Power Laws & Rich Get Richer

Social Computing

Department of Computer Science
University of Massachusetts, Lowell

Hadi Amiri
hadi@cs.uml.edu

Lecture Topics

- Popularity
- Power Laws
- Rich Get Richer model

Popularity

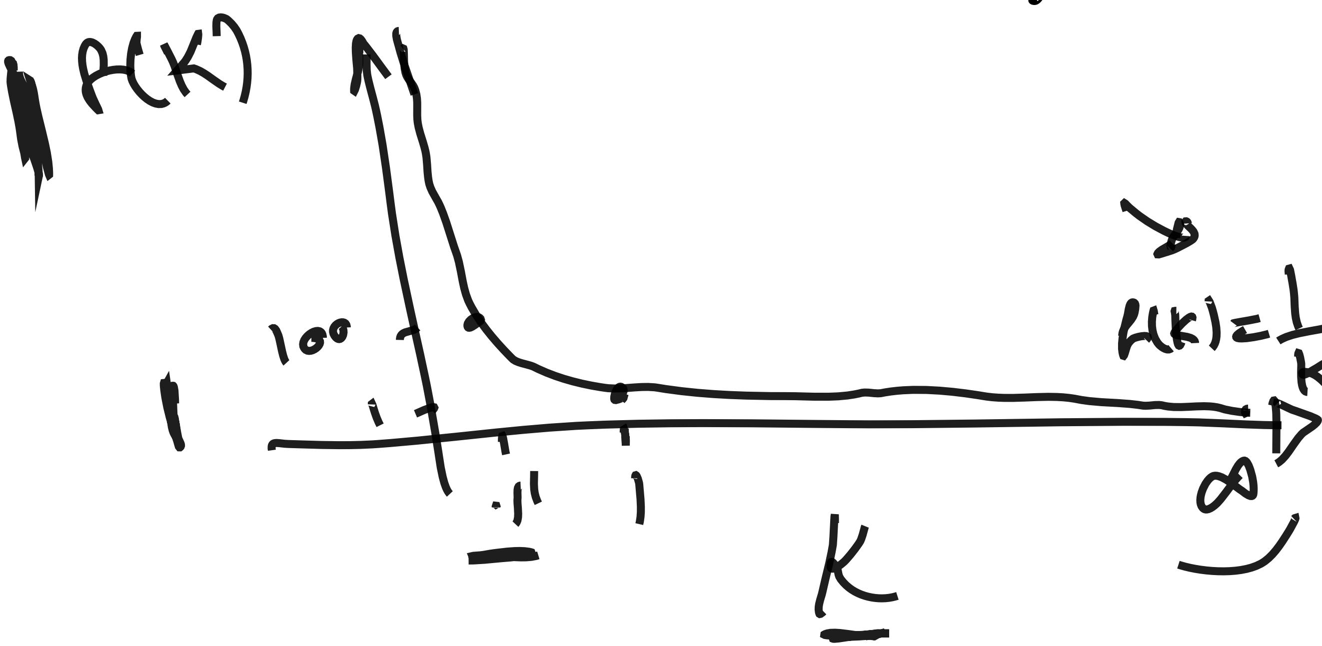
- Popularity can be characterized by **extreme imbalances**!
  - People are known to their immediate social circle!
  - Few people achieve wider visibility!
  - Very few achieve global name recognition.

- Learning objectives:
  - How can we quantify these imbalances?
  - Why do they arise?
  - Can popularity be predicted?

Power Law

- A function that decreases as *k* to some fixed power, e.g. $1/k^2$, is called a **power law**!
  - It allows to see very large values of *k* in data!

- Extreme imbalances are likely to arise!



