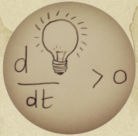
Lecture One (Possibly)

Stories of Complex Sociotechnical Systems: Measurement, Mechanisms, and Meaning
Lipari Summer School, Summer, 2012

Prof. Peter Dodds

Department of Mathematics & Statistics | Center for Complex Systems | Vermont Advanced Computing Center | University of Vermont

Licensed under the *Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License*.



Outline

Computational Story Lab:

Computational Story Lab:

Chris Danforth
Brian Tivnan

- NSF, NASA, MITRE
- 3000 processors + 100 TB storage at the Vermont Advanced Computing Core
- 100 TB storage in Danforth's office.

Something of a plan:

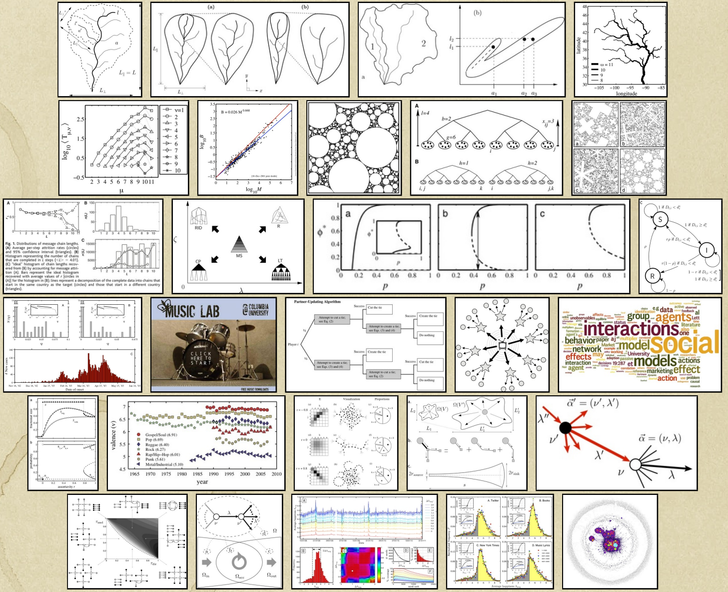
- Lecture 1: Complexity; Networks, and Social Search
  - Theory, Experiments.
- Lecture 2: Measuring Happiness
  - Big Data.
- Lecture 3: Social Contagion and Influence
  - Theory, Experiments, Big Data.

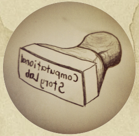
About these slides:

- Three versions (all in pdf):
  1. Presentation,
  2. Flat Presentation,
  3. Compact version (3x2).
- Presentation versions are navigable and hyperlinks are clickable.
- Web links look like this (⊞).
- References in slides link to full citation at end. [2]
- Citations contain links to papers in pdf (if available).
- 60 hours of lectures → 3 hours.
- Brought to you by a concoction of LaTeX, Beamer, perl, and madness.

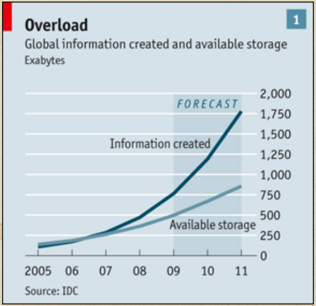
Two graduate level courses:

- Principles of Complex Systems (⊞), University of Vermont
- Complex Networks (⊞), University of Vermont




Computational Story Lab

The Rise of the Data Scientist: (⊞)



- Exponential growth: $\sim$ 60% per year.

Big Data Science:

- 2013: year traffic on Internet estimate to reach 2/3 ZB (1ZB=10^3EB=10^6PB=10^9TB)
- Large Hadron Collider: 40 TB/second.
- 2016—Large Synoptic Survey Telescope: 140 TB every 5 days.
- Facebook: $\sim 10^{11}$ photos
- Twitter: $\sim 10^{11}$ tweets

Data, Data, Everywhere—The Economist, Feb 25, 2010 (⊞)

No really, that's a lot of data

Data inflation 2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: *The Economist*

Basic Science $\simeq$ Describe + Explain:

Lord Kelvin (possibly):

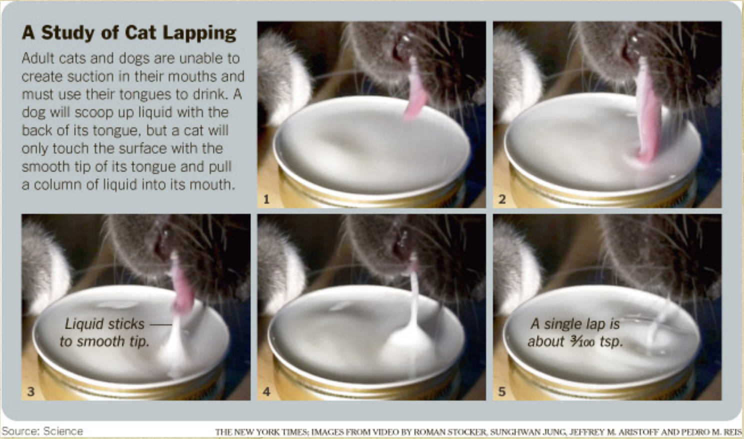
- “To measure is to know.”
- “If you cannot measure it, you cannot improve it.”

Bonus:

- “X-rays will prove to be a hoax.”
- “There is nothing new to be discovered in physics now, All that remains is more and more precise measurement.”

Whimsical but great example of real science:

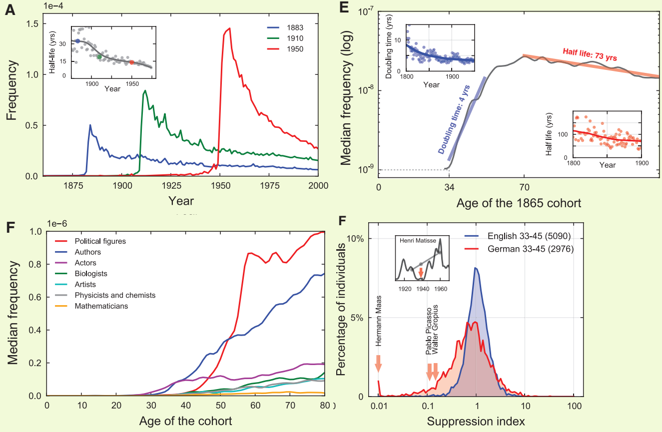
"How Cats Lap: Water Uptake by *Felis catus*" (⊞)
Reis et al., *Science*, 2010.



