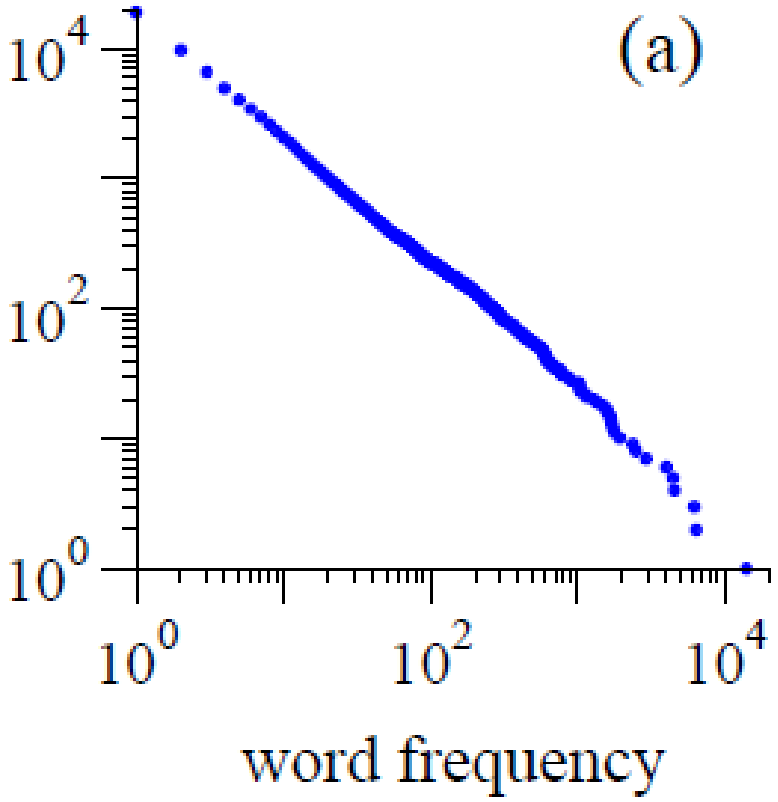
Behavioral Data Mining

Lecture 9
Modeling People

Outline

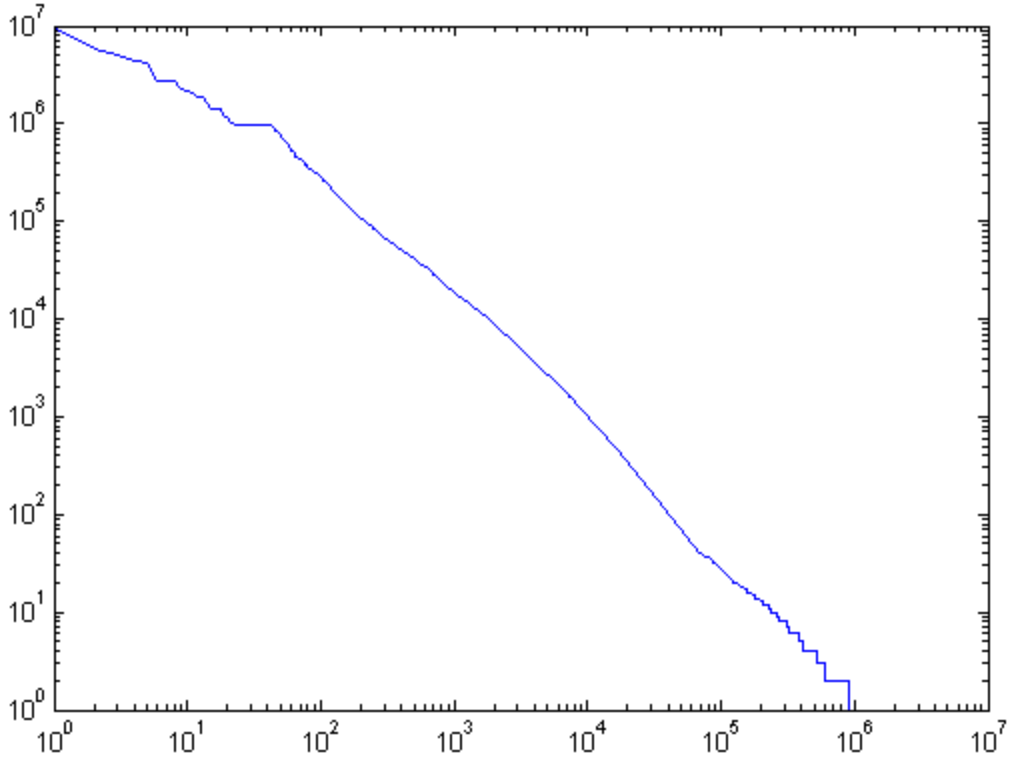
- Power Laws
- Big-5 Personality Factors
- Social Network Structure

Power Laws



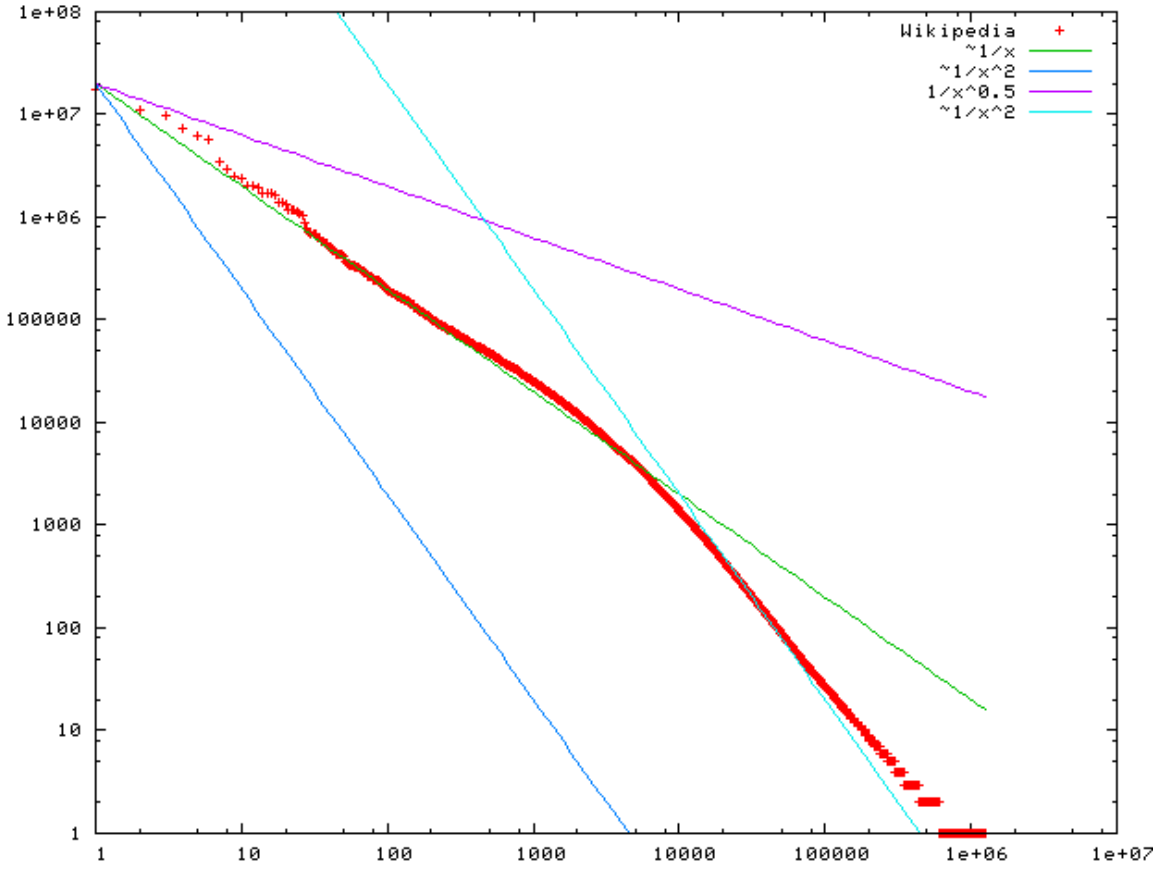
Y-axis = frequency of word, X-axis = rank in decreasing order