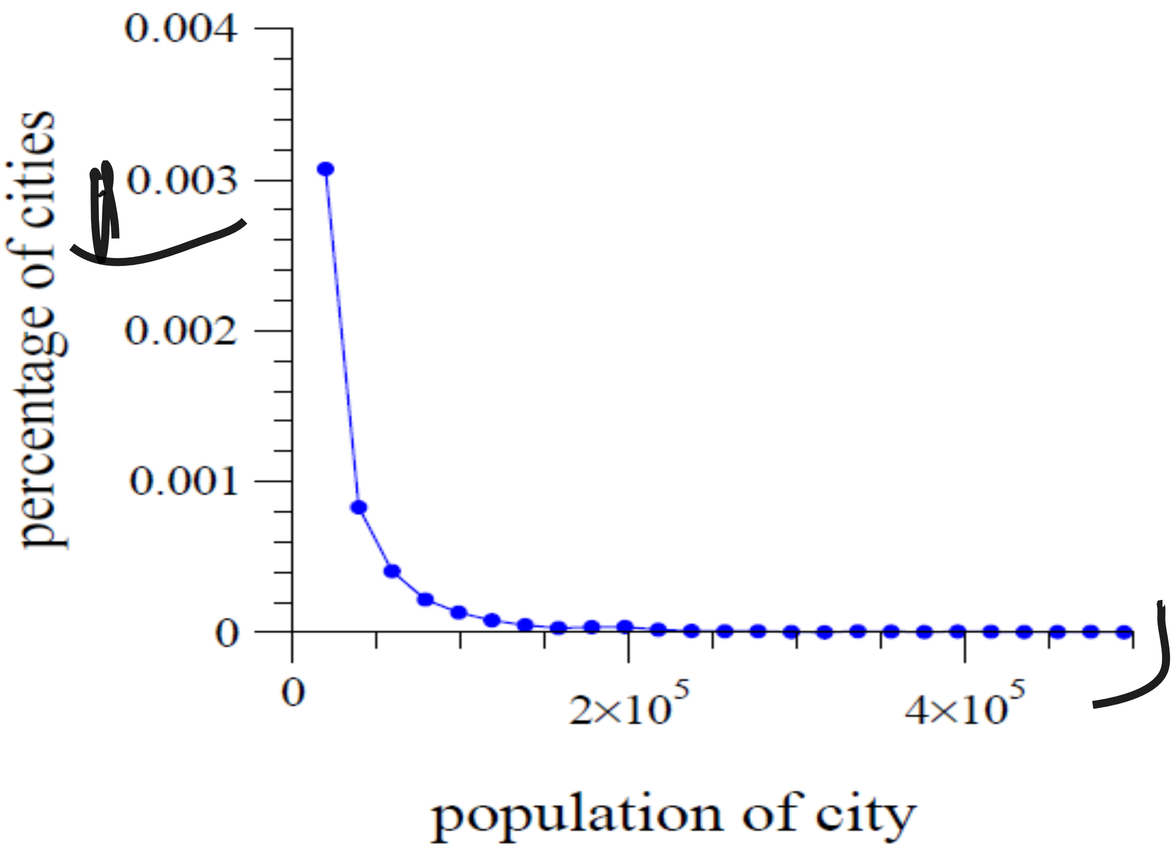
$h(k) = \frac{a}{k^c}$

$h(k) = \frac{1}{k^2}$

$h(k) = \frac{1}{k^{40}}$

Power Law- Cnt.

- Histogram of the populations of all US cities with population of 10,000 or more.



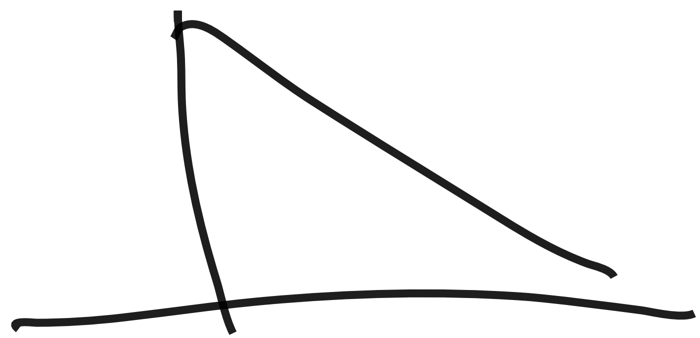
Power Law- Cnt.

- **Power law Test**: Given a dataset, test if it exhibits a power law distribution?
  1. Compute histogram of values wrt a popularity measure (e.g. *#in-links*, *#downloads*, *population of cities, etc.*)
  2. Test if the result approximately estimates a power law a/k^c for some a and c, and if so, estimate a and the exponent c.

Power Law- Cnt.

