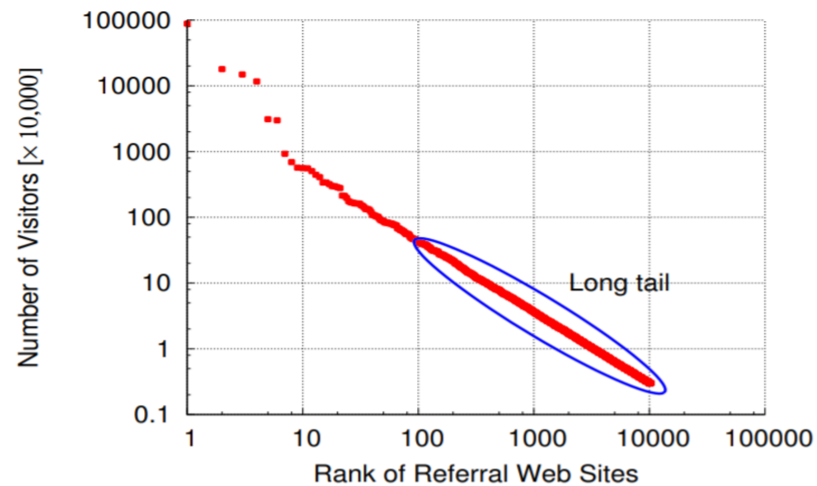
E.g., ACM TWEB, PAM ’11, IFIP Performance ‘11, IPTPS ‘10

Power law, Pareto, and Zipf

- Power-law, Pareto, Zipf (in some sense the same)
  - Power-law: $f(x) \sim x^{-\eta}$ (probability of value x)
  - Pareto: $F(x) = P[X > x] = \int f(x)\,dx \propto x^{-\kappa}$ (cumulative prob.)
  - Zipf: $v_r \propto r^{-\alpha}$ (discrete representation; frequency v_r of rank r)
  - Parameters related as: $\kappa = \eta - 1 = 1/\alpha$
    - E.g., paper and references therein: "A Tale of the Tails: Power-laws in Internet Measurements", IEEE Network, Mahanti et al., 2013



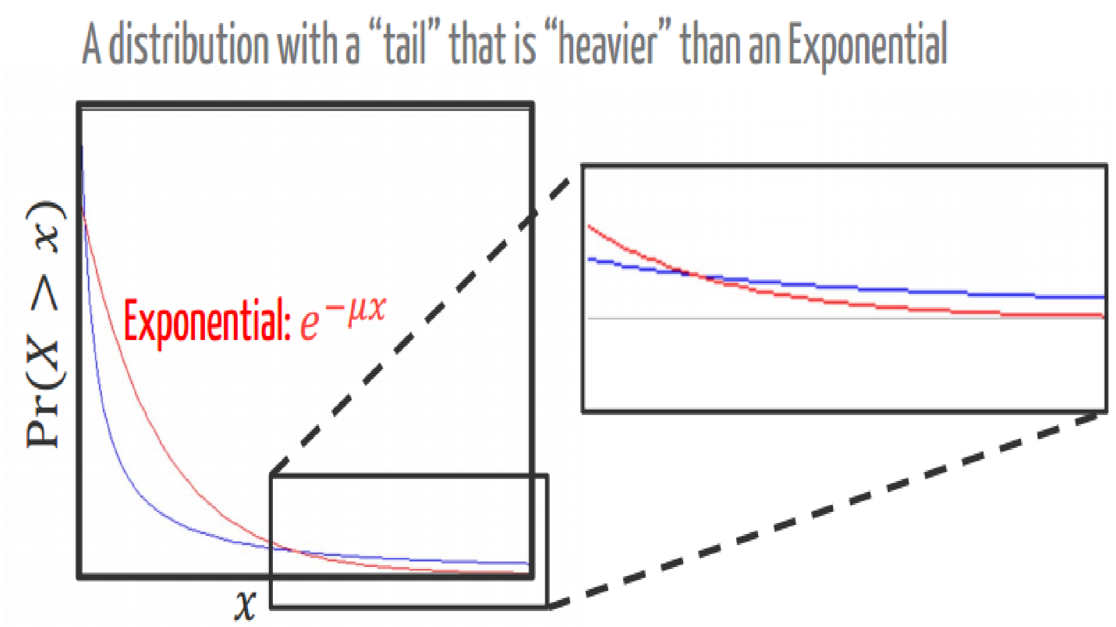
(b) Complementary Cumulative Distribution Function



(a) Rank Size Distribution

Heavy-tail distributions ...

- "A probability distribution is said to have a heavy tail if the tail is not exponentially bounded"
- ... and then there are many many other "heavy tail" distributions, variations and generalizations, including distributions such as log-normal, various generalized Zipf/Pareto distributions, etc.

