Power Laws & Rich Get Richer

Machine Learning with Graphs

Department of Computer Science
University of Massachusetts, Lowell
Spring 2021

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?

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!

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

$$\underbrace{\log(f(k))}_{y} = \underbrace{\log a}_{z} - \underbrace{c\log k}_{cx}$$

Power Law- Cnt.

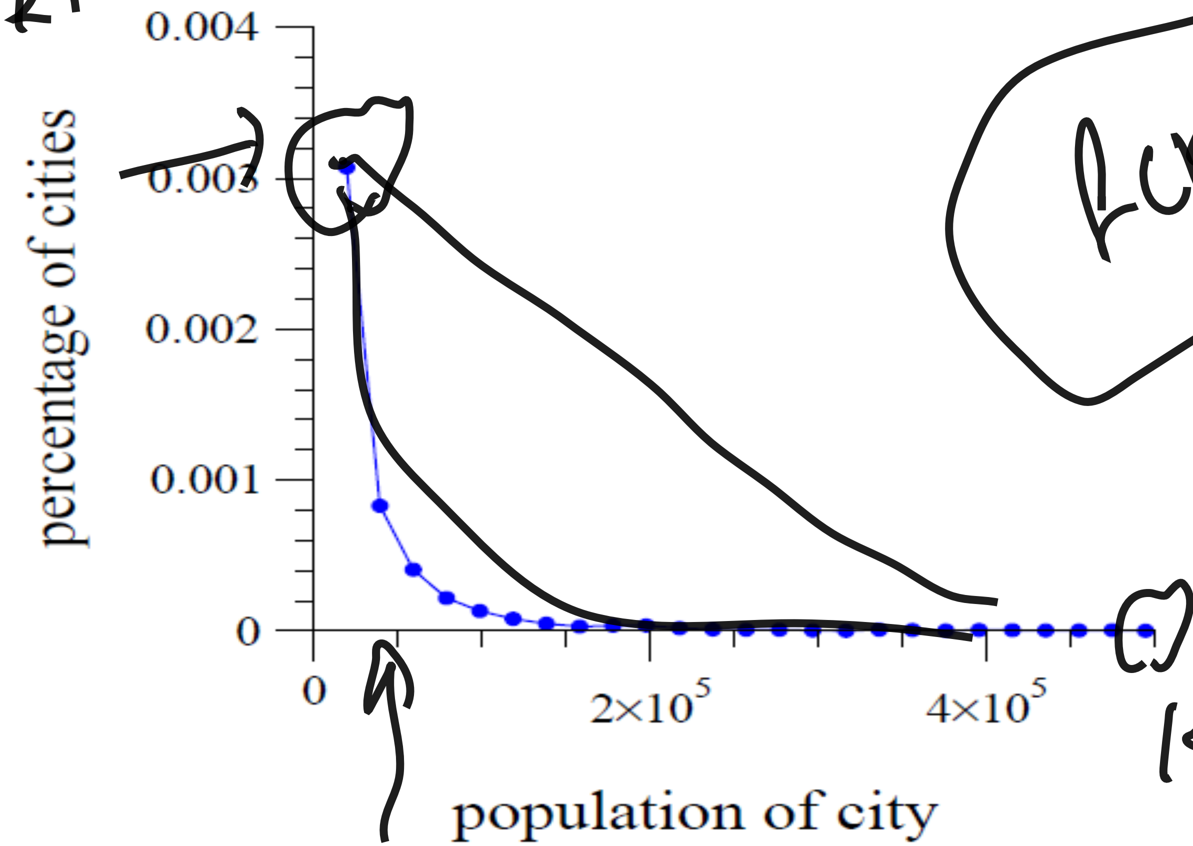
larger c

smaller c

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

c, a

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



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.*)