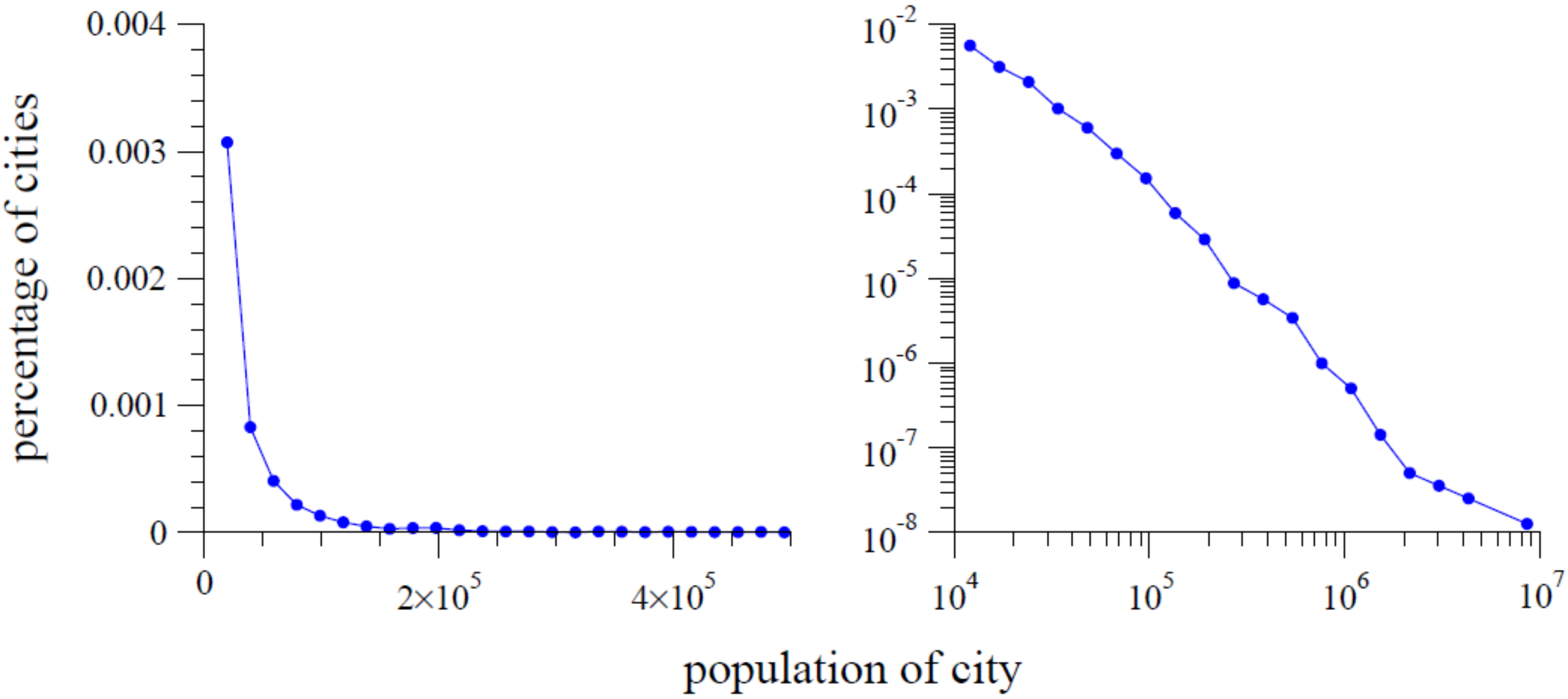
- What should a power law plot look like?
  - $f(k)$: the fraction of items that have value k
  - If power law holds, $f(k) = a/k^c$?
    - for some constant c and a.
  - $f(k) = a/k^c = ak^{-c}$
  - $\log f(k) = \log a - c \log k$
    - **straight line!** "log $f(k)$" as a function of "log k"
      - "c": slope, and
      - "log a": y-intercept.
    - log-log plot!