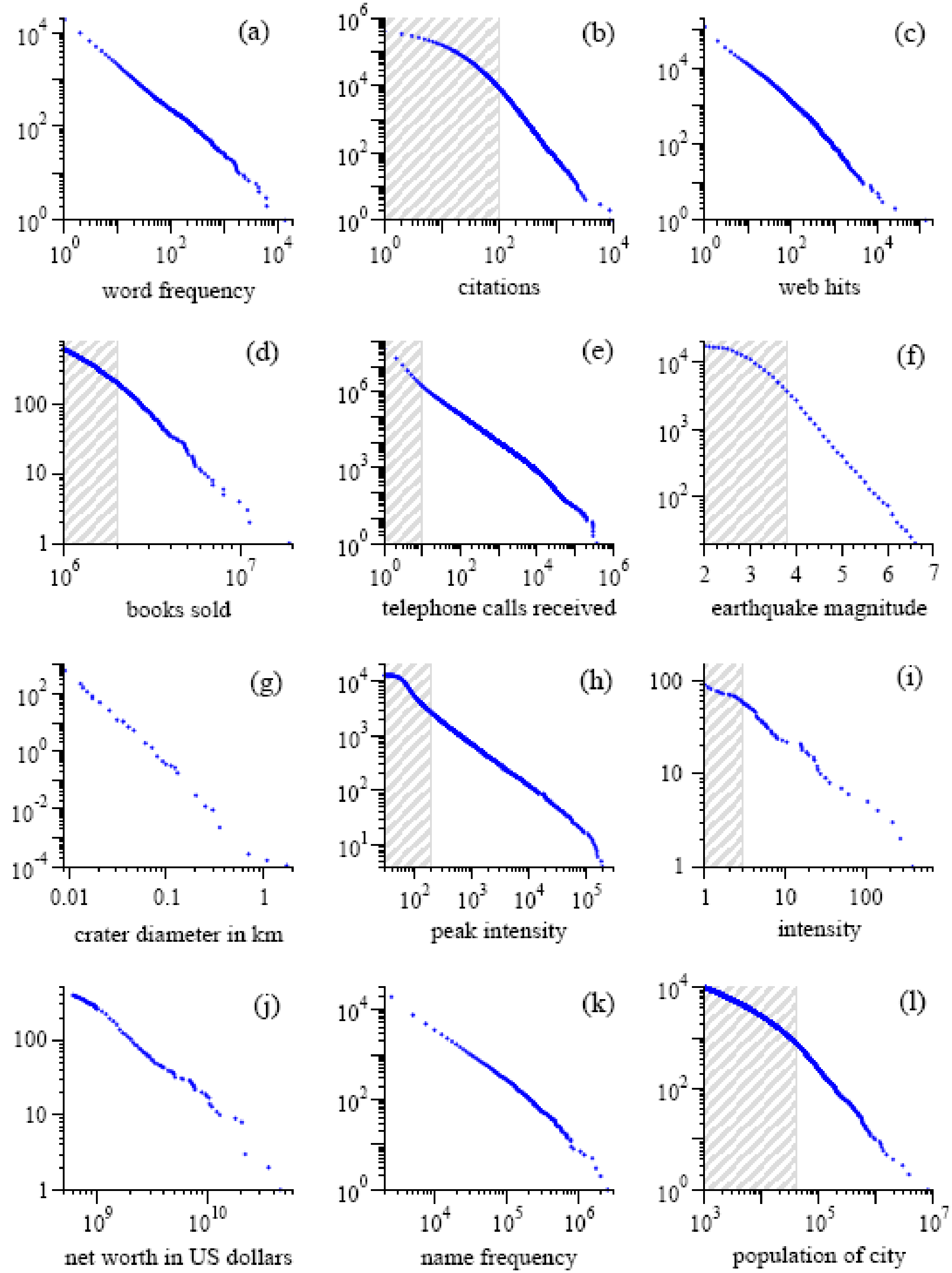


(more) Examples of power laws

a. Word frequency: Estoup.
b. Citations of scientific papers: Price.
c. Web hits: Adamic and Huberman
d. Copies of books sold.
e. Diameter of moon craters: Neukum & Ivanov.
f. Intensity of solar flares: Lu and Hamilton.
g. Intensity of wars: Small and Singer.
h. Wealth of the richest people.
i. Frequencies of family names: e.g. US & Japan not Korea.
j. Populations of cities.

… AND many many more …

The following graph is plotted using Cumulative distributions



M. E. J. Newman, "Power laws, Pareto distribution and Zipf's law", Contemporary physics (2005).

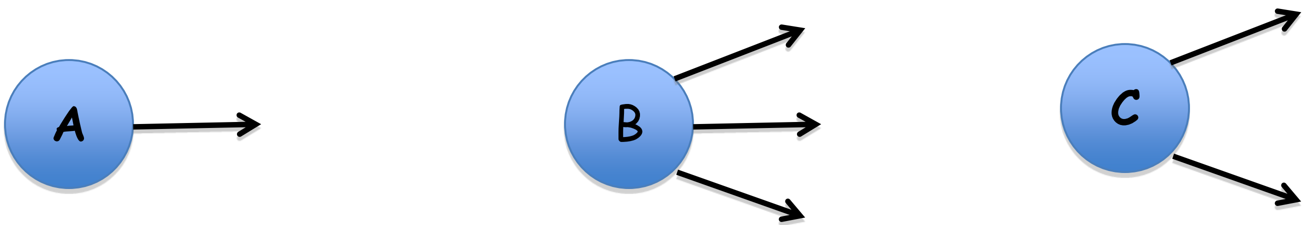
Real world data for x_{min} and α

	x_{min}	α
frequency of use of words	1	2.20
number of citations to papers	100	3.04
number of hits on web sites	1	2.40
copies of books sold in the US	2 000 000	3.51
telephone calls received	10	2.22
magnitude of earthquakes	3.8	3.04
diameter of moon craters	0.01	3.14
intensity of solar flares	200	1.83
intensity of wars	3	1.80
net worth of Americans	$600m	2.09
frequency of family names	10 000	1.94
population of US cities	40 000	2.30

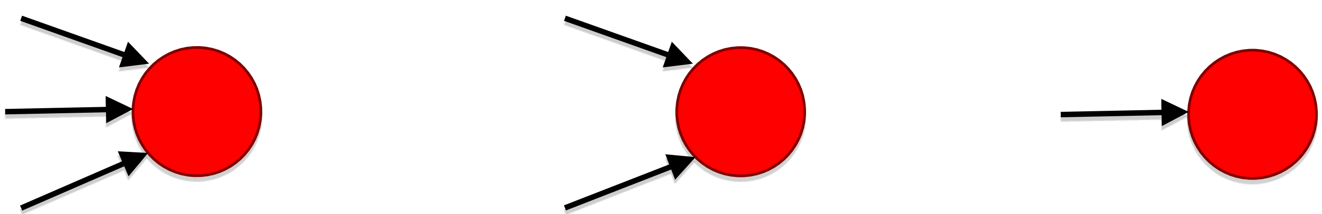
Now, consider a **social network**, the Internet, or some other network ...

Preferential Attachment (PA)

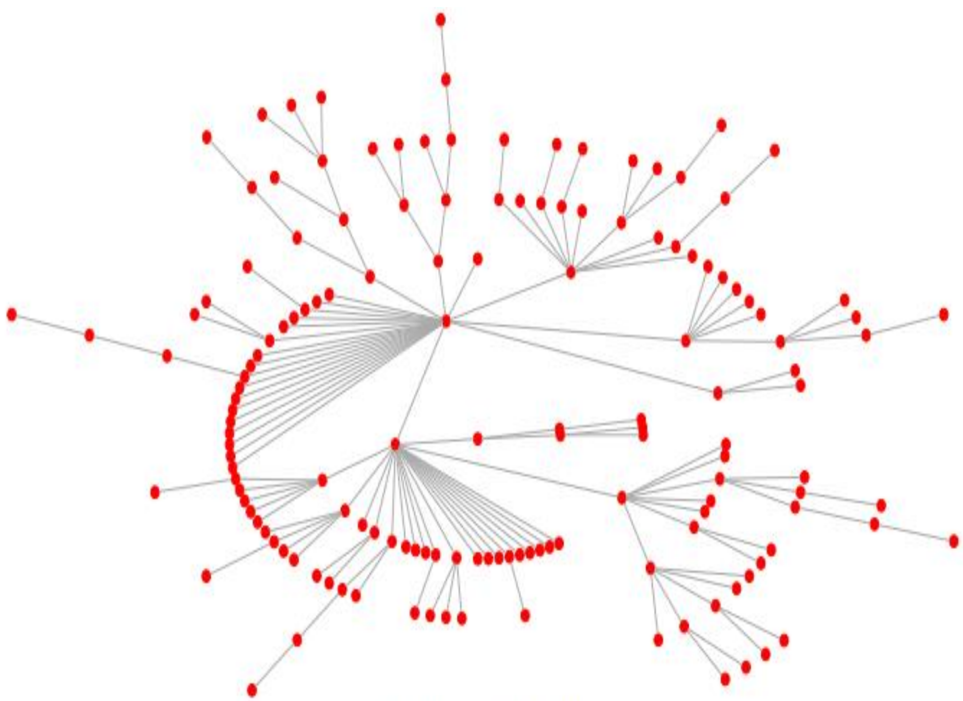
- Link probability proportional to node degree
  - p_i proportional to k_i^{α}
  - For **source node selection** (Out-degree, α = 0.8)