Power Laws



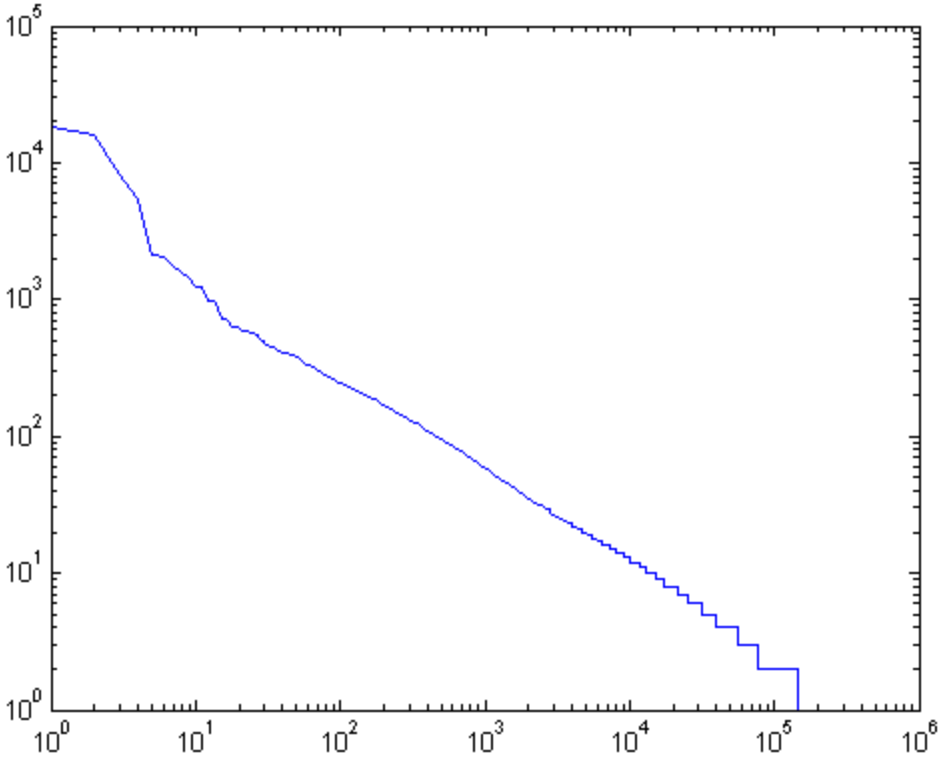
Y-axis = frequency of word, X-axis = rank in decreasing order