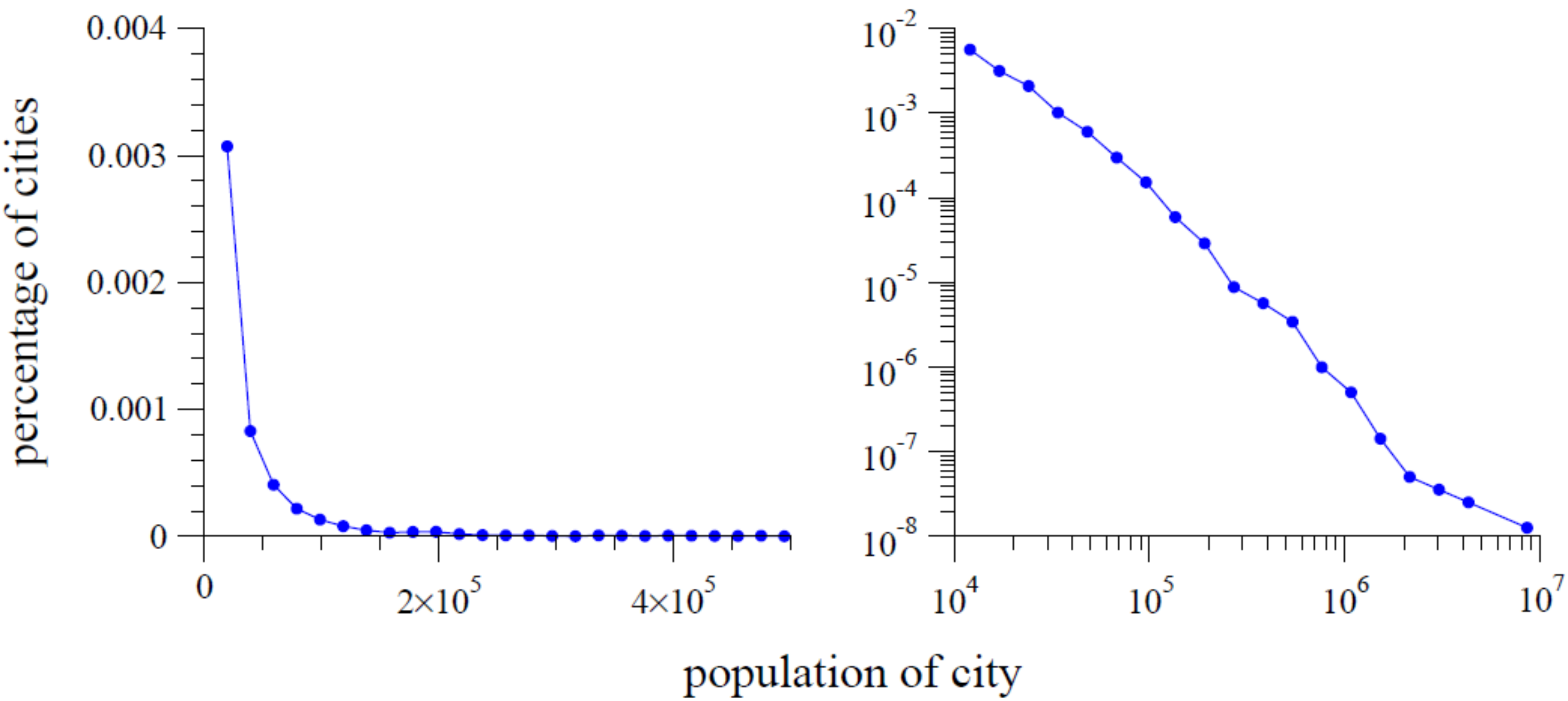
  2. Test if the result approximately estimates a power law a/k^c for some a and c, and if so, estimate the exponent c.