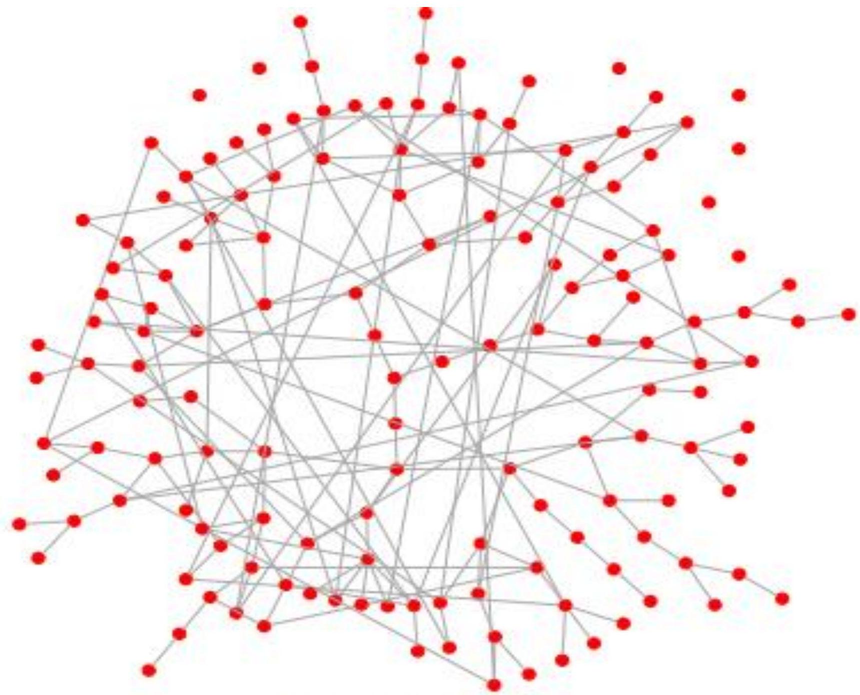
  - For **destination node selection** (In-degree, α = 0.9)



Preferential attachment and Power law



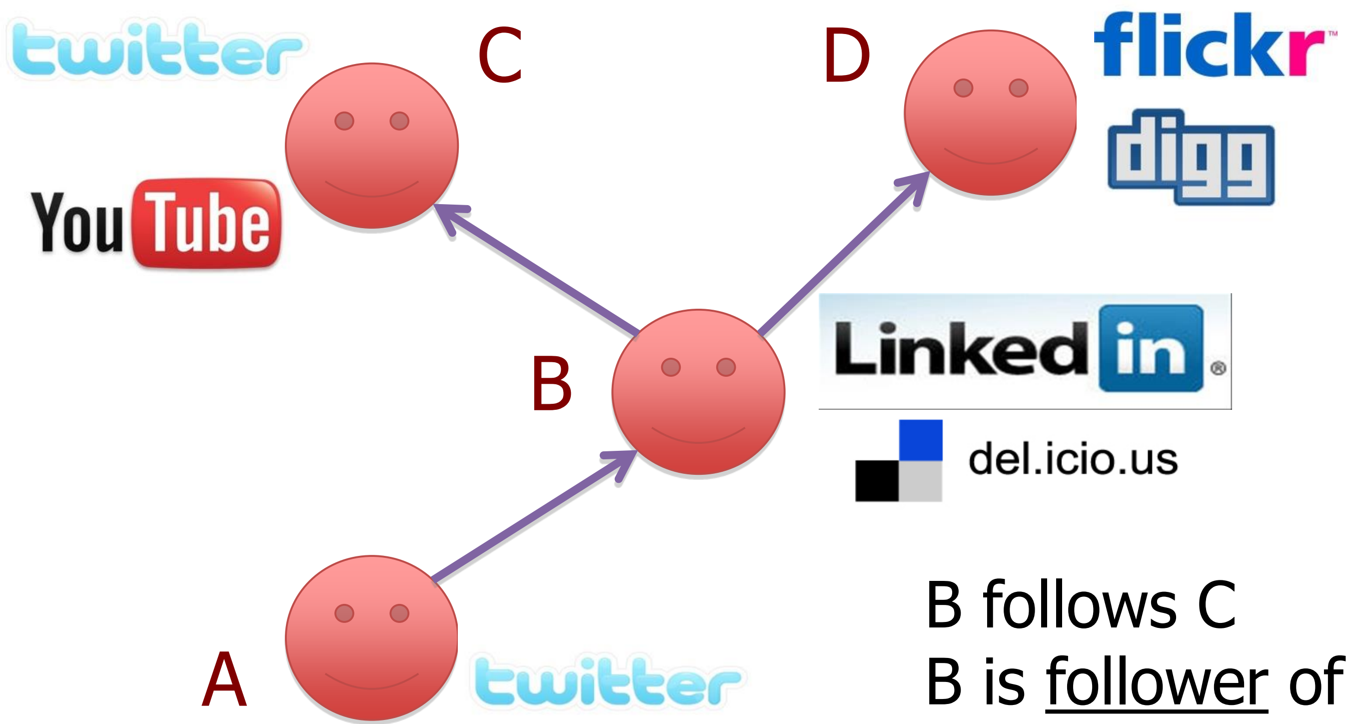
(a) Power-law graph



(b) Random graph

- Preferential attachment (or rich gets richer) have been shown to result in power-law graphs
- In contrast, the Erdös-Rényi random graph has an exponential node degree distribution

friendfeed [Garg et al. IMC '09]