Wikipedia



Y-axis = frequency of word, X-axis = rank in decreasing order

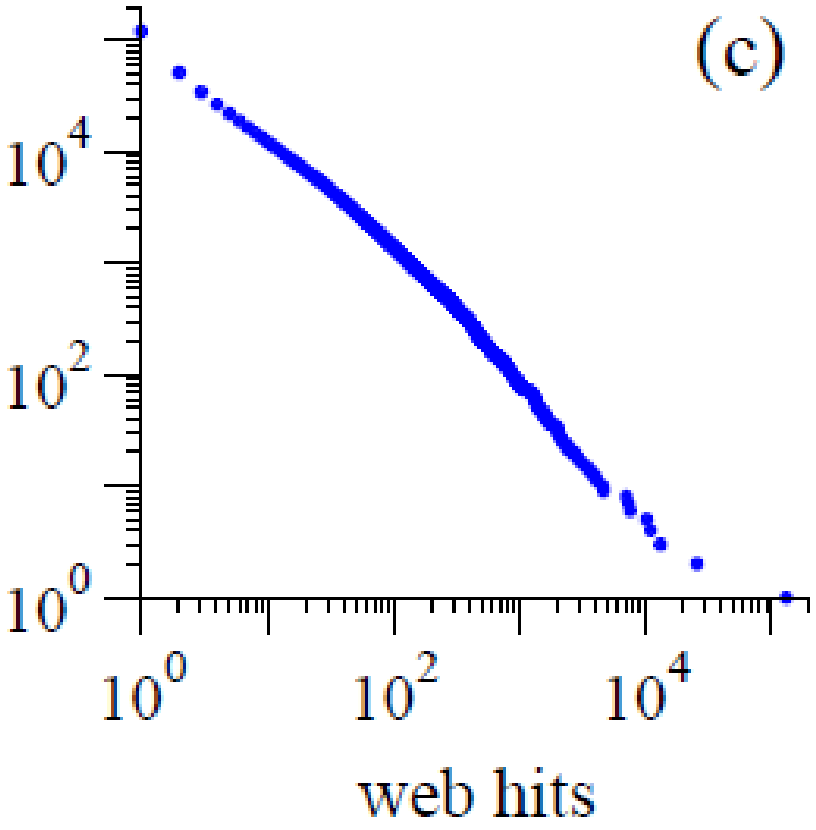
Power Laws – User Activity



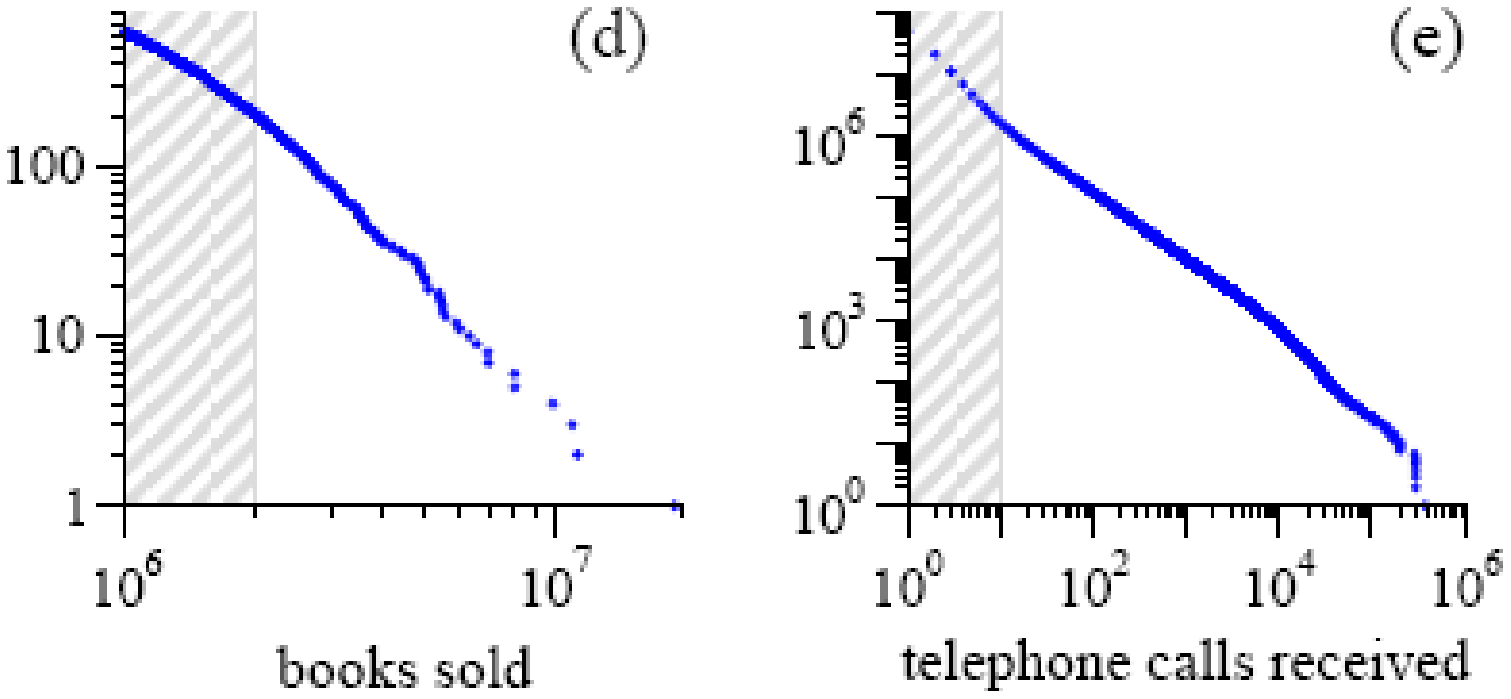
X-axis = rank of user, Y-axis = number of movie reviews

Very similar plots apply for Y = any action by users.

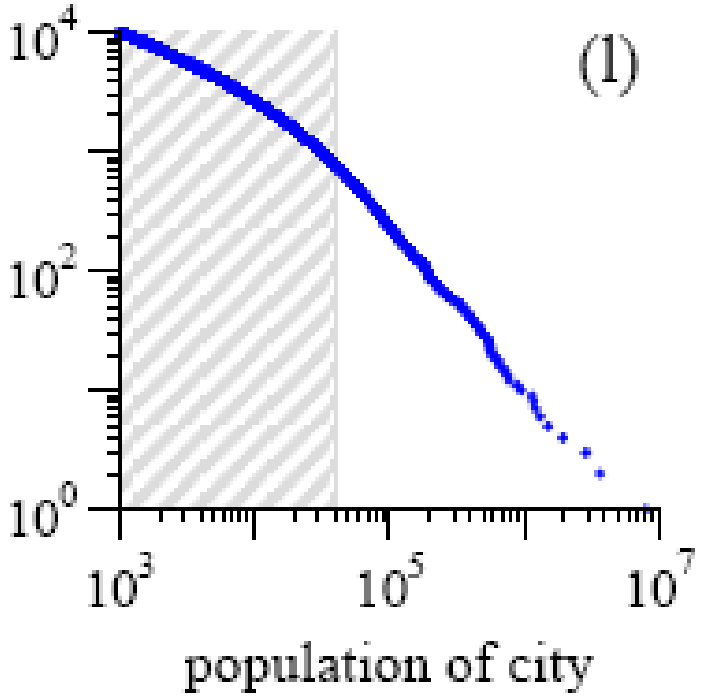
Power Laws



Power Laws


(d)
100
10
1
10^6
10^7
books sold
(e)
10^6
10^3
10^0
10^0
10^2
10^4
10^6
telephone calls received