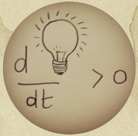
Three pieces: Observation + Experiment + Theory
Amusing interview here (⊞)

Big Data—Culturomics:

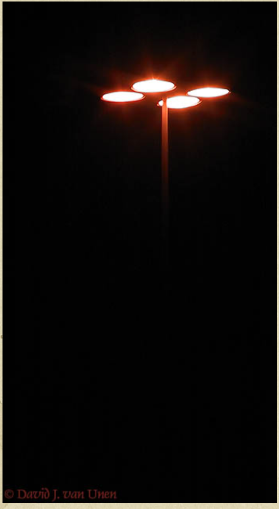
"Quantitative analysis of culture using millions of digitized books" by Michel et al., Science, 2011 [20]



http://www.culturomics.org/ (⊞)
Google Books ngram viewer (⊞)



What matters and what's measurable:



Science in the age of Big Data:

- Goal: Match Observation with Theory with Experiment.
- Traditional Engine: Cycle of hypothesis formation and testing.
- The boost: Data driven detection of stories.

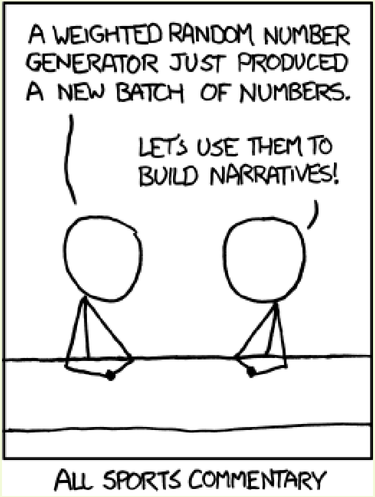
Four Thinkings for Big Data Storytellers:

1. Probabilistic Thinking (statistics)
2. Mechanistic Thinking (statistical physics)
3. Algorithmic Thinking (computer science)
4. Data Visualization Thinking (art, graphic design)

Framing issues:

- “Data Scientist” implies “Describes but does not explain.”

Homo narrativus—We are story-telling machines:



- Mechanisms = Evolution equations, algorithms, stories, ...
- "Also, all financial analysis. And, more directly, D&D."

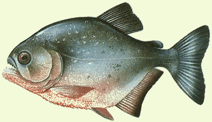
http://xkcd.com/904/ (⊞)

Complexity Manifesto:

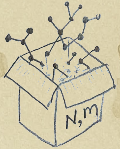
1. Systems are ubiquitous and systems matter.
2. Consequently, much of science is about understanding how pieces dynamically fit together.
3. 1700 to 2000 = Golden Age of Reductionism.
   - Atoms!, sub-atomic particles, DNA, genes, people, ...
4. Understanding and creating systems (including new ‘atoms’) is the greater part of science/engineering.
5. Universality: systems with quantitatively different micro details exhibit qualitatively similar macro behavior.
6. Computing advances make the Science of Complexity possible:
   - 6.1 We can measure and record enormous amounts of data, research areas continue to transition from data scarce to data rich.
   - 6.2 We can simulate, model, and create complex systems in extraordinary detail.

Revolution: Big Data & Complex Networks

- Many complex systems can be viewed as complex networks of physical or abstract interactions.
- Opens door to mathematical and numerical analysis.
- Mindboggling amount of work published on complex networks since 1998...
- Why all this ‘new’ research on networks?
- Answer: Incredible Amounts of Data.
- ... largely due to your typical theoretical physicist:



- *Piranha physicus*
- Hunt in packs.
- Feast on new and interesting ideas (see chaos, cellular automata, ...)



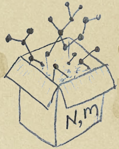
Popularity according to Google Scholar:

"Collective dynamics of 'small-world' networks" [31]

- Watts and Strogatz
  Nature, 1998
- Cited 16,157 times (as of June 19, 2012)

"Emergence of scaling in random networks" [3]

- Barabási and Albert
  Science, 1999
- Cited 13,984 times (as of June 19, 2012)



The UNIVERSITY of VERMONT

Networks and creativity:

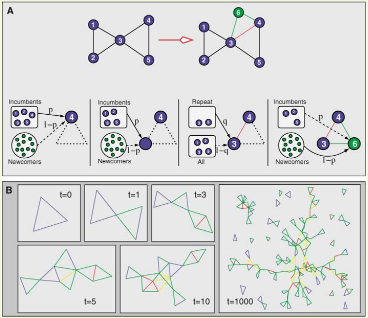
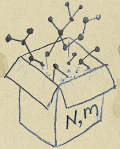


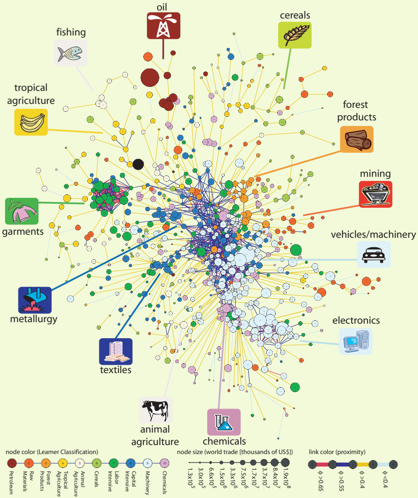
Fig. 2. Modeling the emergence of collaboration networks in creative enterprises. (**A**) Creation of a team with $m = 3$ agents. Consider, at time zero, a collaboration network comprising five agents, all incumbents (blue circles). Along with the incumbents, there is a large pool of newcomers (green circles) available to participate in new teams. Each agent in a team has a probability p of being drawn from the pool of incumbents and a probability $1 - p$ of being drawn from the pool of newcomers. For the second and subsequent agents selected from the incumbents' pool: (i) with probability q, the new agent is randomly selected from among the set of collaborators of a randomly selected incumbent already in the team; (ii) otherwise, he or she is selected at random among all incumbents in the network. For concreteness, let us assume that incumbent 4 is selected as the first agent in the new team (leftmost box). Let us also assume that the second agent is an incumbent, too (center-left box). In this example, the second agent is a past collaborator of agent 4, specifically agent 3 (center-right box). Lastly, the third agent is selected from the pool of newcomers; this agent becomes incumbent 6 (rightmost box). In these boxes and in the following panels and figures, blue lines indicate newcomer-newcomer collaborations, green lines indicate newcomer-incumbent collaborations, yellow lines indicate new incumbent-incumbent collaborations, and red lines indicate repeat collaborations. (**B**) Time evolution of the network of collaborations according to the model for $p = 0.5$, $q = 0.5$, and $m = 3$.