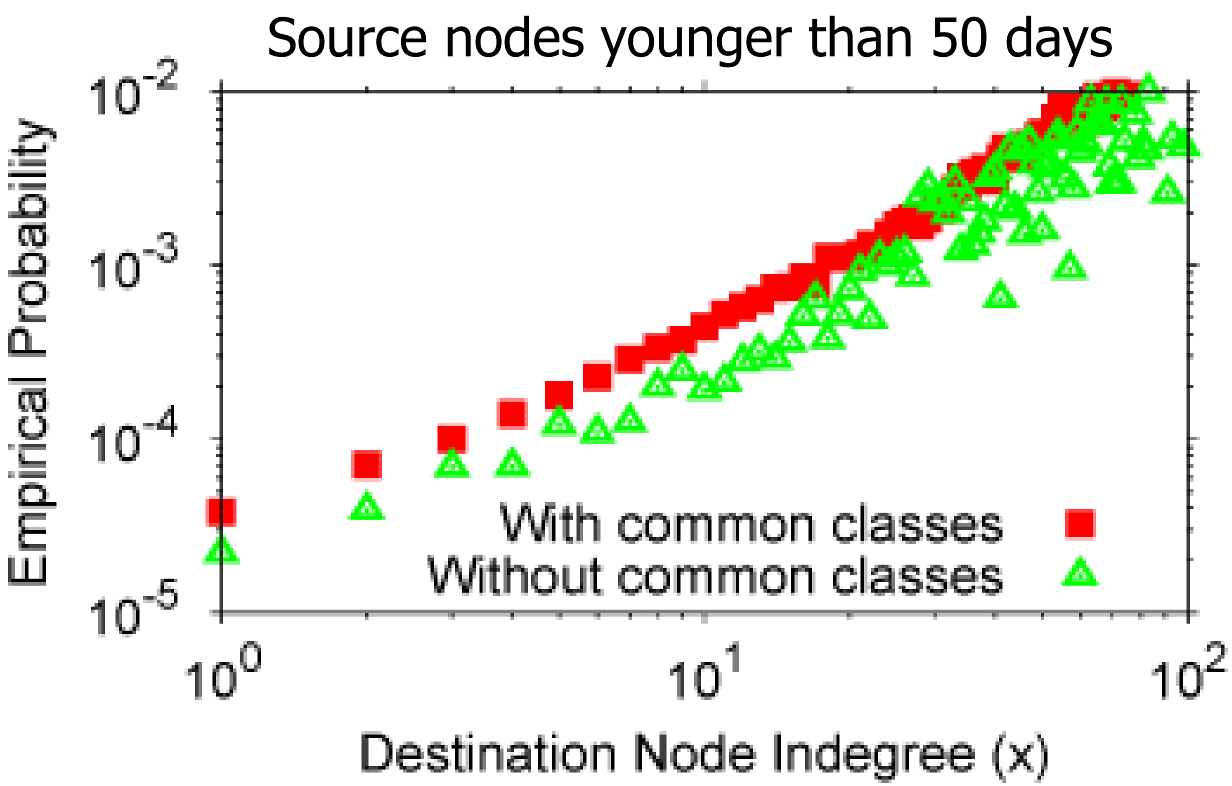


B follows C
B is <u>follower</u> of C
C is <u>friend</u> of B

[Garg et al. IMC '09]

Group Affiliation & Link Formation

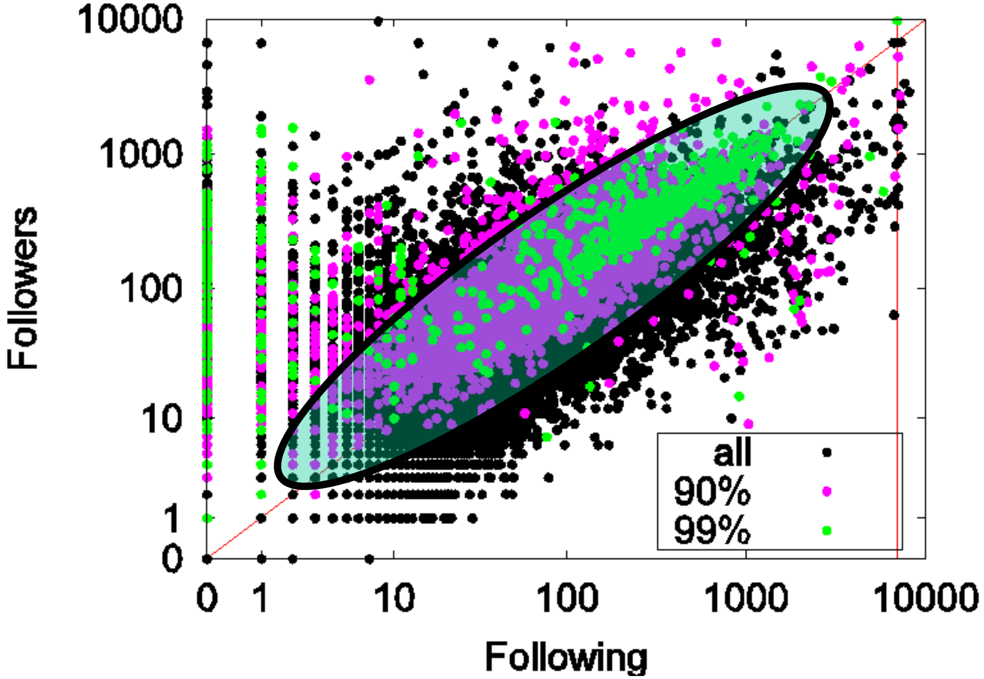
- Does PA explain the observed data? Yes!
- Does subscription to common services (common interest) biases the preference? Yes!