Examples of Power Laws

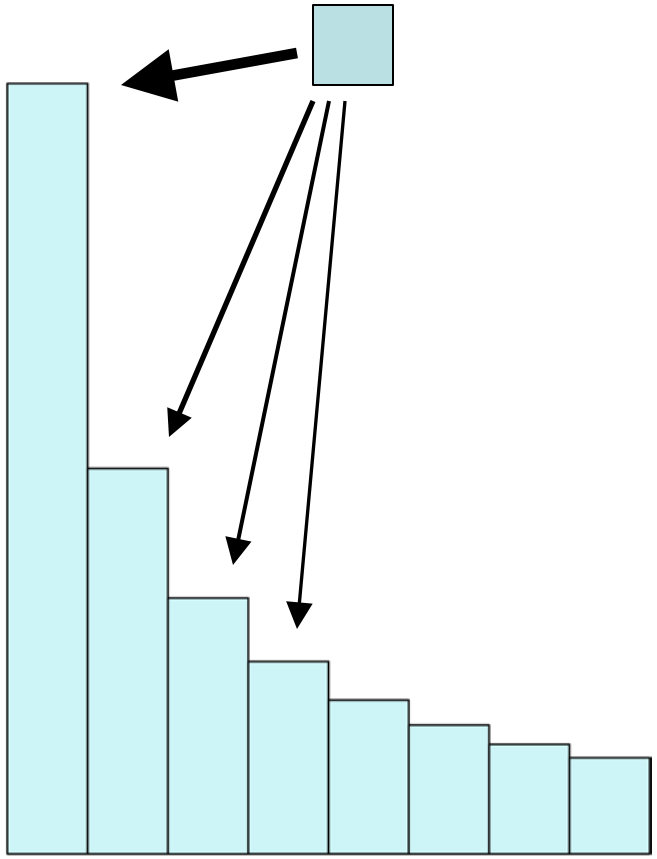


Examples of Power Laws

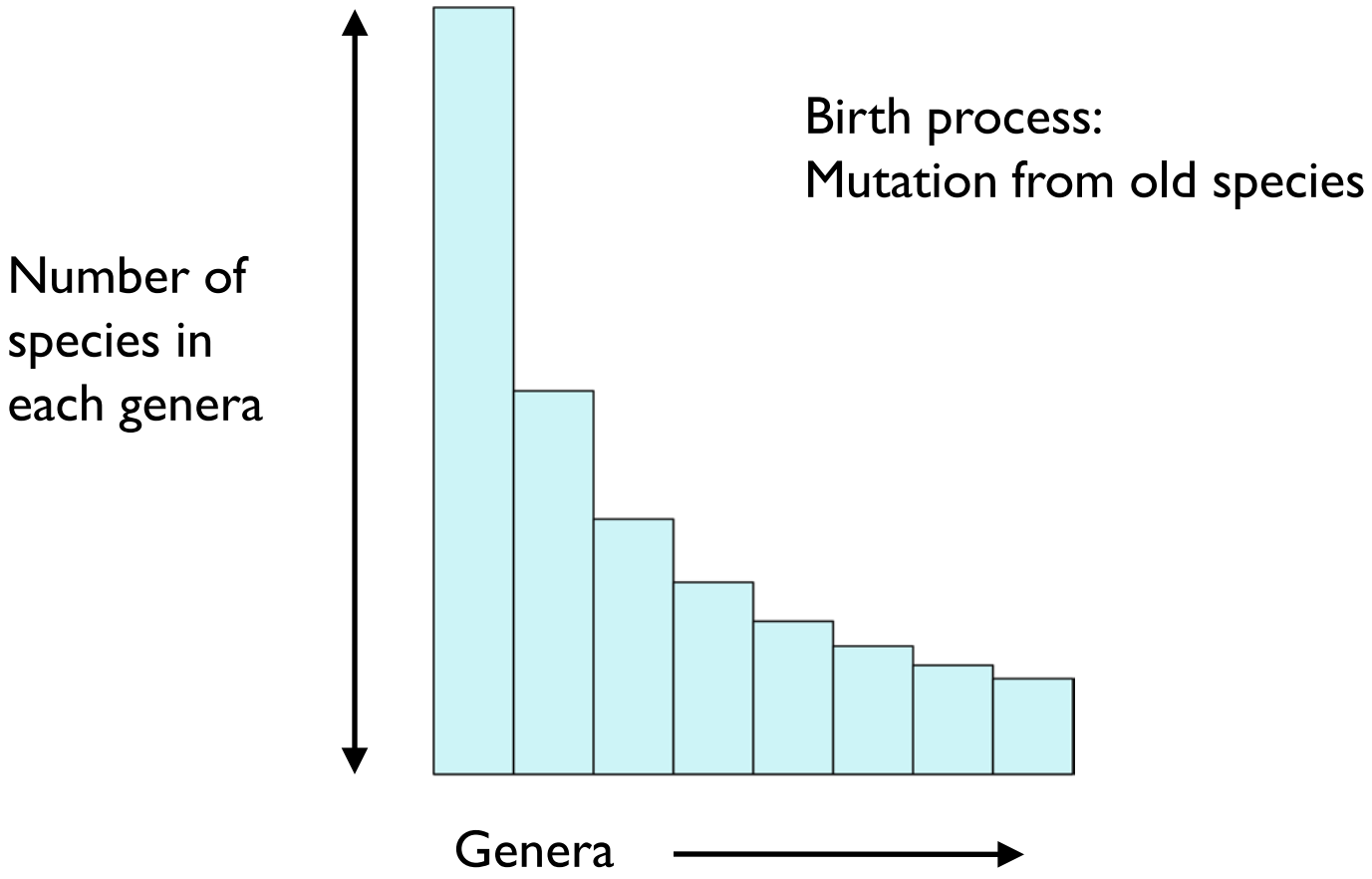
Also

- Number of users' Facebook friends
- Size of corporations
- The popularity of Facebook apps
- Net worth of individuals
- Number of pages in web sites
- Number of speakers of a language
- Number of links into a web site
- Number of sales of books, records, DVDs etc.
- Number of links out of a web site

Preferential Attachment



Yule's law (1925)



Simon's model of texts

Text is built by sampling earlier texts:

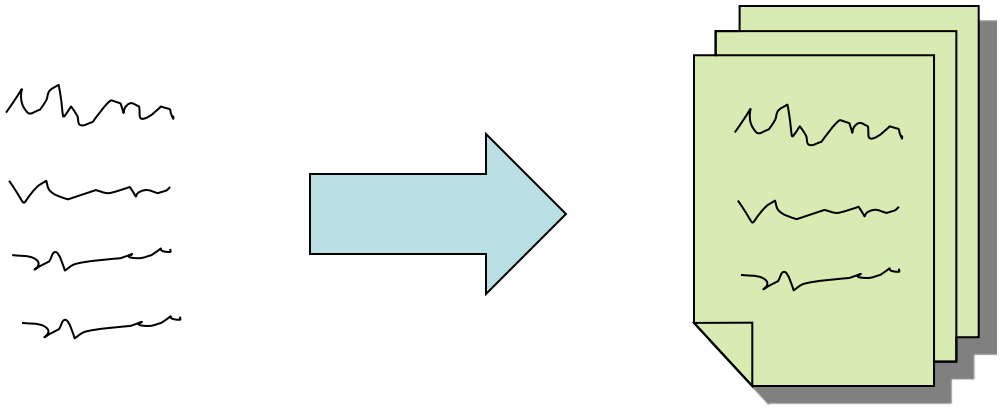
- Association: sampling earlier passages in the same corpus.
- Imitation: *"sampling segments of word sequences from other works he has written, from works of other authors, and, of course, from sequences he has heard."*

Applies to many other power law domains. The basic process is replication with a small amount of noise.

Simon's model of texts

Statified sampling:

Sampling and re-assembly of small segments of text. The choice of which segments to assemble does not have to be random.



Two forms of Power Law

Frequency vs. Rank: the examples we saw so far

Frequency vs. Count: Kleinberg's chapter and Simon's formulation

Count vs. Rank: $F \propto 1/R^{\alpha}$

Count vs. Frequency: $F \propto 1/C^{\beta}$

The two formulations are equivalent, and the constants are related by: $\beta = 1 + 1/\alpha$

Many processes cluster around $\alpha = 1$, or $\beta = 2$.

Working with Power Laws

Mean and variance may not be bounded. They are very unstable statistics in any case for power laws.

Median is not helpful either, its close to 1 for any $\alpha = 1$ distribution.

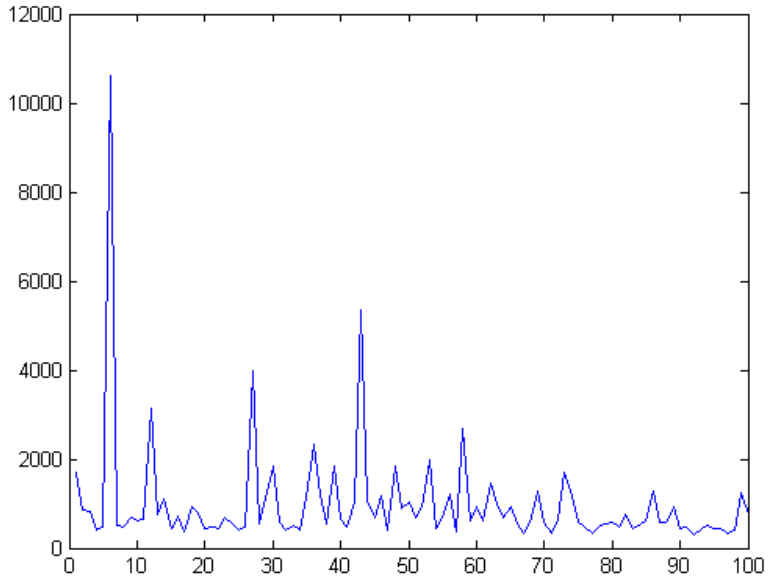
Best parametrizations are α or β and the Y-intercept.