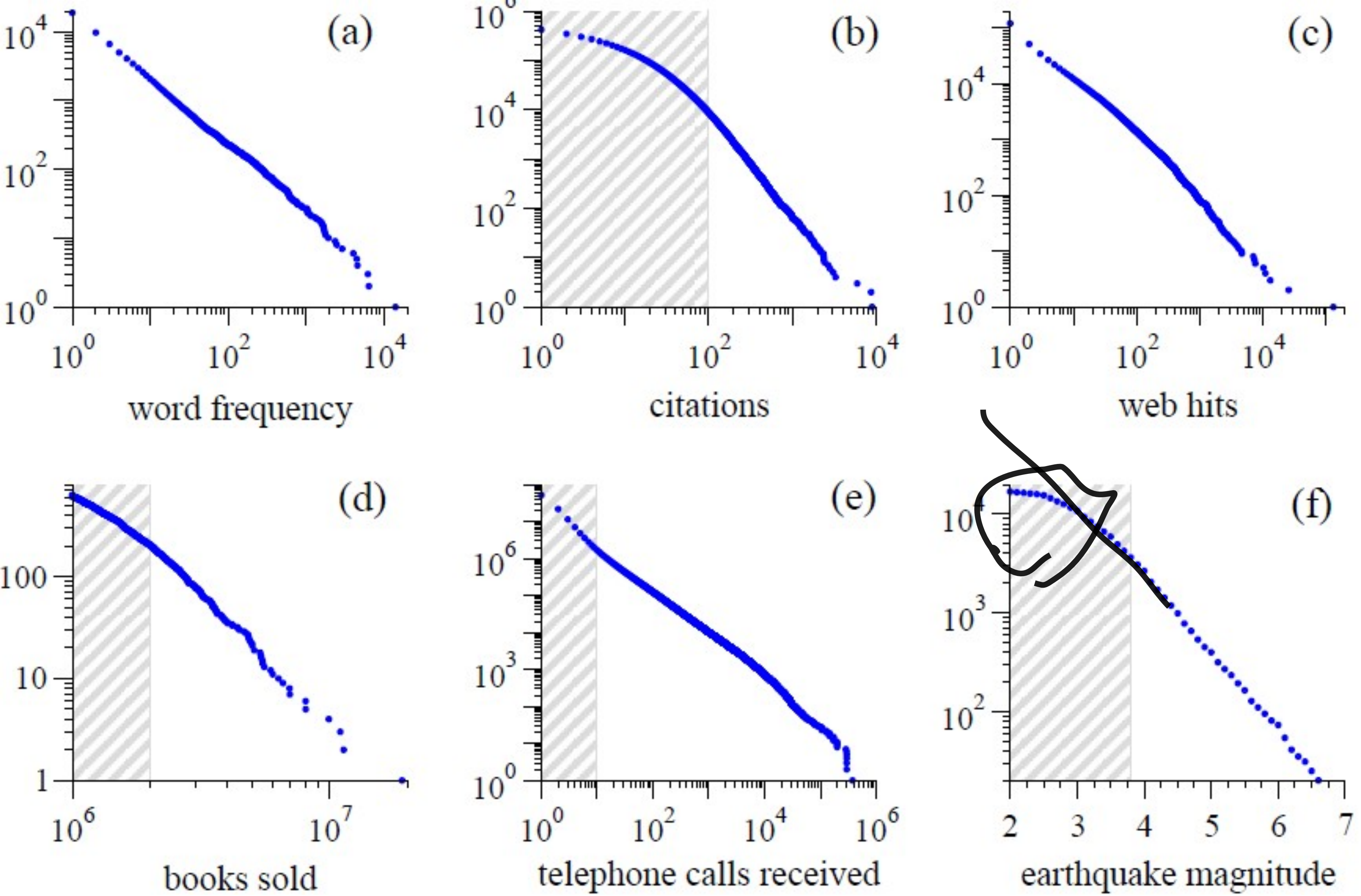
Power Law- Cnt.

- 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!