A few chirps about Twitter

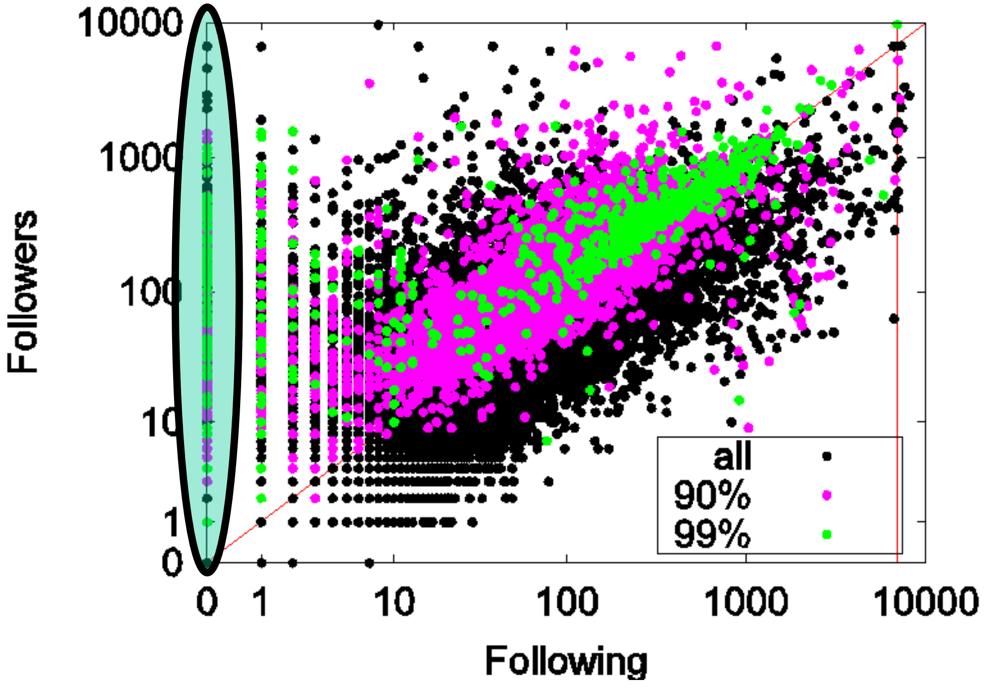
- … by Krishnamurthy, Gill, and Arlitt

Aside: User relationships on Twitter



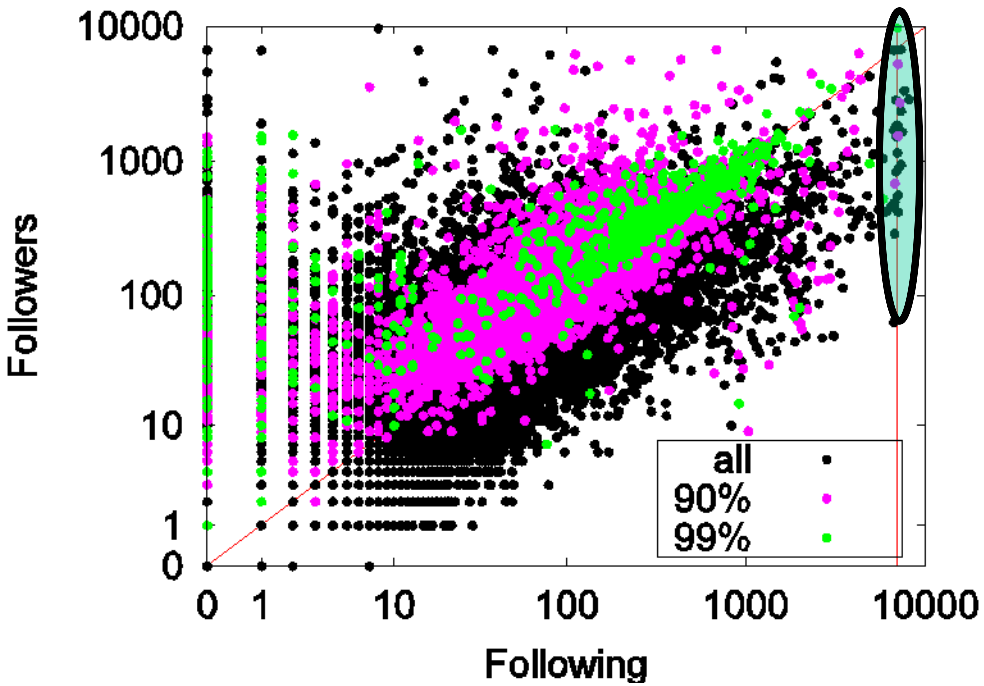
- **Acquaintances**
- Similar number of followers and following
- Along the diagonal
- Green portion is top 1-percentile of tweeters

Aside: User relationships on Twitter



- **Broadcasters**
- News outlets, radio stations
- No reason to follow anyone
- Post playlists, headlines

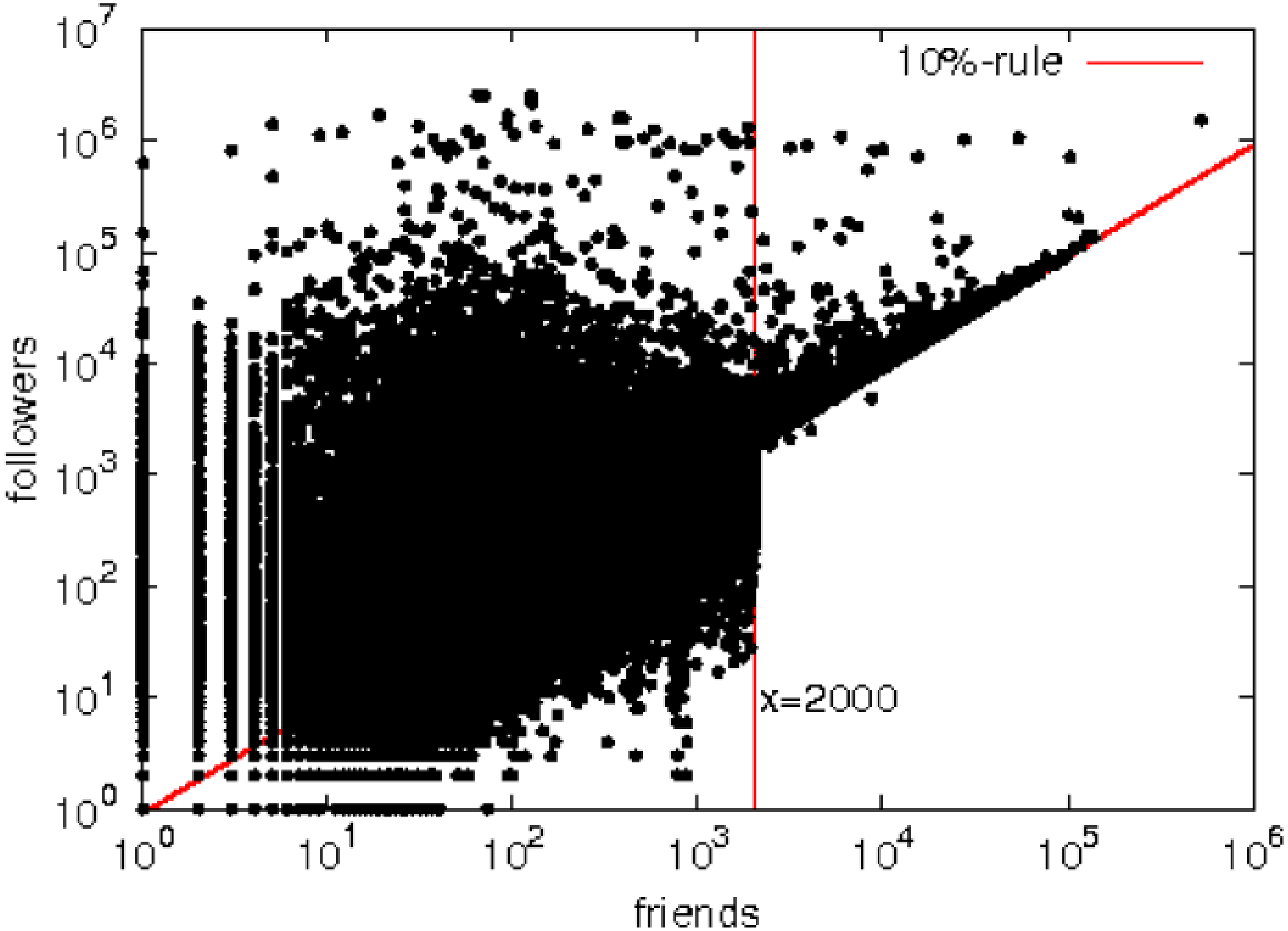
Aside: User relationships on Twitter



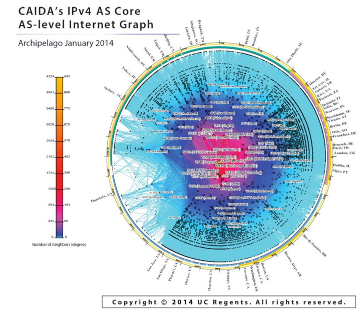
- **Miscreants?**
- Some people follow many users (programmatically)
- Hoping some will follow them back
- Spam, widgets, celebrities (at top)