Several estimation schemes. Simple curve fitting on the loglog plot ignores the density difference at large ranks.

A better approach is to use the Pareto cumulative distribution $\Pr(X > x)$, which inverts and smooths the frequency/rank plot.

Partitioning Power-Law Data

Usually, randomly partitioning a large dataset into pieces works well, as long as the partitions are large enough. Not so for power law data. We find instead that:

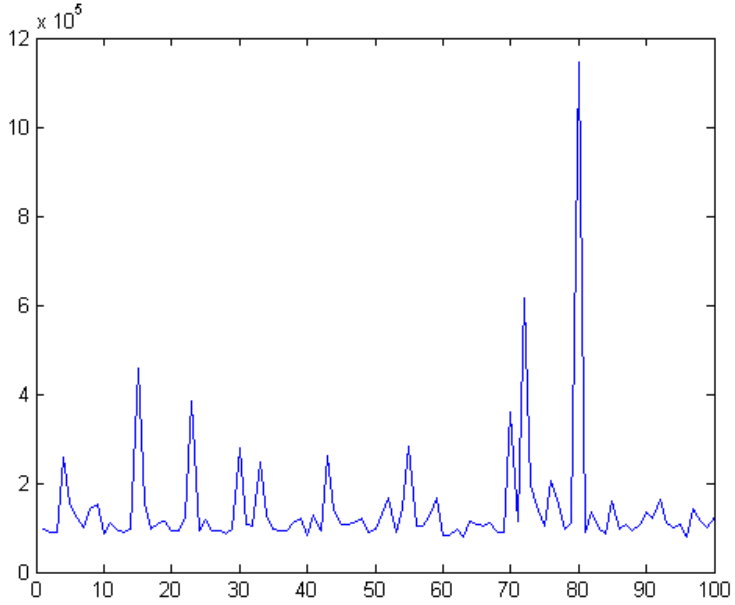


Bin totals for 10k power law values randomly split into 100 bins

Largest bin total about 15x times the average.

Partitioning Power-Law Data

Usually, randomly partitioning a large dataset into pieces works well, as long as the partitions are large enough. Not so for power law data. We find instead that:

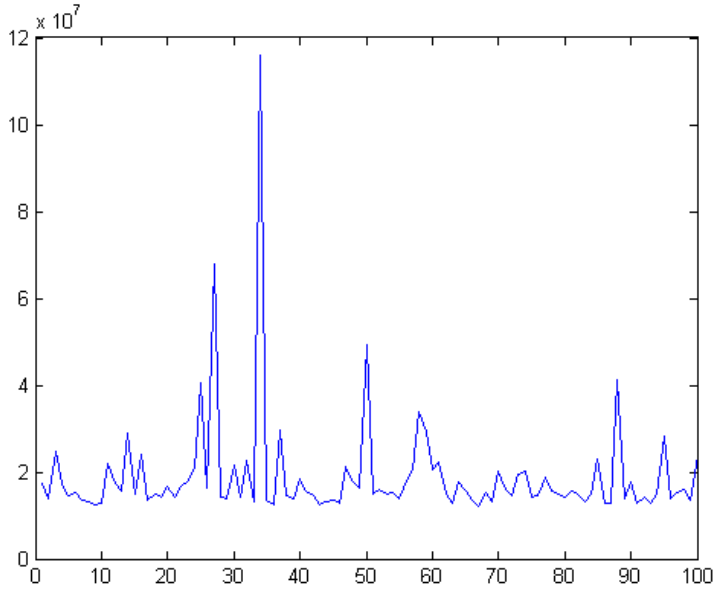


Bin totals for 1m power law values randomly split into 100 bins

Largest bin total approx 7 times the average.

Partitioning Power-Law Data

Usually, randomly partitioning a large dataset into pieces works well, as long as the partitions are large enough. Not so for power law data. We find instead that:

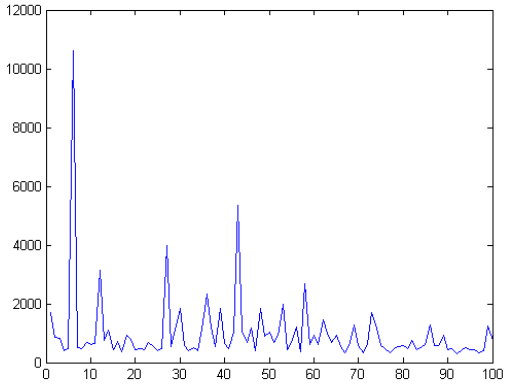


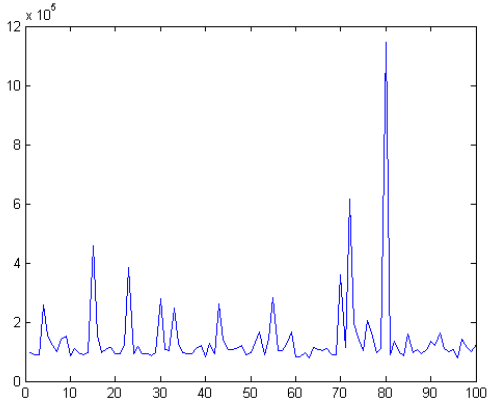
Totals for 100m power law values randomly split into 100 bins

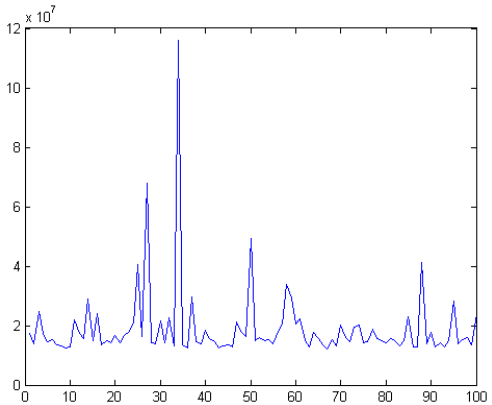
Largest bin total approx 6 times the average.

Partitioning Power-Law Data

Summary:







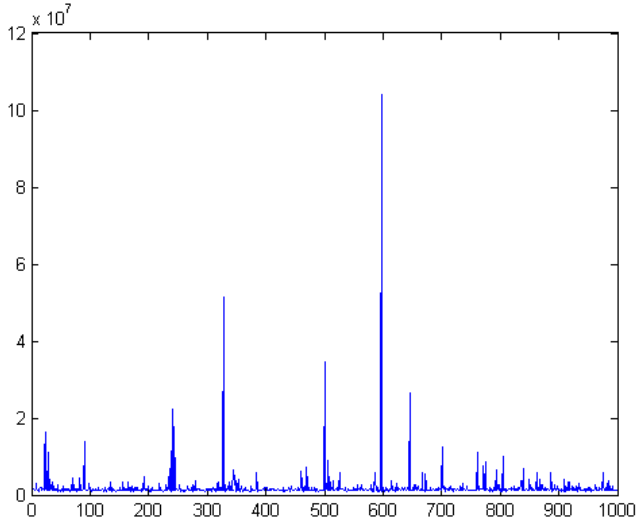
10k values 1million values 100 million values

With true power law data, this leads to "stragglers" in MapReduce algorithms which take many times longer than avg

What about using more bins?