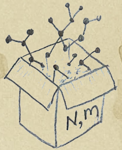
- Guimerà et al., Science 2005: [15] "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance"
- Broadway musical industry
- Scientific collaboration in Social Psychology, Economics, Ecology, and Astronomy.



The Evolution of Economies:

- Hidalgo et al.'s "The Product Space Conditions the Development of Nations" [16]
- How do products depend on each other, and how does this network evolve?
- How do countries depend on each other for water, energy, people (immigration), investments?

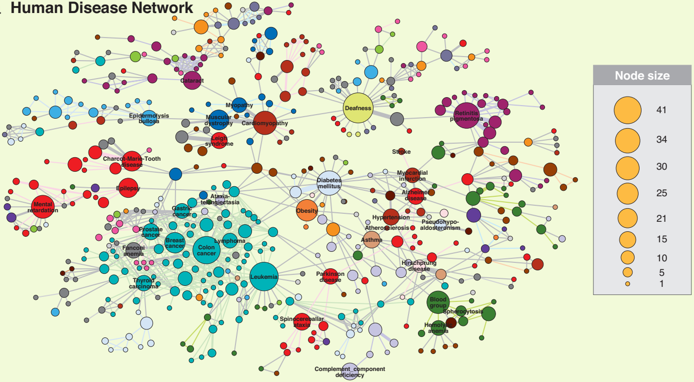


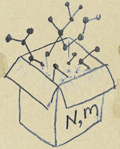


Networks of diseases:

- The human disease and disease gene networks (Goh *et al.*, 2007):

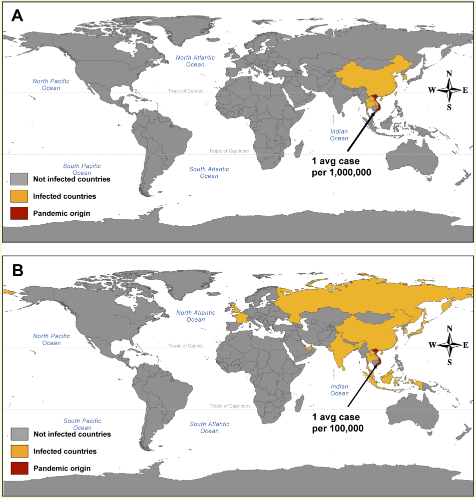
a **Human Disease Network**

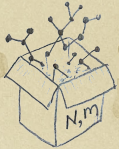




Disease contagion:

"Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions" Colizza *et al.*, PLoS Medicine 2007. [10]





Social Contagion:

Controversial work by Fowler and Christakis et al. on social contagion of:

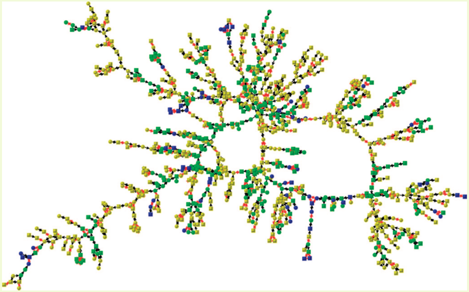
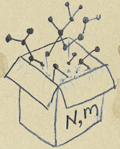


Figure 1. Loneliness clusters in the Framingham Social Network. This graph shows the largest component of friends, spouses, and siblings at Exam 7 (centered on the year 2000). There are 1,019 individuals shown. Each node represents a participant, and its shape denotes gender (circles are female, squares are male). Lines between nodes indicate relationship (red for siblings, black for friends and spouses). Node color denotes the mean number of days the focal participant and all directly connected (Distance 1) linked participants felt lonely in the past week, with yellow being 0–1 days, green being 2 days, and blue being greater than 3 days or more. The graph suggests clustering in loneliness and a relationship between being peripheral and feeling lonely, both of which are confirmed by statistical models discussed in the main text.

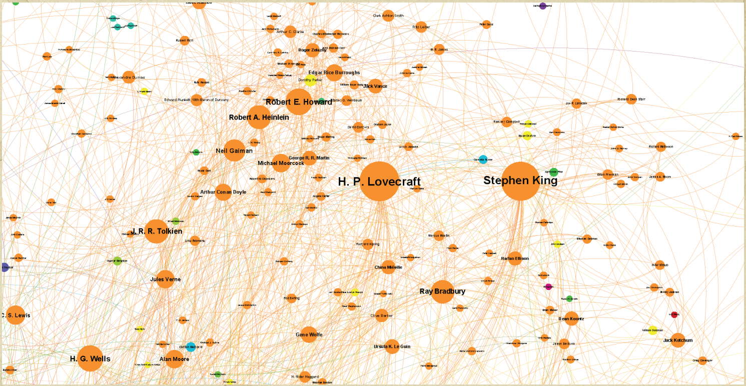
- Obesity [8]
- Smoking cessation [9]
- Happiness [13]
- Loneliness [7]

One of many questions:

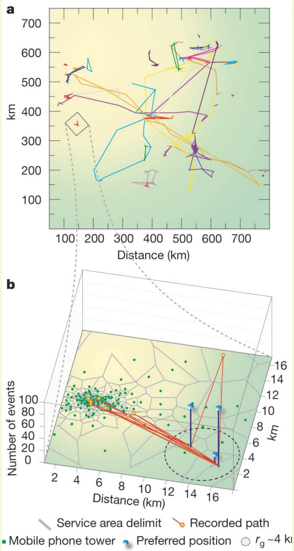
How does the (very) sparse sampling of a real social network affect their findings?

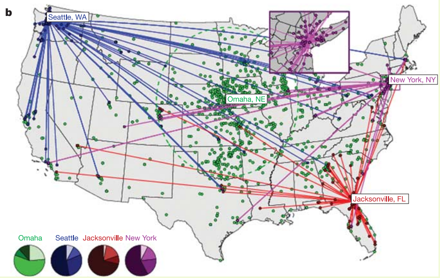


From here (⊞), the linking of people (roughly) according to the Wikipedia:

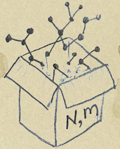


How people move around:





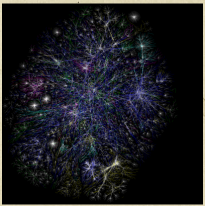
- Study movement and interactions of people.
- Brockmann *et al.* [6] “Where’s George” study.
- Barabasi’s group: tracking movement via cell phones [14].