Aside: User relationships on Twitter

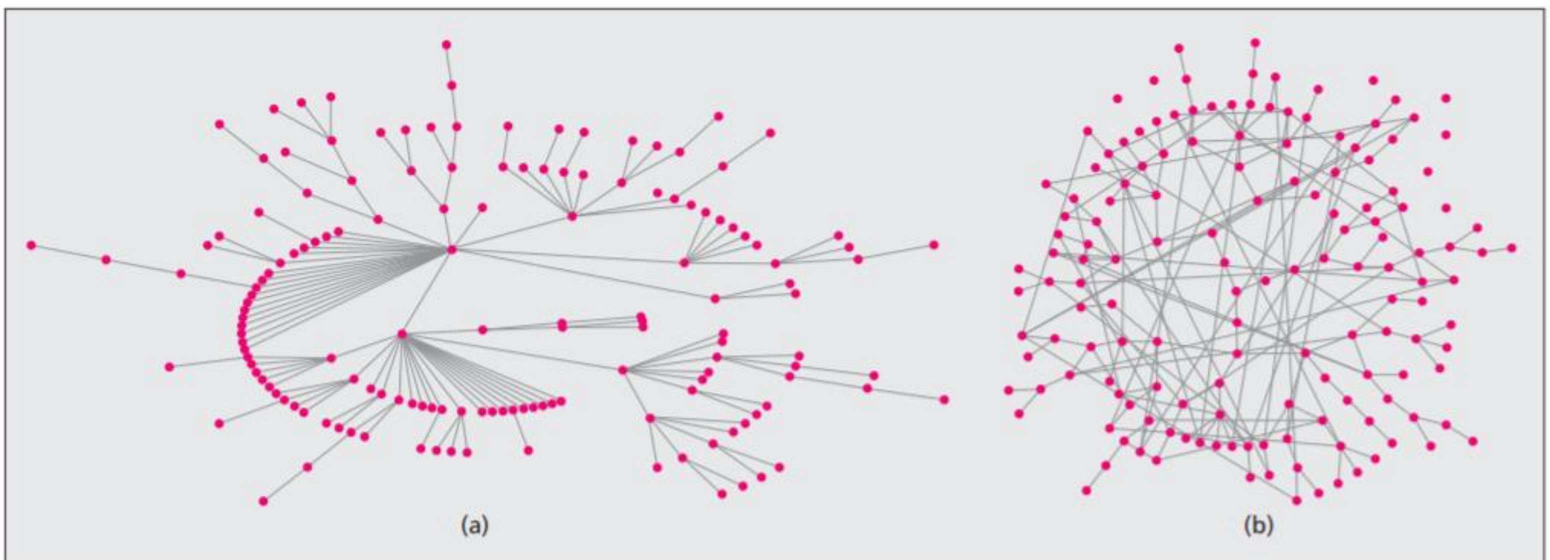
Twitter noticed the miscreants…
… enacted the 10% rule (you can follow 10% more people than follow you)



Are Scale-Free Networks Better?



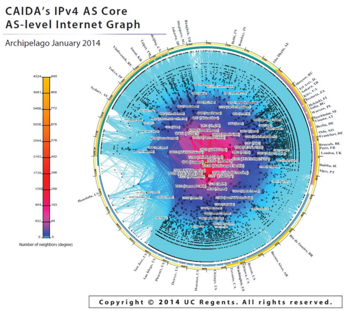
- Scale-free networks have power-law degree distribution (at least asymptotically)
- Average diameter lower in scale-free (SF) than in exponential (E) graphs



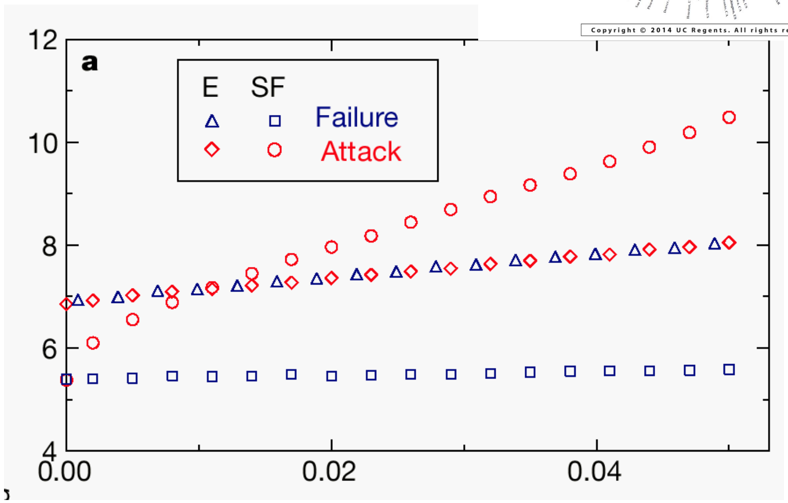
Mahanti et al. 2013

Figure 5. *Comparison of power-law and random graphs: Each graph consists of 150 vertices. A vertex is represented by a red dot and the edge is shown using a solid grey line. The graphs were simulated using the NetworkX package in Python, and visualized using* `Graphviz`.

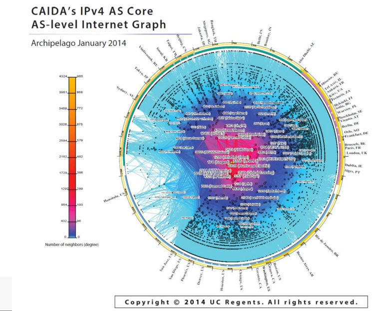
Are Scale-Free Networks Better?



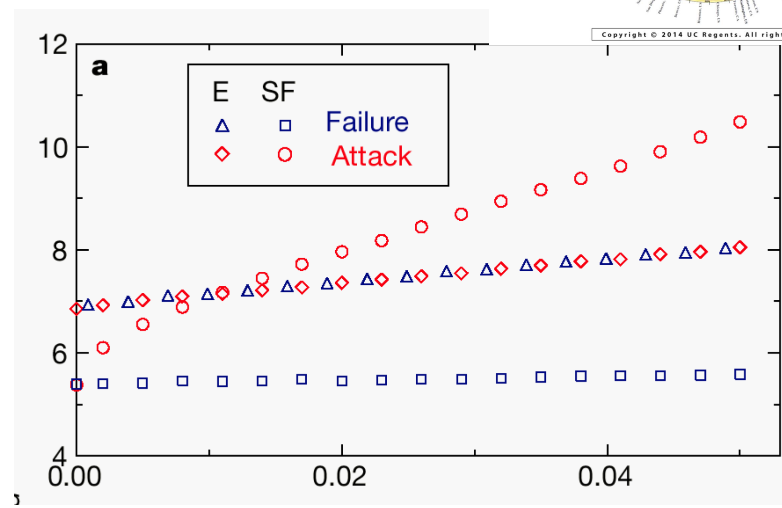
- Scale-free networks have power-law degree distribution (at least asymptotically)
- Average diameter lower in scale-free (SF) than in exponential (E) graphs
- What if nodes are removed?
  - at random: scale free keeps lower diameter
  - by knowledgable attacker (nodes of highest degree removed first): scale-free diameter grows quickly
-