Partitioning Power-Law Data

Try 100m values into 1000 bins:



Max/mean ratio = 50

Caveat: Sensitive to power law exponent, <1 is better, >1 worse

- Exponent often smaller for user partitioning
- Maybe larger for feature partitioning (examples earlier)

Ideas for workarounds?

Partitioning Power-Law Data

For very unbalanced data, it may work to randomly split large objects into several small ones – e.g. user data.

These objects can be randomly assigned to different partitions, and aggregated later.

Removing very large objects is often the best strategy (i.e. robot filtering). Especially if the power law for number of actions gets "steeper" at the high end indicating another statistical process in a mixture.

Outline

- Power Laws
- Big-5 Personality Factors
- Social Network Structure

Personality Factors

The "Big-5" are a set of personality traits derived from many different surveys and taxonomies.

- E Extroversion
- A Agreeableness
- C Conscientiousness
- N Neuroticism
- O Openness

Personality Factors

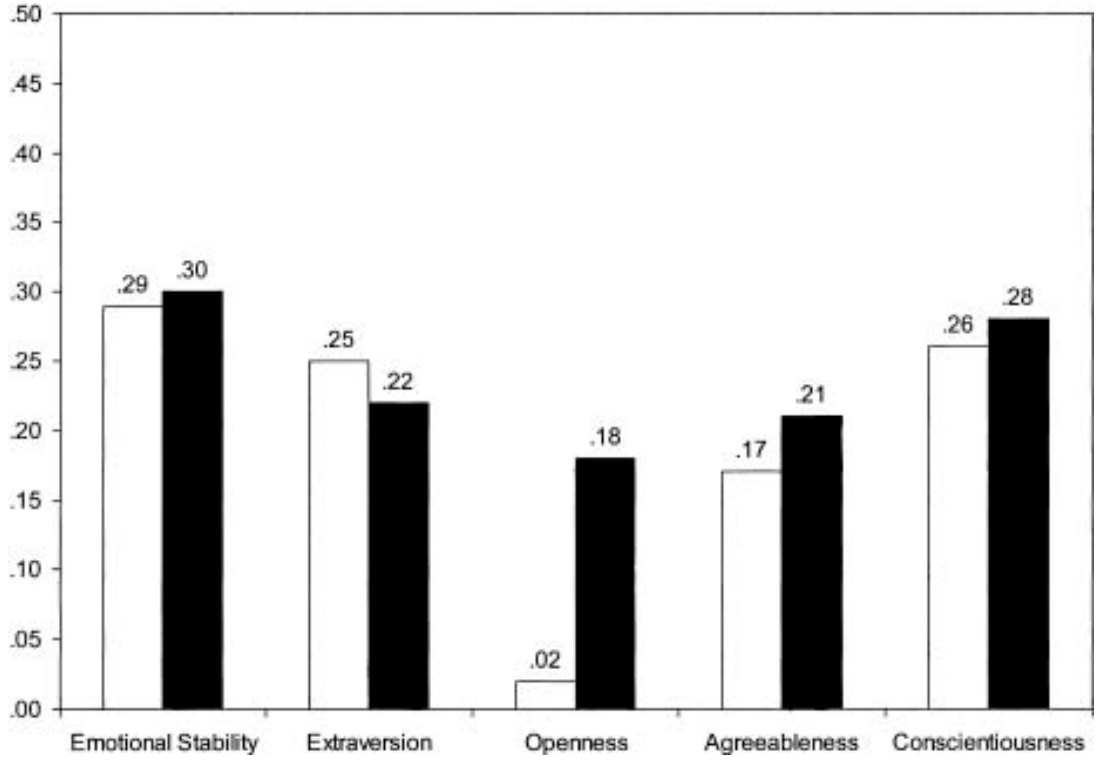
Big-5 predict many outcomes including:

- Group (interpersonal) dynamics
- Career outcomes – qualified by area
- Health, wellness, risk-taking, coping with disease
- Most effective persuasive messages:
  - Halko, S. and J.A. *Kientz*. "*Personality* and Persuasive Technology: An Exploratory Study on *Health*-Promoting Mobile Applications“
- They are mostly stable over one’s lifetime, but can change over many years.

Personality Factors

Solid bars – life satisfaction

Open bars – job satisfaction



Personality Factors

Big-5 factors show up in almost any sufficiently rich measurement domain. E.g. web sites visited, movies seen etc.

There are domain-specific aspects of user choice as well, but Big-5 will transfer to other domains as well.

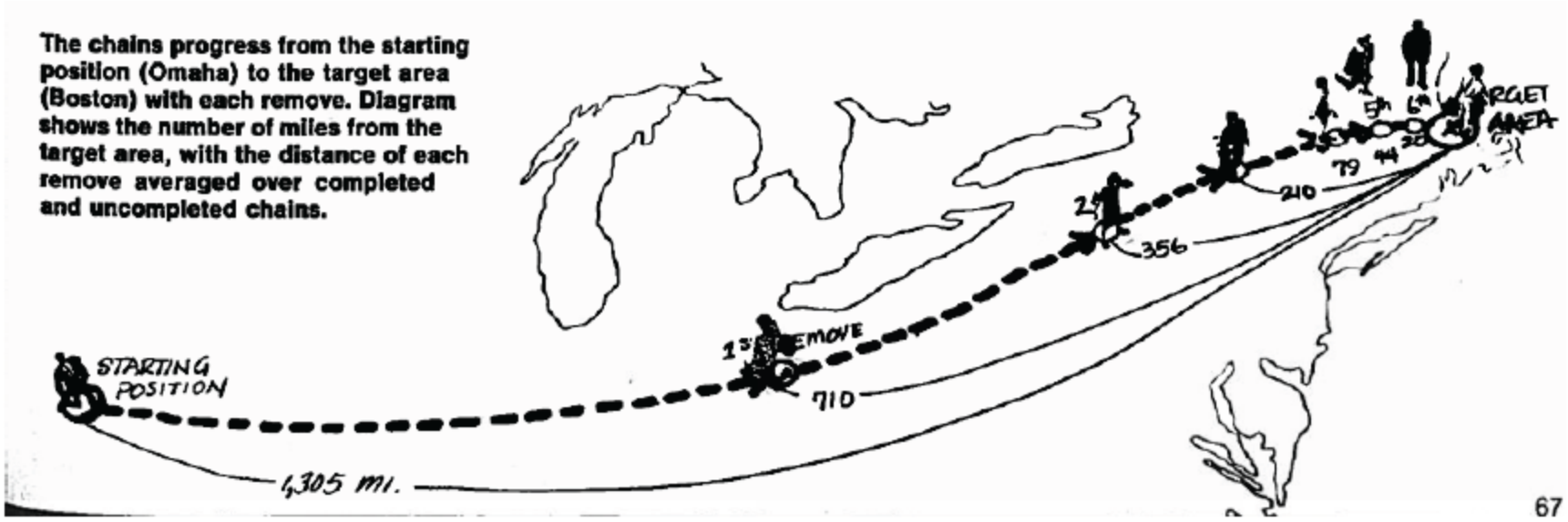
Big-5 are normally computed from questionnaires which limits their reach. But personality data from calibrated users can be used to estimate factors for uncalibrated users, e.g. by modeling web sites visited.