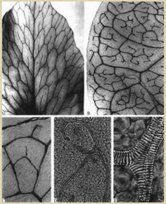


Three broad network classes:

1. Physical networks

- River networks
- Neural networks
- Trees and leaves
- Blood networks
- The Internet
- Road networks
- Power grids





- Distribution (branching) vs. redistribution (cyclical)

Three broad network classes:

2. Interaction networks

- Biochemical networks
- Gene-protein networks
- Food webs: who eats whom
- The World Wide Web (?)
- Airline networks
- The Media
- Paper citations

datamining.typepad.com (⊞)

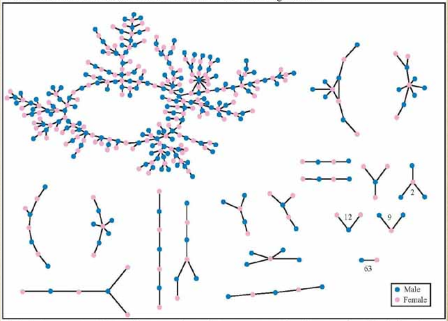
Three broad network classes:

2. Interaction networks: social networks

- Snogging
- Friendships

- Boards and directors
- Organizations
- Facebook
- Twitter

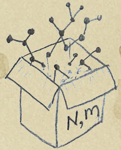
The Structure of Romantic and Sexual Relations at "Jefferson High School"



Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

(Bearman *et al.*, 2004)

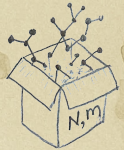
- 'Remotely sensed' by: email activity, instant messaging, phone logs (*cough*).

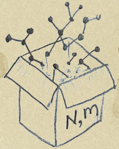


Four broad network classes:

3. Relational networks

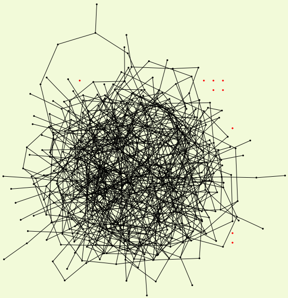
- Consumer purchases (Wal-Mart: $>$ petabyte $= 10^{15}$ bytes)
- Thesauri: Networks of words generated by meanings
- Knowledge/Databases/Ideas
- Metadata—Tagging: flickr (⊞) bit.ly (⊞),





A notable feature of large-scale networks:

- Graphical renderings are often just a big mess.



$\Leftarrow$ Typical hairball

- number of nodes $N = 500$
- number of edges $m = 1000$
- average degree $\langle k \rangle = 4$

- And even when renderings somehow look good: "That is a very graphic analogy which aids understanding wonderfully while being, strictly speaking, wrong in every possible way" said Ponder [Stibbons] —*Making Money*, T. Pratchett.
- We need to extract digestible, meaningful aspects.

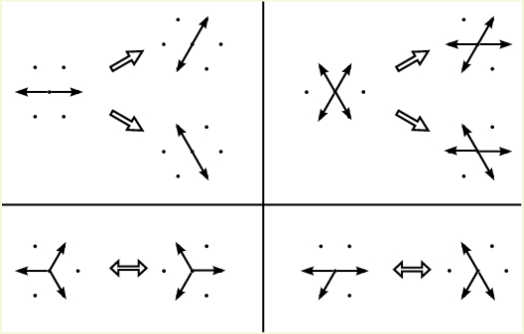
The Theory of Anything:

Fluids mechanics:

- Fluid mechanics = One of the great successes of understanding complex systems.
- Navier-Stokes equations: micro-to-macro system evolution.
- Yesness: Observations + Experiment + Theory
- Works for many very different ‘fluids’:
  - the atmosphere,
  - oceans,
  - blood,
  - galaxies,
  - the earth’s mantle...
  - and ball bearings on lattices...?

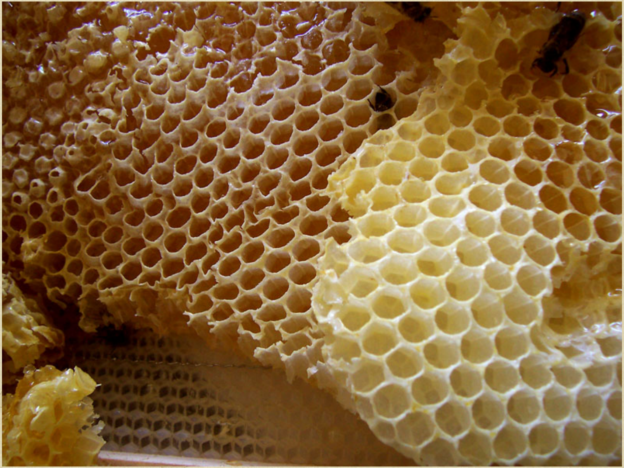
Lattice gas models

Collision rules in 2-d on a hexagonal lattice:



- Lattice matters... Only hexagonal lattice works in 2-d.
- No 'good' lattice in 3-d.
- Upshot: play with 'particles' of a system to obtain new or specific macro behaviours.

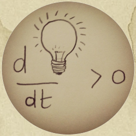
Hexagons—Honeycomb: (⊞)



- Orchestrated? Or an accident of bees working hard?
- See "On Growth and Form" by D'Arcy Wentworth Thompson (⊞).[27, 28]

Hexagons—Giant's Causeway: (⊞)

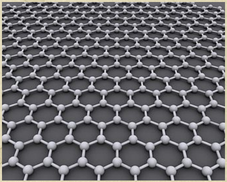
http://newdesktopwallpapers.info



Hexagons—Giant's Causeway: (⊞)

http://www.physics.utoronto.ca/

Hexagons run amok:



- Graphene (⊞): single layer of carbon molecules in a perfect hexagonal lattice (super strong).
- Chicken wire (⊞) ...

Symmetry Breaking

Philip Anderson (⊞)—"More is Different," Science, 1972 [2]

- Argues against idea that the only real scientists are those working on the fundamental laws.
- Symmetry breaking → different laws/rules at different scales...