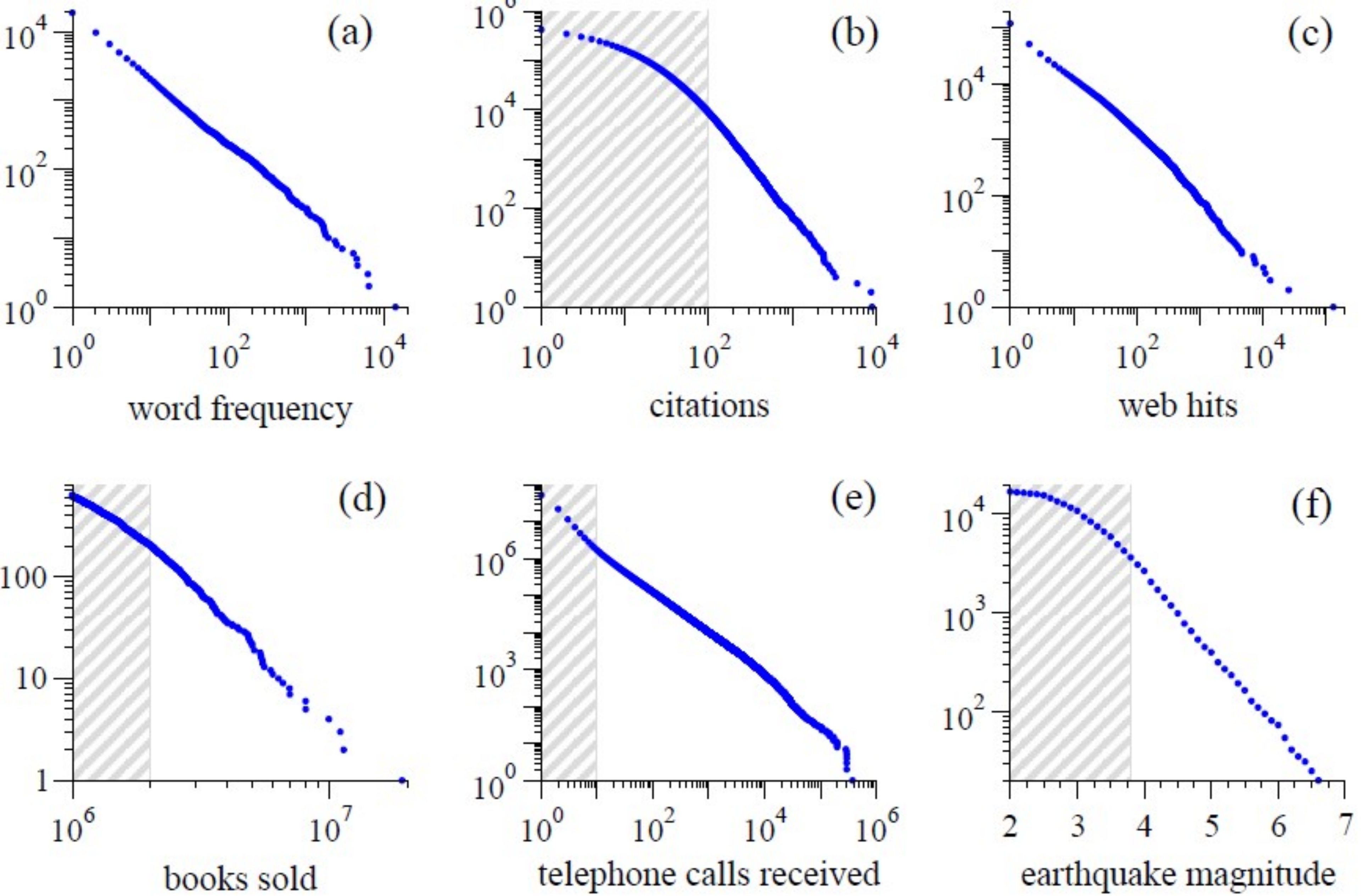
$y = b + cx$

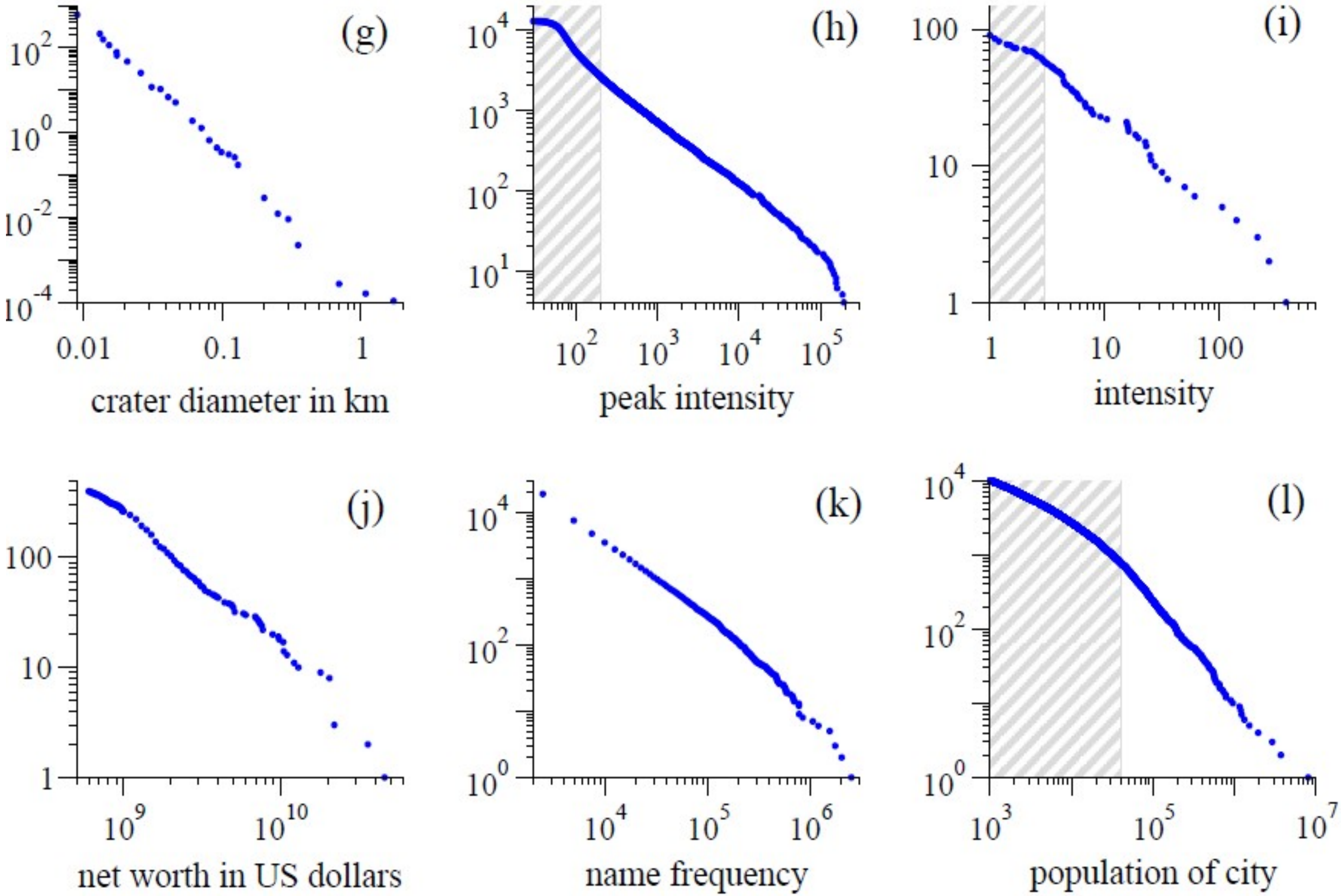


Power Law- Cnt.

- If power-law holds, the "log -log" plot should be a **straight line**.




(a)
word frequency
(b)
citations
(c)
web hits
(d)
books sold
(e)
telephone calls received
(f)
earthquake magnitude



Source: Power laws, Pareto distributions and Zipf's law. Newman. Contemporary physics 46.5. 2005.

Popularity

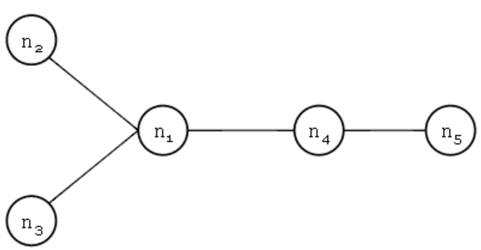
- Let's focus on the Web in which we can measure popularity accurately!
  - Popularity of a page

Popularity- Cnt.