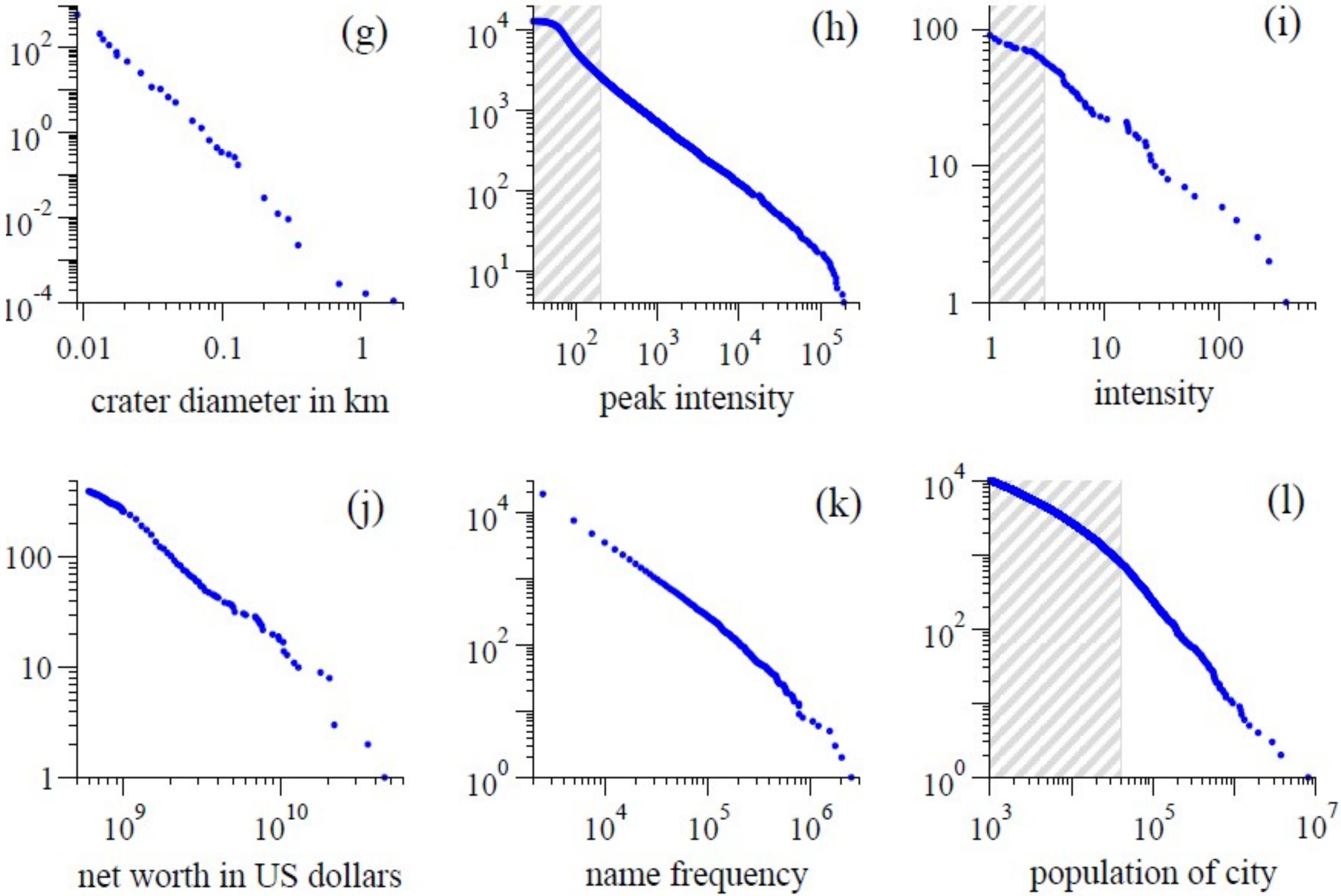
Power Law- Cnt.

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





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


(g)
crater diameter in km
(h)
peak intensity
(i)
intensity
(j)
net worth in US dollars
(k)
name frequency
(l)
population of city

Popularity

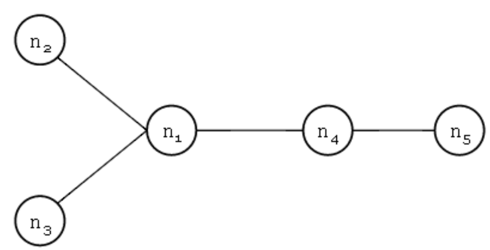
- Let’s focus on the Web where we can quantify popularity!
  - Popularity of a page

Popularity- Cnt.