Outline

- Power Laws
- Big-5 Personality Factors
- Social Network Structure

6 Degrees of Separation

- Milgram's experiment found a median distance of 6 people for forwarding a letter to an unknown recipient in the US.

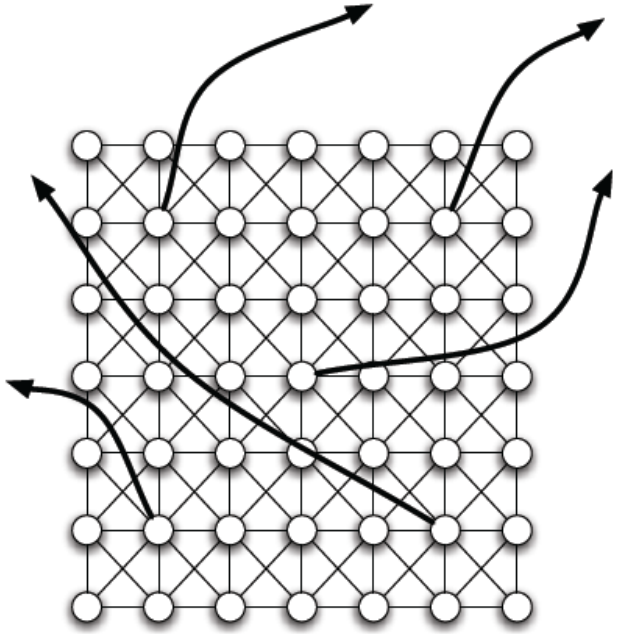


6 Degrees

- Milgram's experiment found a median distance of 6 people for forwarding a letter to an unknown recipient in the US.
- The most surprising aspect of this result was not the small diameter per se, but that people were able to find a short path with only local information.
- The links were found through a combination of geography and occupation.
- More recent experiments, e.g. by Microsoft Research, established the diameter at less than 7 world-wide, using instant messaging.

Small World Networks

- Watts-Strogatz model. Dense local ties plus a small number of random links creates a small-world network, i.e. Facebook + LinkIn
- Builds on the qualitative idea of "weak ties" (Granovetter) from social network theory.



Small World Networks

- The original Watts-Strogatz model only considers reachability.
- Kleinberg considered "searchability" assuming a greedy algorithm based on geography.
  - Each user forwards to their acquaintance who is as close as possible in geographic distance to the target.
- Kleinberg's random graph model uses a "Clustering Exponent" α and defines the probability of a link between two nodes (u,v) as

$$\Pr(l(u,v)) \propto d(u,v)^{-\alpha}$$

- He shows that such networks have uniquely short search time when $\alpha = 2$.

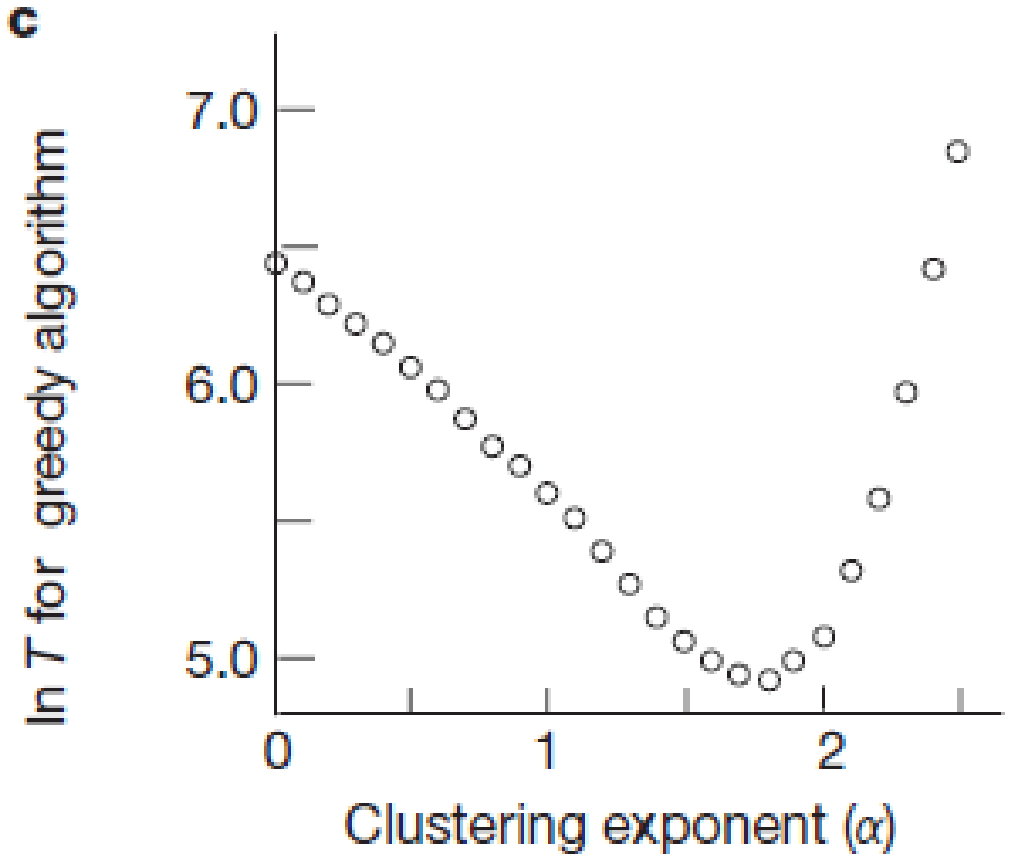
Small World Networks

- The original Watts-Strogatz model only considers reachability.
- Kleinberg considered "searchability" assuming a greedy algorithm based on geography.
  - Each user forwards to their acquaintance who is as close as possible in geographic distance to the target.
- Kleinberg's random graph model uses a "Clustering Exponent" α and defines the probability of a link between two nodes (u,v) as

$$\Pr(l(u,v)) \propto d(u,v)^{-\alpha}$$

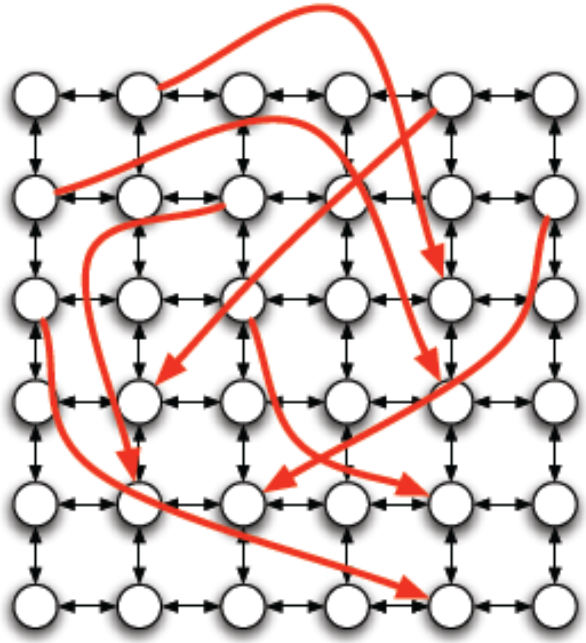
- He shows that such networks have uniquely short search time when $\alpha = 2$.