(2006 study → "most creative physicist in the world" (⊞))

Symmetry Breaking

“Elementary entities of science X obey the laws of science Y”

- X
- solid state or many-body physics
- chemistry
- molecular biology
- cell biology

vdots

- psychology
- social sciences

- Y
- elementary particle physics
- solid state many-body physics
- chemistry
- molecular biology

⋮

- physiology
- psychology

Symmetry Breaking

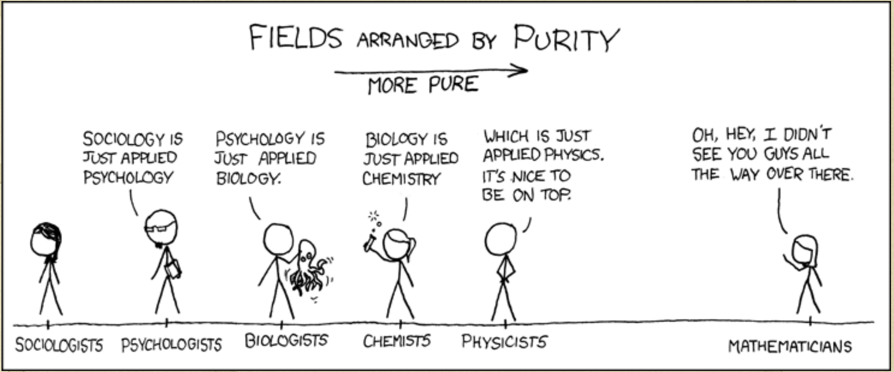
Anderson:

[the more we know about] "fundamental laws, the less relevance they seem to have to the very real problems of the rest of science."

Scale and complexity thwart the constructionist hypothesis.

Accidents of history and path dependence (⊞) matter.

More is different:



http://xkcd.com/435/ (⊞)

A real science of complexity:

A real theory of ~~everything~~ anything:

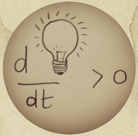
1. Is not just about the ridiculously small stuff...
2. It's about the increase of complexity

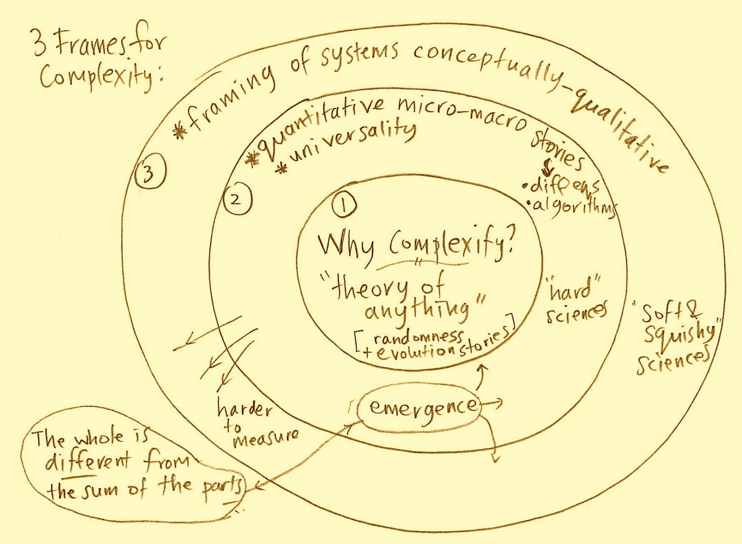
Symmetry breaking/ Accidents of history vs. Universality

- Second law of thermodynamics: we're toast in the long run.
- So how likely is the local complexification of structure we enjoy?
- How likely are the Big Transitions?

Complexification—the Big Transitions:

- Big Bang.
- Big Randomness.
- Big Replicate.
- Big Life.
- Big Evolve.
- Big Word.
- Big Story.
- Big Number.
- Big God.
- Big Make.
- Big Science.
- Big Data.
- Big Information.
- Big Algorithm.
- Big Connection.
- Big Social.
- Big Awareness.




3 Frames for Complexity:
*framing of systems conceptually-qualitative
*quantitative micro-macro stories
*universality
stories
•diffeqs
•algorithms
③
②
①
Why Complexify?
"theory of anything"
[randomness + evolution stories]
"hard" sciences
"Soft & squishy" sciences
harder to measure
emergence
The whole is different from the sum of the parts

Complex Sociotechnical systems:

Sociotechnical phenomena and algorithms:

- Change: How do social movements begin & evolve?
- Performance: How does collective problem solving best work?
- Contagion: How does information move through social networks?
- Elevation: Which rules give the best 'game of society?'

- What can people and computers do together? (Google!)
- Play Project: Use Play + Crunch (or Carbon and Silicon) to solve problems. Which problems?

Play and Crunch—Foldit:

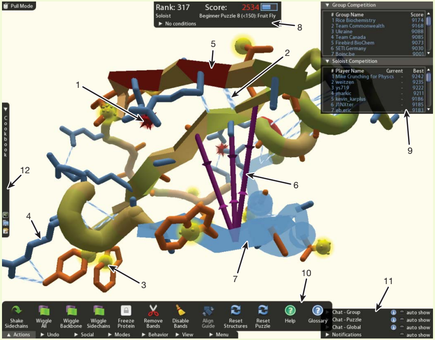
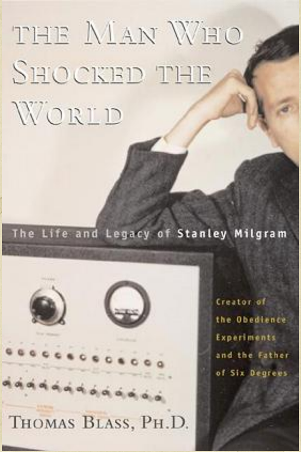


Figure 1 | Foldit screenshot illustrating tools and visualizations. The visualizations include a clash representing atoms that are too close (arrow 1); a hydrogen bond (arrow 2); a hydrophobic side chain with a yellow blob because it is exposed (arrow 3); a hydrophilic side chain (arrow 4); and a segment of the backbone that is red due to high residue energy (arrow 5). The players can make modifications including 'rubber bands' (arrow 6), which add constraints to guide automated tools, and freezing (arrow 7), which prevents degrees of freedom from changing. The user interface includes information about the player's current status, including score (arrow 8); a leader board (arrow 9), which shows the scores of other players and groups; toolbars for accessing tools and options (arrow 10); chat for interacting with other players (arrow 11); and a 'cookbook' for making new automated tools or 'recipes' (arrow 12).

- "Predicting protein structures with a multiplayer online game." Cooper et al., Nature, 2010. [11]
- Also: Chess, zooniverse (⊞), ESP game (⊞), captchas (⊞).

Milgram's social search experiment (1960s)



http://www.stanleymilgram.com

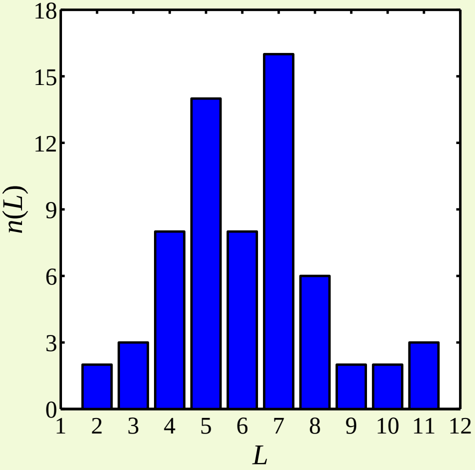
- Target person = Boston stockbroker.
- 296 senders from Boston and Omaha.
- 20% of senders reached target.
- chain length $\simeq 6.5$.