- Let’s focus on the Web where we can quantify popularity!
  - 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



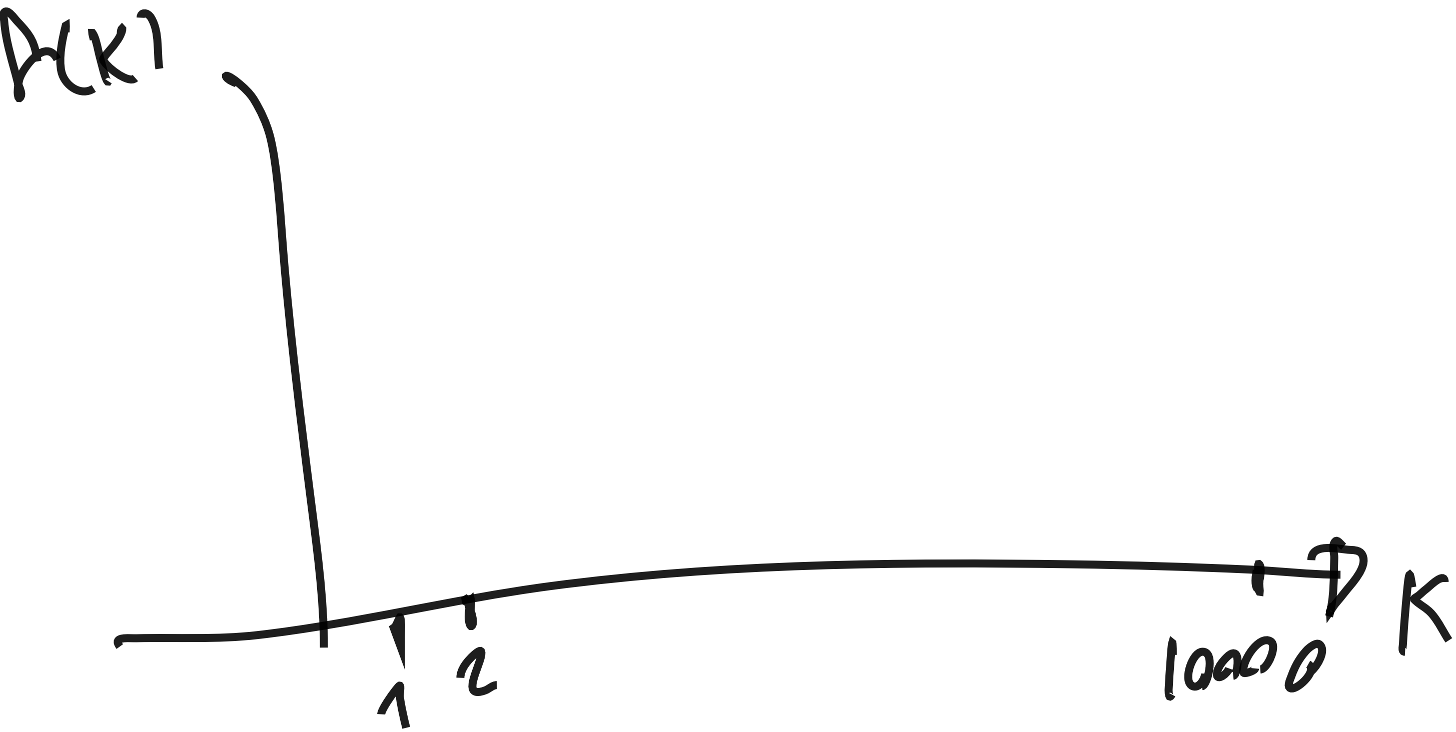
$$C(n_1) = \sum_{j=1}^{n} A_{1j} = \sum_{i=1}^{n} A_{i1} = 3$$