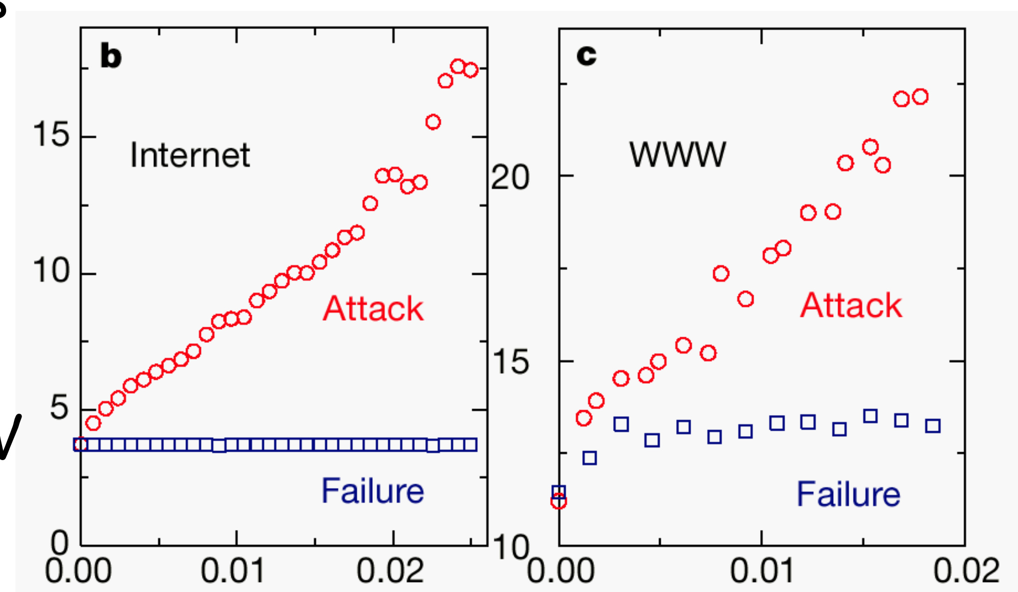


Are Scale-Free Networks Better?



- Scale-free networks have power-law degree distribution (at least asymptotically)
- Average diameter lower in scale-free (SF) than in exponential (E) graphs
- What if nodes are removed?
  - at random: scale free keeps lower diameter
  - by knowledgable attacker (nodes of highest degree removed first): scale-free diameter grows quickly
- Same results apply using sampled Internet and WWW graphs (that happen to be scale-free)



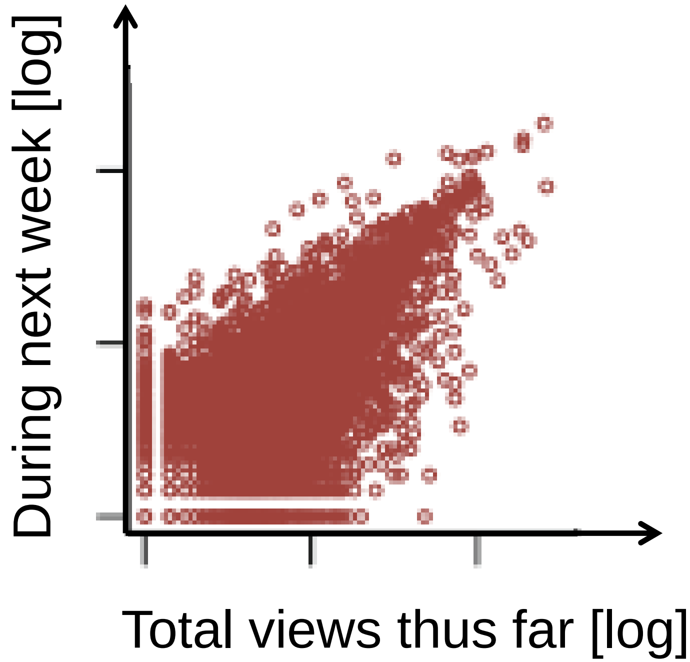


YouTube

… and back to the video example again …

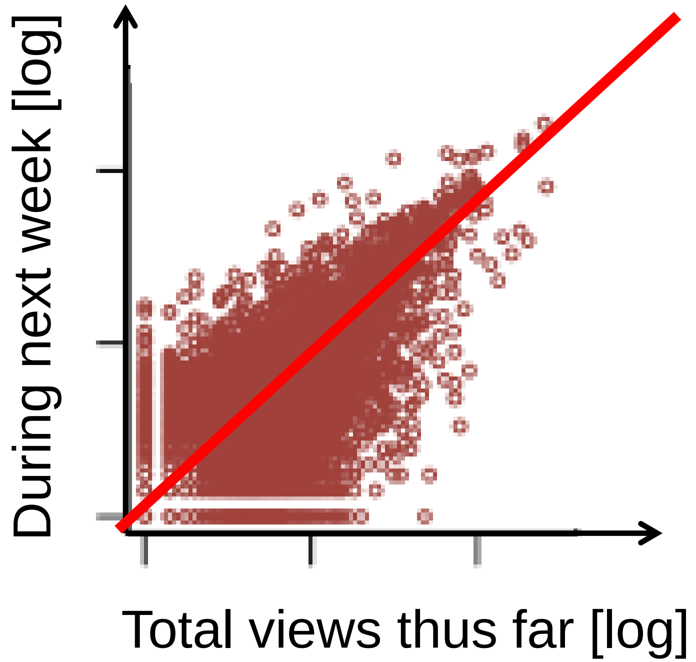
E.g., ACM KDD ’12, IFIP Performance ‘11, ACM TWEB