- Let’s focus on the Web in which we can measure popularity accurately!
  - Popularity of a page ~ number of its **in-links**
    - Easy to count!

Degree Centrality- Cnt.

- A node is central if it has ties to many other nodes
  - Look at the node degree



$$C(n_1) = \sum_{j=1}^{n} A_{1j} = \sum_{i=1}^{n} A_{i1} = 3$$

	$n1$	n_2	n_3	n_4	n_5	$\sum_{j=1}^{n} A_{ij}$
n_1	0	1	1	1	0	**3**
n_2	1	0	0	0	0	**1**
n_3	1	0	0	0	0	**1**
n_4	1	0	0	0	1	**2**
n_5	0	0	0	1	0	**1**
$\sum_{i=1}^{n} A_{ij}$	**3**	**1**	**1**	**2**	**1**	

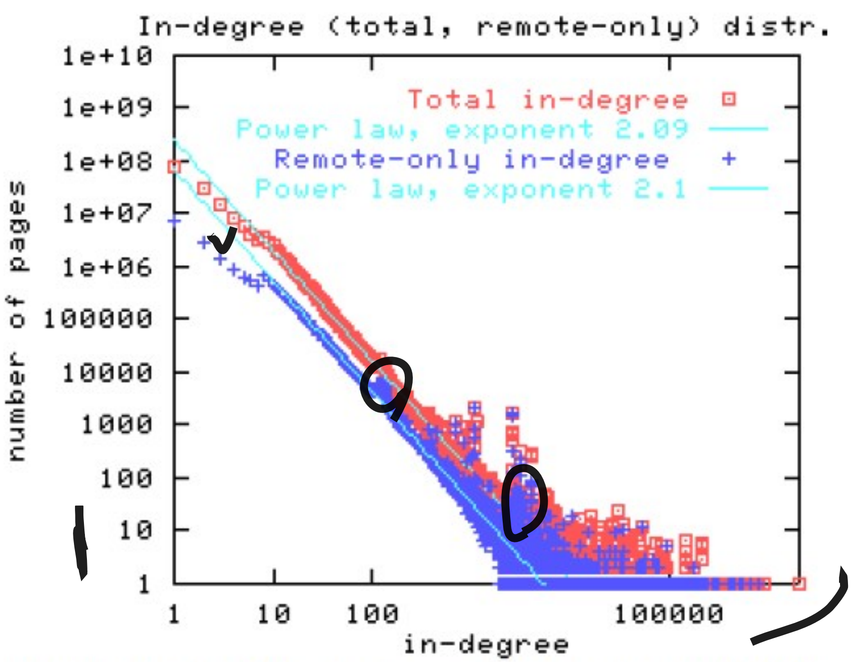
Adjacency Matrix (A)