	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)

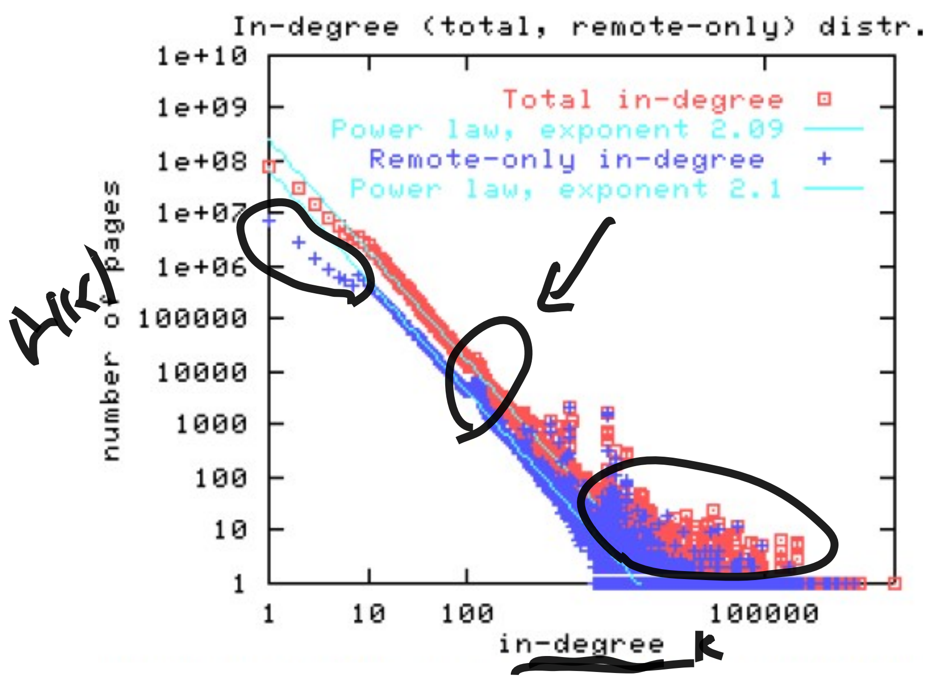
Popularity- Cnt.

- 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?



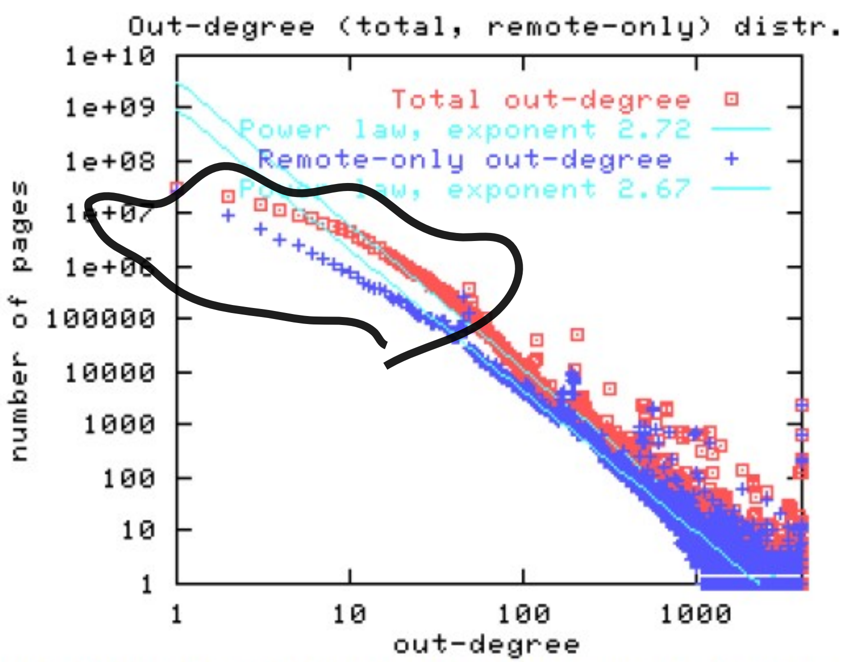
Remote-only: older crawl

- c ~= 2.1
- Straight lines are linear regressions for the best power law fit.
- 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.