Popular terms:

- The Small World Phenomenon;
- "Six Degrees of Separation."

Milgram's social search experiment (1960s)

Lengths of successful chains:



From Travers and Milgram (1969) in Sociometry:[29]
"An Experimental Study of the Small World Problem."

The Small World Problem:

Two features characterize a social 'Small World':

1. Short paths exist.
2. People are good at finding them.

Social Search—the Columbia experiment

Milgram's small world experiment with email:

- "An Experimental study of Search in Global Social Networks"
  P. S. Dodds, R. Muhamad, and D. J. Watts, *Science*, Vol. 301, pp. 827–829, 2003. [12]

Social search—the Columbia experiment

Experiment details:

- Word of mouth + accidental global media coverage: 60,000+ participants in 166 countries.
- 18 targets in 13 countries including:
  - a professor at an Ivy League university,
  - an archival inspector in Estonia,
  - a technology consultant in India,
  - a policeman in Australia,
  - and a veterinarian in the Norwegian army.
- 24,000+ search chains.

Social search—the Columbia experiment

Participation rates:

- Milgram's experiment: $\approx$ 75% participation rate.
- Email version (different era): $\approx$ 37% participation rate.
- Probability of a chain of length 10 getting through:

$$.37^{10} \simeq 5 \times 10^{-5}$$

- Columbia experiment: Only 384 completed chains (1.6% of all chains).

Upshot:

- Motivation/Incentives/Perception matter.
- Distant influence in networks is hard.

Social search—Algorithmic choices matter:

Successful chains disproportionately used:

- weak ties (Granovetter)
- professional ties (34% vs. 13%)
- ties originating at work/college
- target's work (65% vs. 40%)

. . . and disproportionately avoided

- hubs (8% vs. 1%) (+ no evidence of funnels)
- family/friendship ties (60% vs. 83%)

Geography $\rightarrow$ Work

Demographics are of minimal importance:

Senders of successful messages showed little absolute dependency on

- age,
- gender,
- country of residence,
- income,
- religion,
- relationship to recipient.

Range of completion rates for subpopulations:

- 30% to 40%

Social search—the Columbia experiment

Basic results:

- $\langle L \rangle = 4.05$ for all completed chains
- L_* = Estimated 'true' median chain length (zero attrition)
- Intra-country chains: $L_* = 5$
- Inter-country chains: $L_* = 7$
- All chains: $L_* = 7$
- Milgram: $L_* \simeq 9$

Usefulness:

Harnessing social search:

- Can distributed social search be used for something big/good?
- What about something evil? (Good idea to check.)
- What about socio-inspired algorithms for information search?
- For real social search, we have an incentives problem.
- Which kind of influence mechanisms/algorithms would help propagate search?
- Fun, money, prestige, ... ?
- Must be 'non-gameable.'

Red balloons:

A Grand Challenge:

- 1969: The Internet is born (⊞) (the ARPANET (⊞)—four nodes!).
- Originally funded by DARPA who created a grand Network Challenge (⊞) for the 40th anniversary.
- Saturday December 5, 2009: DARPA puts 10 red weather balloons up during the day.
- Each 8 foot diameter balloon is anchored to the ground somewhere in the United States.
- Challenge: Find the latitude and longitude of each balloon.
- Prize: $40,000.

*DARPA = Defense Advanced Research Projects Agency (⊞).

Where the balloons were:

Finding red balloons:

The winning team and strategy:

- MIT's Media Lab (⊞) won in less than 9 hours. [21]
- Pickard et al. "Time-Critical Social Mobilization," [21] Science Magazine, 2011.
- People were virally recruited online to help out.
- Idea: Want people to both (1) find the balloons and (2) involve more people.
- Recursive incentive structure with exponentially decaying payout:
  - \$2000 for correctly reporting the coordinates of a balloon.
  - \$1000 for recruiting a person who finds a balloon.
  - \$500 for recruiting a person who recruits the balloon finder, etc.

Finding balloons:

Clever scheme:

- Max payout = $4000 per balloon.
- Individuals have clear incentives to both
  1. involve/source more people (spread), and
  2. find balloons (goal action).
- Gameable?
- Limit to how much money a set of bad actors can extract.

Extra notes:

- MIT's brand helped greatly.
- MIT group first heard about the competition a few days before. Ouch.
- A number of other teams did well (⊞).
- Worthwhile looking at these competing strategies.

The social world appears to be small... why?

Theory: how do we understand the small world property?

- Connected random networks have short average path lengths:

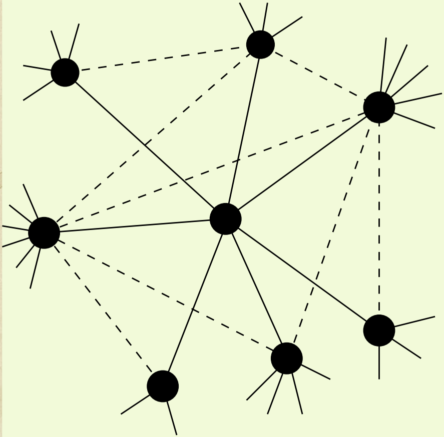
$$\langle d_{AB} \rangle \sim \log(N)$$

N = population size,

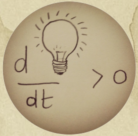
d_{AB} = distance between nodes A and B.

- But: social networks aren't random...