Popularity- Cnt.

- Question:
  - What fraction of pages on the Web have *k* in-links?

Popularity- Cnt.

- Question:
  - What fraction of pages on the Web have *k* in-links?



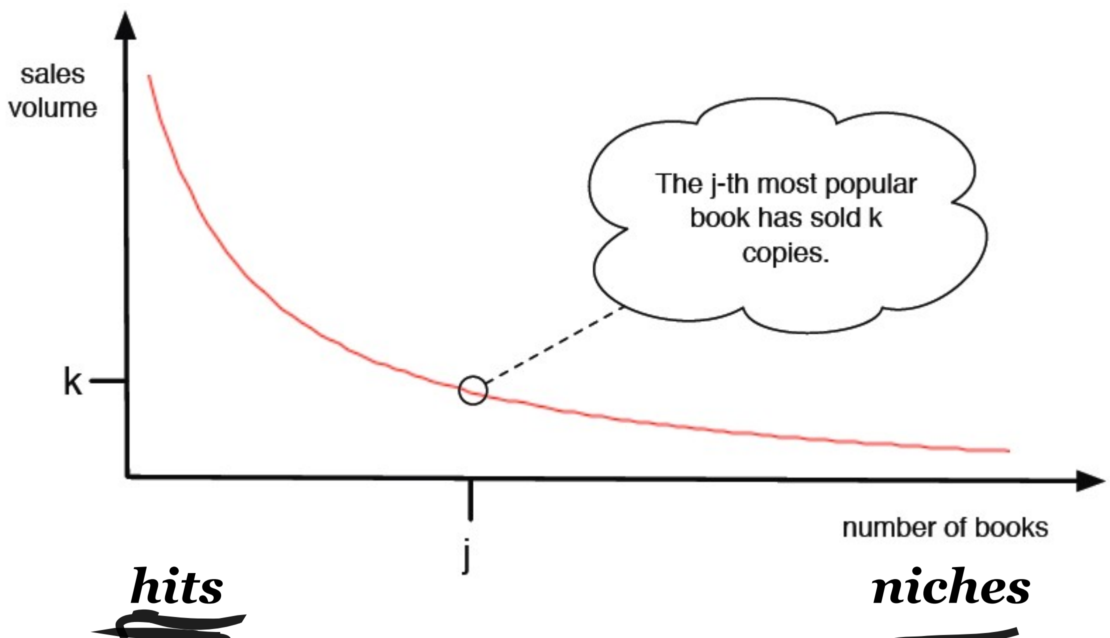
Remote-only: older crawl

- c ~= 2.1
- The anomalous bump at 120 on the x-axis is due to a large *clique** formed by a single spammer.

* Subset of nodes such that every two distinct nodes are adjacent (directly connected).