Popularity- Cnt.

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

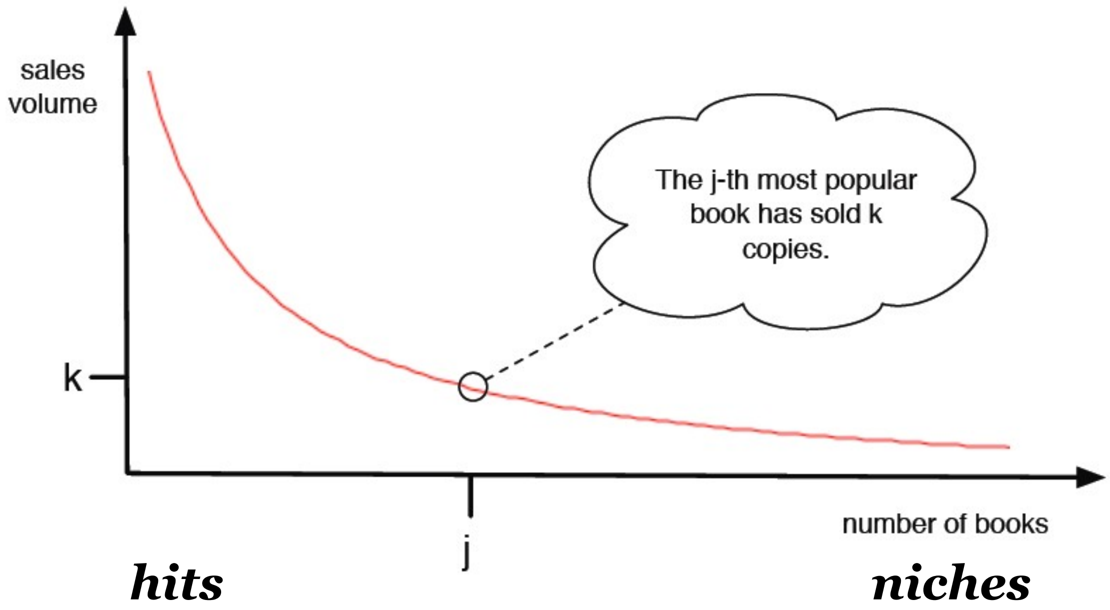


Remote-only: older crawl

- c ~= 2.7
- Initial segment of the out-degree distribution deviates significantly from the power law:
  - pages with low out-degree follow a different distribution.