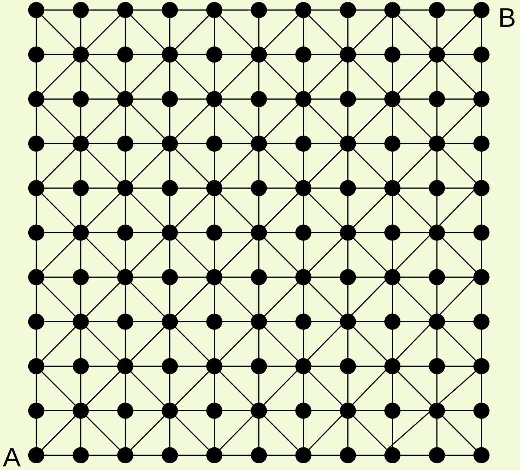
Simple socialness in a network:



Need "clustering" (your friends are likely to know each other):

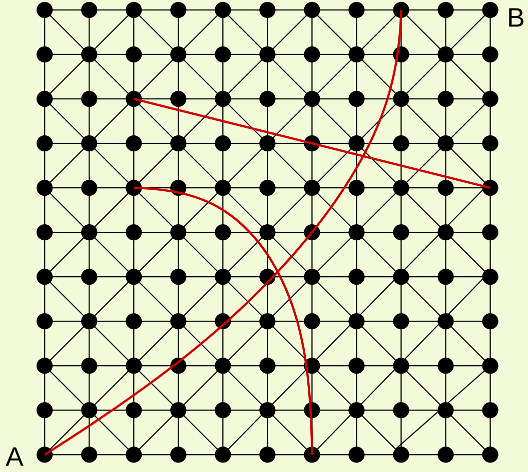


Non-randomness gives clustering:

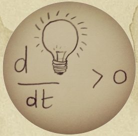


$d_{AB} = 10 \rightarrow$ too many long paths.

Randomness + regularity



Now have $d_{AB} = 3$ $\langle d \rangle$ decreases overall



Small-world networks

Introduced by Watts and Strogatz (Nature, 1998) [31]
"Collective dynamics of 'small-world' networks."

Small-world networks were found everywhere:

- neural network of C. elegans,
- semantic networks of languages,
- actor collaboration graph,
- food webs,
- social networks of comic book characters,...

Very weak requirements:

- local regularity
  + random short cuts

Previous work—finding short paths

But are these short cuts findable?

Nope...

Nodes cannot find each other quickly
with any local search method.

Need a more sophisticated model...

Previous work—finding short paths

- What can a local search method reasonably use?
- How to find things without a map?
- Need some measure of distance between friends and the target.

Some possible knowledge:

- Target's identity
- Friends' popularity
- Friends' identities
- Where message has been

Previous work—finding short paths

Jon Kleinberg (Nature, 2000) [17]
"Navigation in a small world."

Allowed to vary:

1. Local search algorithm
2. Network structure.

Theoretical optimal search:

- "Greedy" algorithm.
- Number of connections grow logarithmically (slowly) in space: $\alpha = d$.
- Social golf.

Previous work—finding short paths

- If networks have hubs can also search well: Adamic et al. (2001) [1]

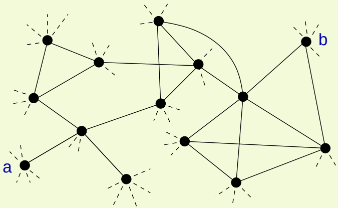
$$P(k_i) \propto k_i^{-\gamma}$$

where k = degree of node i (number of friends).

- Basic idea: get to hubs first (airline networks).
- But: hubs in social networks are limited.

The problem

If there are no hubs and no underlying lattice, how can search be efficient?



Which friend of a is closest to the target b?

What does 'closest' mean?

What is 'social distance'?

Models

One approach: incorporate identity.

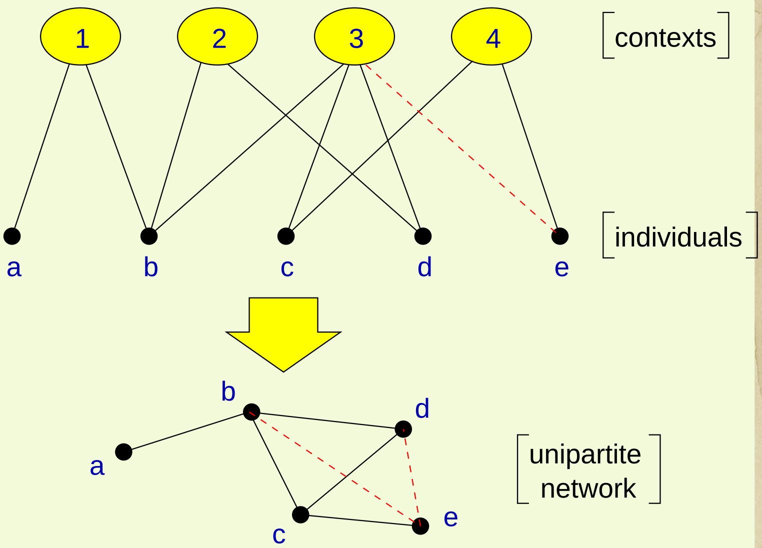
Identity is formed from attributes such as:

- Geographic location
- Type of employment
- Religious beliefs
- Recreational activities.

Groups are formed by people with at least one similar attribute.

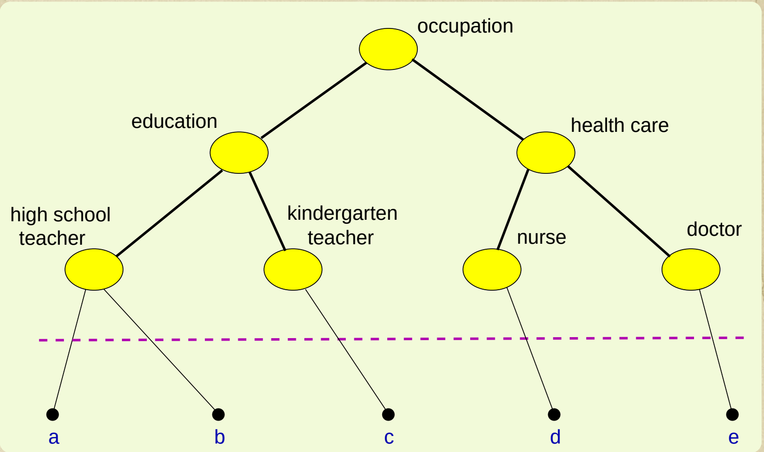
Attributes ⇔ Contexts ⇔ Interactions ⇔ Networks.

Social distance—Bipartite affiliation networks

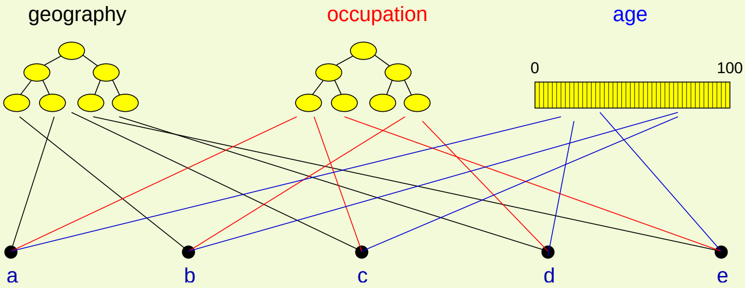


Bipartite affiliation networks: boards and directors, movies and actors.