Popularity- The Long Tail

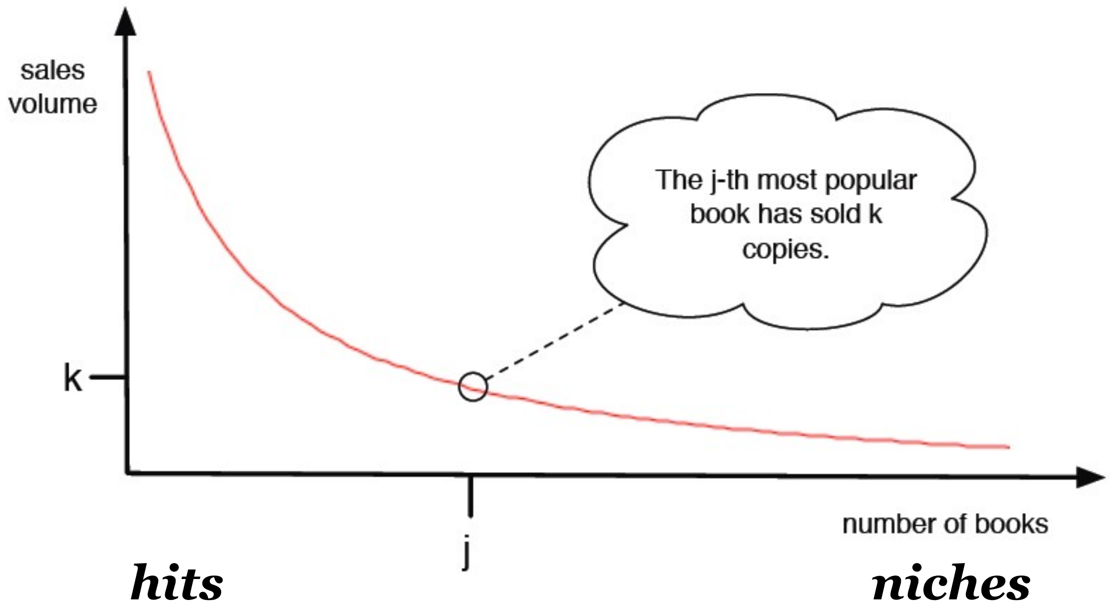
- **Question**: Are most sales generated by a
  - **small set** of **popular items** (***hits***), or
  - **large set** of **less popular items** (***niches***)?



Check if this curve is changing shape over time, adding more area under the right at the expense of the left!

Popularity- The Long Tail

- **Question**: Would personalization be useful?
  - E.g. through exposing people to items that (may not be popular but) match with their interests!



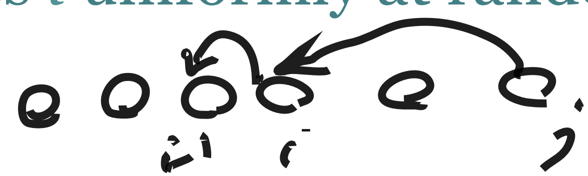
Popularity- Cause

- What is causing Power laws / Popularity?

Rich Get Richer (RGR)

Rich-Get-Richer: A simple model for the creation of links as a basis for power laws!

1. Pages (1,..,N) are created in order.
2. Each page creates one link to an **earlier** page.
3. When page *j* is created, it links to **page *i* < j**:
   a) With probability $\boxed{p}$, *j* chooses *i* uniformly at random & creates **a link to *i***.



   b) With probability (1- *p*), *j* chooses *i* uniformly at random & copies *i*'s decision (creates **a link to the page that *i* points to**).

RGR - Implication of Rule b

b) With probability (1- p), j chooses i uniformly at random & copies i's decision (creates **a link to the page that *i* points to**).

RGR - Implication of Rule b

b) With probability (1- p), j chooses i uniformly at random & copies i's decision (creates **a link to the page that *i* points to**)

The probability of linking to a page is proportional to the total number of pages that currently link to that page!

Copying mechanism → RGR

RGR – Power Law

- We observe power law, if we run this model for many pages
  - the fraction of pages with *k* in-links will be distributed according to a power law $1/k^c$!
  - Value of the exponent *c* depends on the choice of p.
- Relation between *c* and p?

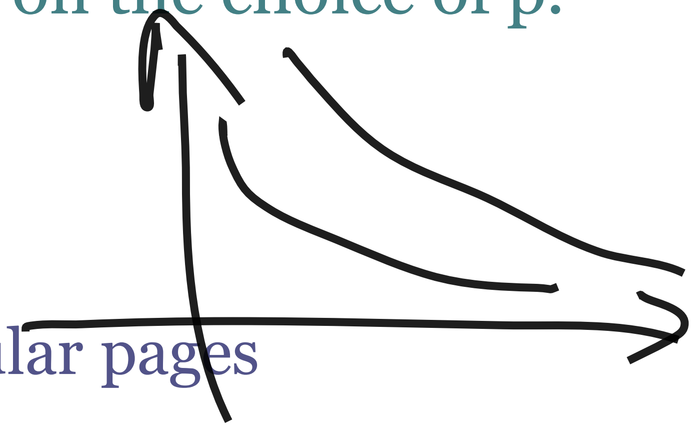
RGR

1 − p

RGR

RGR - Power Law

- We observe power law, if we run this model for many pages
  - the fraction of pages with *k* in-links will be distributed according to a power law $1/k^c$!
  - Value of the exponent *c* depends on the choice of p.
- Relation between *c* and p?
  - Smaller p / larger 1-p
    - Copying becomes more frequent
    - More likely to see extremely popular pages
    - *c* gets larger



RGR - Preferential Attachment

- Due to copying mechanism: the probability of linking to a page is proportional to the total number of pages that currently link to that page!