Popularity- The Long Tail

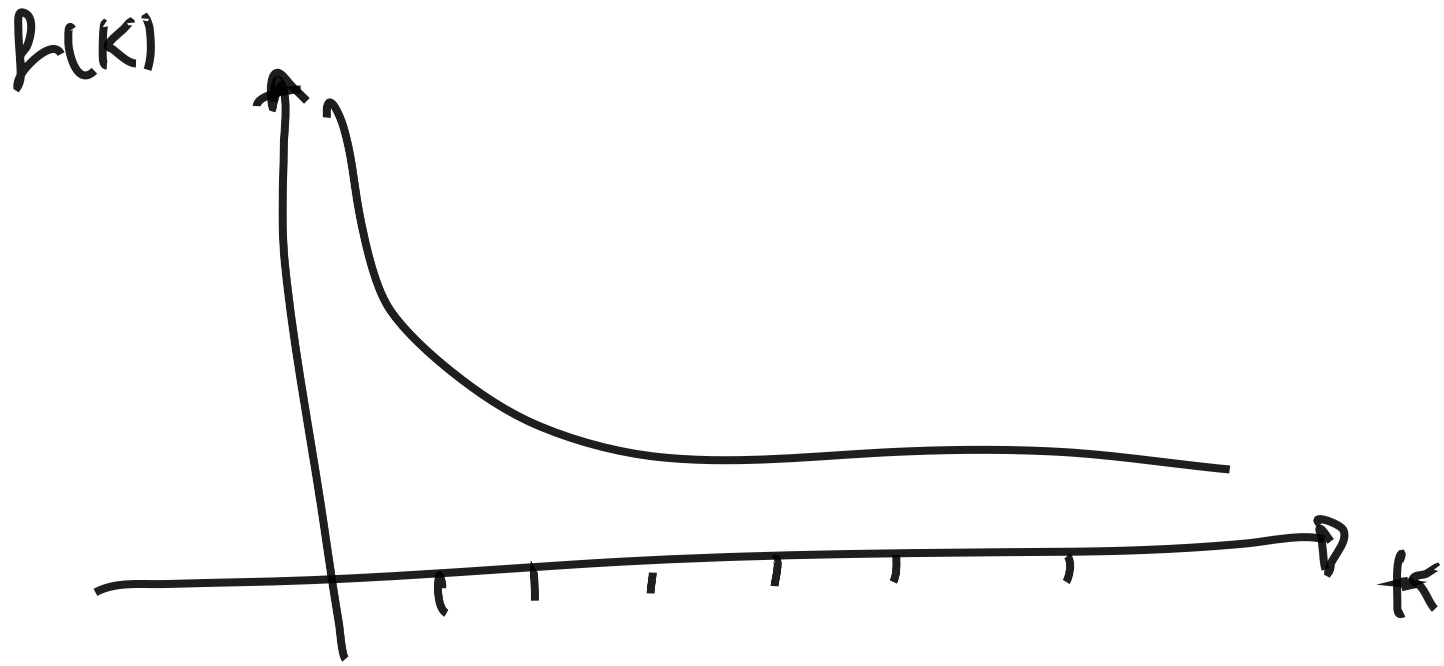
- **Question**: Are most sales generated by a
  - **small set** of **popular items** (***hits***), or
  - **large set** of **less popular items** (***niches***)?



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

Popularity- Cause

- What is causing Power laws / Popularity?



Rich Get Richer (RGR)

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

1. Pages are created in order and named 1, 2, …, N.
2. When page *j* is created, it produces a link to an **earlier page *i* < j** according to the following rules:
   a) With probability *p*, page *j* chooses page *i* uniformly at random, and creates **a link to *i***.
   b) With probability (1- *p*), page *j* chooses page *i* uniformly at random and creates **a link to the page that *i* points to** (copies decision made by *i*).

- Let's assume that each page creates just 1 link
  - We can extend this model to multiple links as well.

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.
- Correlation between *c* and p?

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.
- Correlation between *c* and p?
  - Smaller 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: restating rule 2 (b):
  - **b)** With probability (1- *p*), page *j* chooses page *i* with probability **proportional to *i*'s current number of in-links** and creates a link to *i*.
    - links are formed "preferentially" to pages that already have high popularity.

RGR - Preferential Attachment

Rich-Get-Richer:

1. Pages are created in order and named 1, 2, …, N.
2. When page *j* is created, it produces a link to an **earlier page *i* < j** according to the following rules:
   a) With probability *p*, page *j* chooses page *i* uniformly at random and creates **a link to *i***.
   b) With probability (1- *p*), page *j* chooses page *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

j 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

> - Explain power laws using the Rich-Get-Richer model!
> - 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?

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.

t+1

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

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 $x_j(t) >= k$?

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

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 Rich-Get-Richer model:
  - Fraction of phone receiving *k* calls per day: $1/k^2$
  - Fraction of books bought by *k* people: $1/k^3$
  - Fraction of papers with *k* citations: $1/k^3$
  - Fraction of cities with population *k*: $1/k^c$
    - 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]