Small World Networks

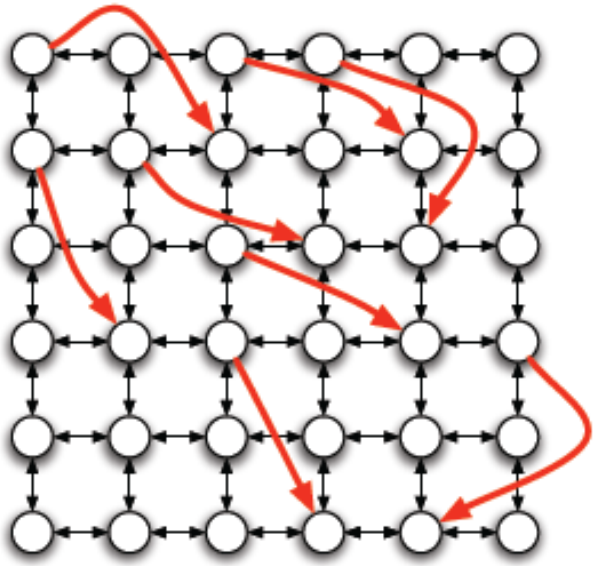


Small World Networks

- Kleinberg's model at α=2 actually has a larger number of short links compared to Watts-Strogatz (α=1)



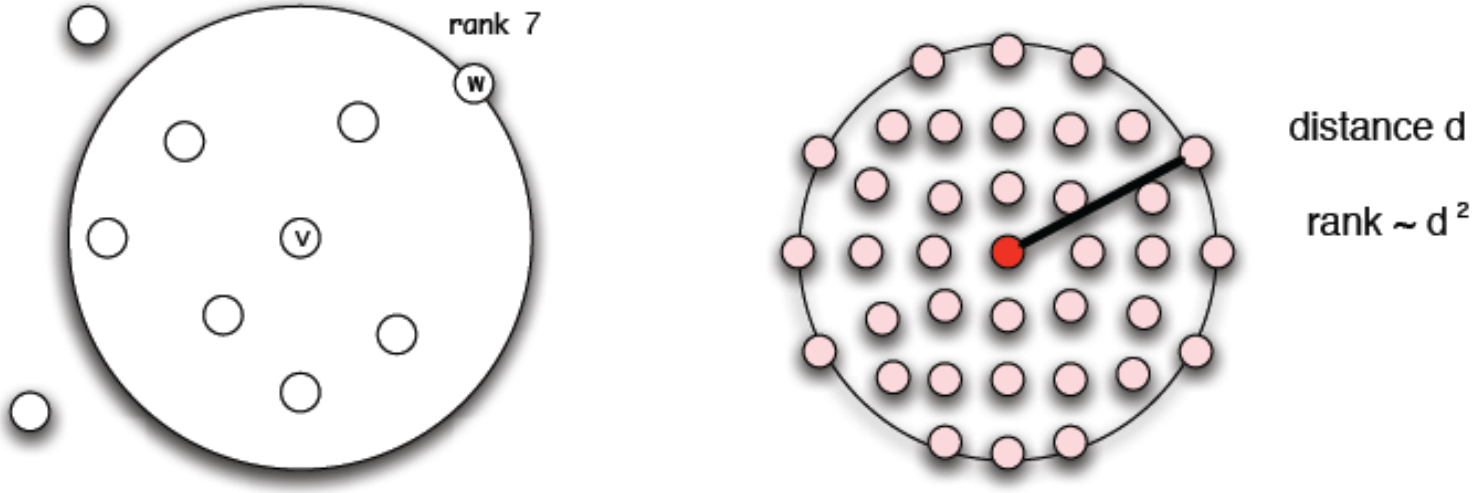
(a) *A small clustering exponent*



(b) *A large clustering exponent*

Small World Networks

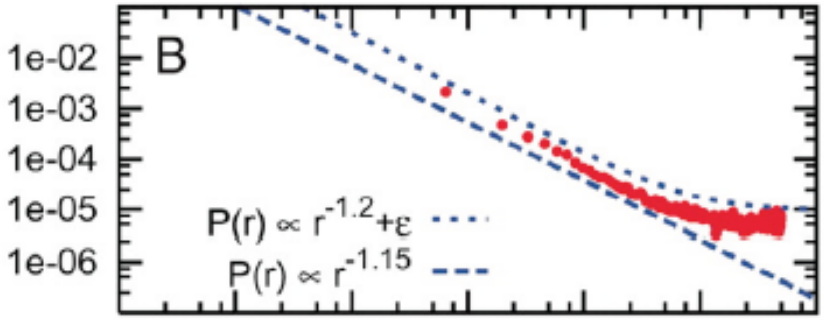
- Generalizing to real social networks – address non-uniformity of geography using rank distance.



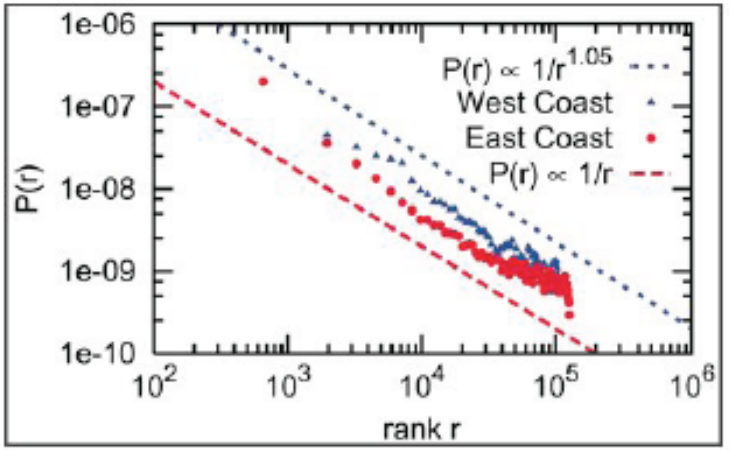
- The optimal cluster exponent in terms of rank distance is $\alpha = 1$, abstracting the original cluster exponent formula.

Small World Networks

- Empirical data from Livejournal matches this model fairly well: the exponent is around 1.15



(a) *Rank-based friendship on LiveJournal*



(b) *Rank-based friendship: East and West coasts*

Small World Networks

Uses of small-world network theory:

- Understanding the spread of ideas on social networks
- Understanding use of weak ties and indirect links for social ends
- Geographic and social models of how these networks are established.

Summary

- Power Laws: Examples, two formulations, and properties. Implications for work partitioning.
- Big-5 Personality Factors: Where they come from, implications.
- Social Network Structure: Small-world phenomena, navigation of networks.