Rich-gets-richer …
… and churn



E.g., Borghol et al.
IFIP Performance ‘11

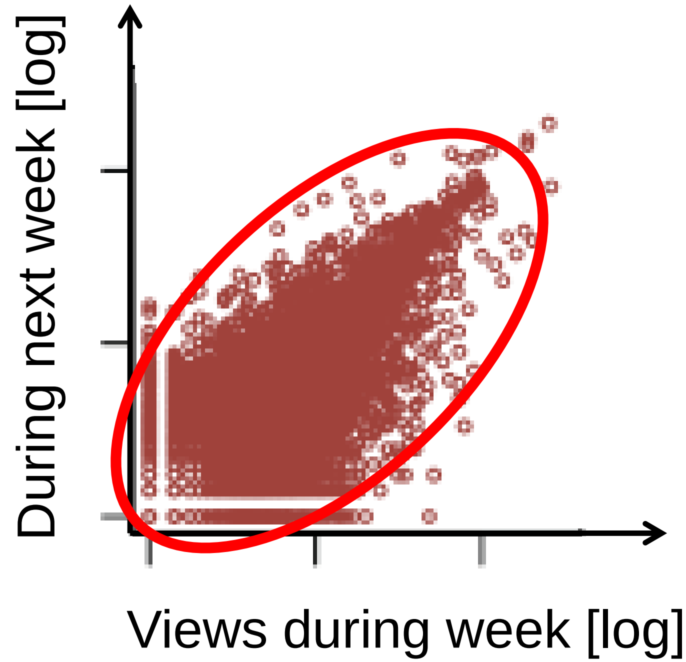
Rich-gets-richer ...
... and churn



- The more views a video has, the more views it is likely to get in the future

E.g., Borghol et al.
IFIP Performance ‘11

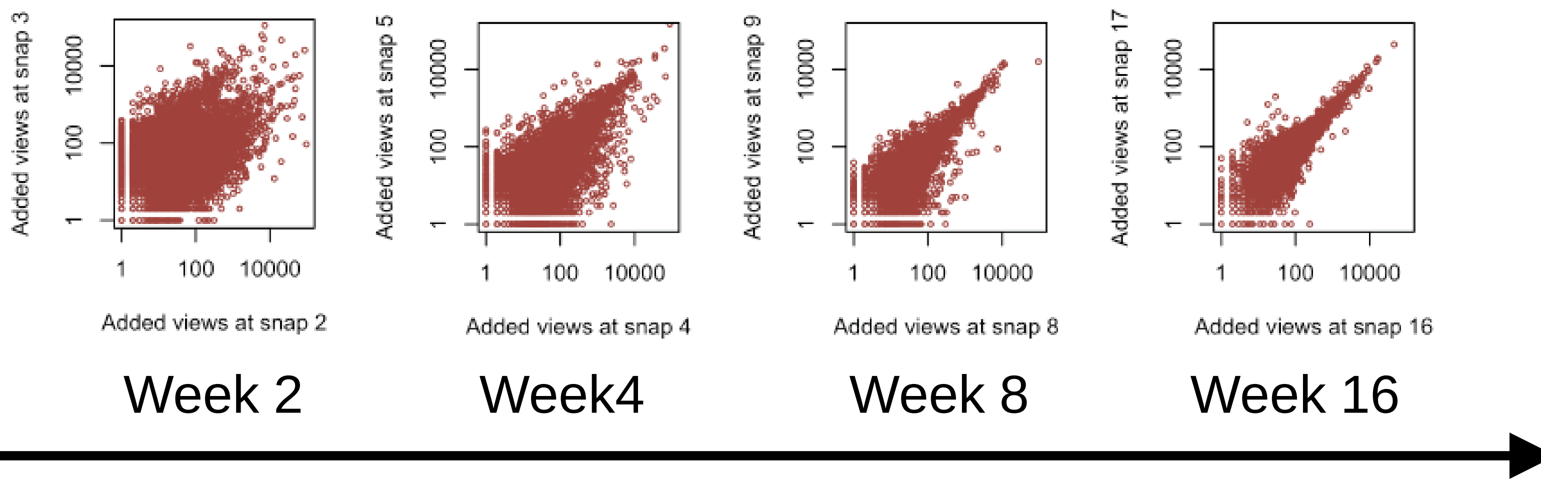
Rich-gets-richer …
… and churn



- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary

E.g., Borghol et al.
IFIP Performance ‘11

Rich-gets-richer ...
... and churn



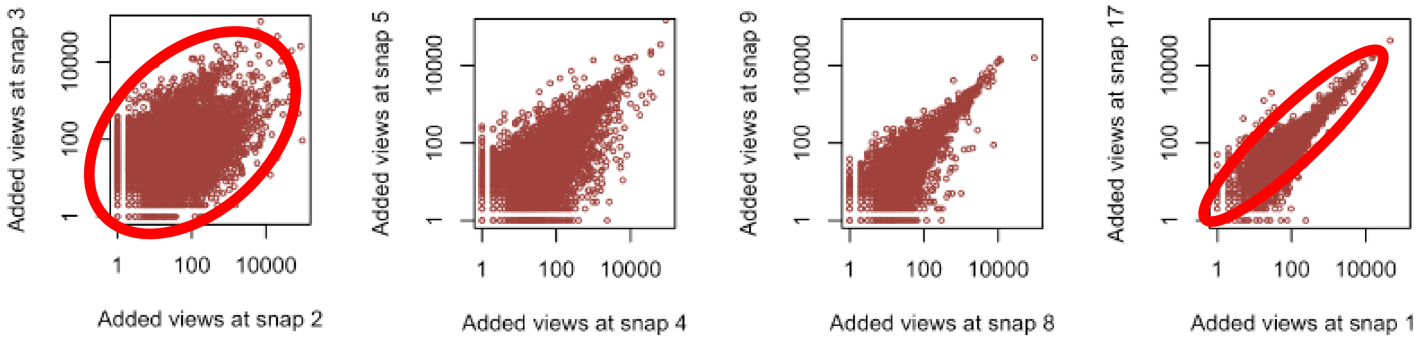
Young videos

Old videos

- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary

E.g., Borghol et al.
IFIP Performance '11

Rich-gets-richer ...
... and churn



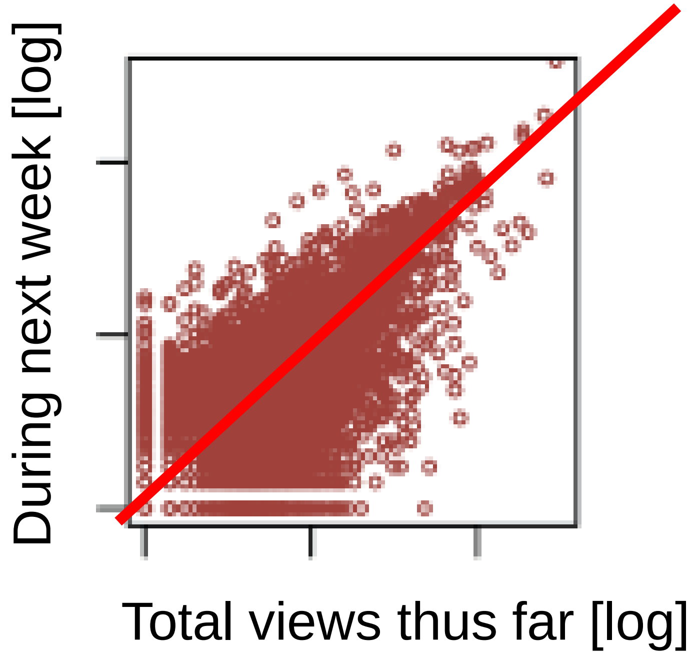
Week 2 Week4 Week 8 Week 16

Young videos Old videos

- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary
- Some long-term popularity

E.g., Borghol et al.
IFIP Performance '11

Rich-gets-richer ...
... and churn



- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary
- Some long-term popularity

E.g., Borghol et al.
IFIP Performance ‘11