- Preferential Attachment:
  - **b)** With probability (1- p), j chooses i with probability **proportional to *i*'s current number of in-links** and creates a link to i.
    - links are formed preferentially to pages with high popularity.

RGR - Preferential Attachment

Rich-Get-Richer:

1. Pages (1,..,N) are created in order.
2. Each page creates one link to an **earlier** page.
3. When page *j* is created, it links to **page *i* < j**:
   a) With probability *p*, *j* chooses *i* uniformly at random and creates **a link to *i***.

   b) With probability (1- *p*), *j* chooses *i* with probability **proportional to *i*'s current number of in-links** and creates a link to *i*.

RGR - Probabilistic Model

- Probabilistic model
  - $X_j(t)$: number of in-links to node j at a time t

- Two points about $X_j(t)$
  1. Value of $X_j(t)$ at time t=j
     - $X_j(j) = 0$
       - node j starts with 0 in-link when it's first created at time j!
  2. Expected Change to $X_j(.)$ over time

Compute the probability that node j gains an in-link in step t+1?

RGR - Probabilistic Model

- Expected Change to $X_j(.)$ over time
  - Probability that node *j* gains an in-link in step t+1?
    - Happens if the newly created node *t+1* points to node *j*.
    - Two cases:
      1. With probability p, node t+1 links to an earlier node chosen uniformly at random:
         - Thus, node t + 1 links to node j with probability 1/t
      2. With probability 1 – p, node t+1 links to an earlier node with probability proportional to the node's current number of in-links.
         - At time t+1:
           - total number of links in the network?
             - t (one out of each prior node)
           - How many of them point to node j?
             - $X_j(t)$ (based on the definition)
           - Thus, node t + 1 links to node j with probability $X_j(t)/t$.

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}.$$

RGR - Probabilistic Model

- Deterministic approximation
  - Approximate $X_j(t)$—the # of in-links of node j—by a continuous function of time $x_j(t)$.

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}.$$

  - Model for rate of growth:

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}. \longrightarrow x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right].$$

RGR - Probabilistic Model

- Identifying power law in DA

$$x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right].$$

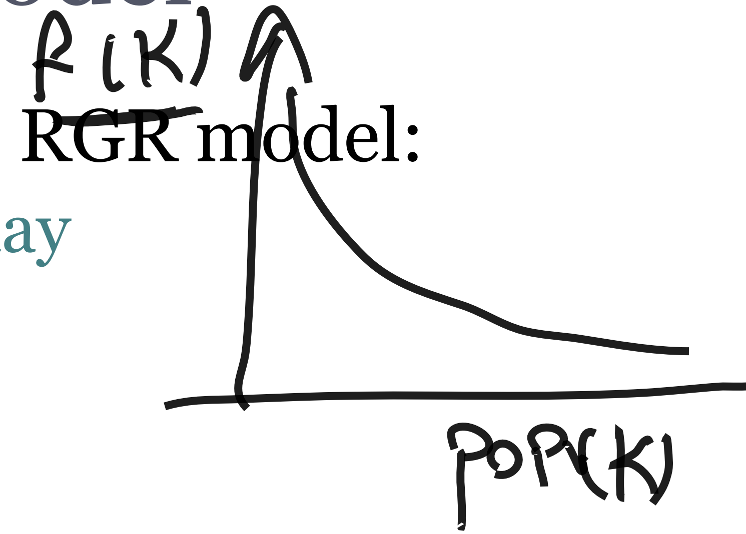
  - For a given value of *k* and time *t*, what fraction of nodes have at least *k* in-links at *t*, OR
  - For a given value of *k* and time t, what fraction of all *j*s satisfy $x_j(t) >= k$?

$$\left[\frac{q}{p} \cdot k + 1\right]^{-1/q}.$$

Power law:
The fraction of nodes with *at least k* in-links is proportional to $k^{-1/q}$.

RGR - Probabilistic Model

- Explain power laws using the RGR model:
  - phones receiving *k* calls per day
  - books bought by *k* people
  - papers with *k* citations
  - cities with population *k*
    - Cities grow in proportion to their size, simply as a result of people having children!



- Once an item becomes popular, the rich-get-richer dynamics are likely to push it even higher!

Reading

- Ch.18 Power Laws and Rich-Get-Richer Phenomena [NCM]