Social distance—Context distance



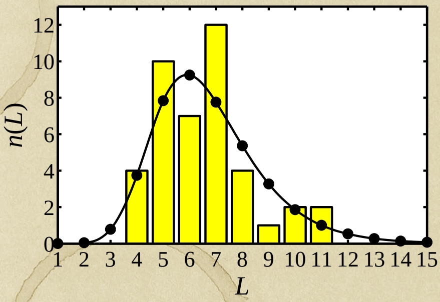
Generalized affiliation networks



- Blau & Schwartz [4], Simmel [25], Breiger [5], Watts *et al.* [30]

The model-results

Milgram's Nebraska-Boston data:



Model parameters:

- $N = 10^8$,
- $z = 300$, $g = 100$,
- $b = 10$,
- $\alpha = 1$, $H = 2$;

- $\langle L_{\text{model}} \rangle \simeq 6.7$
- $L_{\text{data}} \simeq 6.5$

Nutshell for Small-World Networks:

- Bare networks are typically unsearchable.
- Paths are findable if nodes understand how network is formed.
- Importance of identity (interaction contexts).
- Improved social network models.
- Improved peer-to-peer networks.
- Construction of searchable information databases through tagging (experts versus hoi polloi).

Size distributions

The sizes of many systems' elements appear to obey an inverse power-law size distribution:

$$P(\text{size} = x) \sim c\, x^{-\gamma}$$

where $x_{\min} < x < x_{\max}$ and $\gamma > 1$.

- x can be continuous or discrete.
- Typically, $2 < \gamma < 3$.
- No dominant internal scale between $x_{\min}$ and $x_{\max}$.
- If $\gamma < 3$, variance and higher moments are 'infinite'
- If $\gamma < 2$, mean and higher moments are 'infinite'
- Negative linear relationship in log-log space:

$$\log_{10} P(x) = \log c - \gamma \log_{10} x$$

Examples:

- Earthquake magnitude (Gutenberg Richter law): $P(M) \propto M^{-3}$
- Number of war deaths: $P(d) \propto d^{-1.8}$ [24]
- Sizes of forest fires
- Sizes of cities: $P(n) \propto n^{-2.1}$
- Number of links to and from websites
- Number of citations to papers: $P(k) \propto k^{-3}$.
- Individual wealth (maybe): $P(W) \propto W^{-2}$.
- Distributions of tree trunk diameters: $P(d) \propto d^{-2}$.
- Diameter of moon craters: $P(d) \propto d^{-3}$.
- Word frequency: e.g., $P(k) \propto k^{-2.2}$ (variable)

Note: Exponents range in error;
see M.E.J. Newman arxiv.org/cond-mat/0412004v3 (⊞)

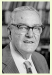
Work of Yore

- 1924: G. Udny Yule [32]:
  # Species per Genus
- 1926: Lotka [18]:
  # Scientific papers per author (Lotka's law)
- 1953: Mandelbrot [19]:
  Optimality argument for Zipf's law for word frequency; focus on language.
- 1955: Herbert Simon [26, 33]:
  Zipf's law, city size, income, publications, and species per genus.
- 1965/1976: Derek de Solla Price [22, 23]:
  Network of Scientific Citations.
- 1999: Barabasi and Albert [3]:
  The World Wide Web, networks-at-large.

Not everyone is happy...

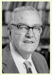
Mandelbrot vs. Simon:

- Mandelbrot (1953): “An Informational Theory of the Statistical Structure of Languages” [19]



- Simon (1955): “On a class of skew distribution functions” [26]

- Mandelbrot (1959): “A note on a class of skew distribution function: analysis and critique of a paper by H. A. Simon”

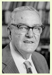


- Simon (1960): “Some further notes on a class of skew distribution functions”

Not everyone is happy... (cont.)

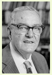
Mandelbrot vs. Simon:

- Mandelbrot (1961): “Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon”



- Simon (1961): “Reply to ‘final note’ by Benoit Mandelbrot”

- Mandelbrot (1961): “Post scriptum to ‘final note’”

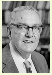


- Simon (1961): “Reply to Dr. Mandelbrot’s post scriptum”

Not everyone is happy... (cont.)

Mandelbrot vs. Simon:

- “We shall restate in detail our 1959 objections to Simon’s 1955 model for the Pareto-Yule-Zipf distribution. Our objections are valid quite irrespectively of the sign of p-1, so that most of Simon’s (1960) reply was irrelevant.”



- “Dr. Mandelbrot has proposed a new set of objections to my 1955 models of the Yule distribution. Like his earlier objections, these are invalid.”

Essential Extract of a Growth Model

Random Competitive Replication (RCR):

1. Start with 1 element of a particular flavor at $t = 1$
2. At time $t = 2, 3, 4, \ldots$, add a new element in one of two ways:
   - With probability ρ, create a new element with a new flavor
     - Mutation/Innovation
   - With probability $1 - \rho$, randomly choose from all existing elements, and make a copy.
     - Replication/Imitation
   - Elements of the same flavor form a group

Random Competitive Replication

Observations:

- Competition for replication between elements is random
- Competition for growth between groups is not random
- Selection on groups is biased by size
- Rich-gets-richer story
- Random selection is easy
- No great knowledge of system needed

Random Competitive Replication

- After some thrashing around, one finds:

$$P_k \propto k^{-\gamma}$$

where

$$\boxed{\gamma = 1 + \frac{1}{(1-\rho)}}$$

- See γ is governed by rate of new flavor creation, ρ.

Evolution of catch phrases

- Yule's paper (1924) [32]:
  "A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S."
- Simon's paper (1955) [26]:
  "On a class of skew distribution functions" (snore)

From Simon's introduction:

It is the purpose of this paper to analyse a class of distribution functions that appear in a wide range of empirical data—particularly data describing sociological, biological and economoic phenomena.

Its appearance is so frequent, and the phenomena so diverse, that one is led to conjecture that if these phenomena have any property in common it can only be a similarity in the structure of the underlying probability mechanisms.

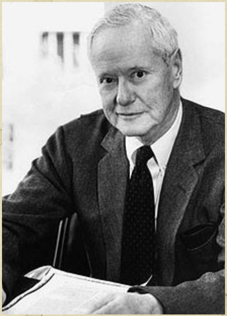
Evolution of catch phrases

- de Solla Price (1965):[22] Cumulative Advantage (better)
  "Networks of scientific papers"

- Robert K. Merton: (⊞) the Matthew Effect (⊞)
- Studied careers of scientists and found credit flowed disproportionately to the already famous

  From the Gospel of Matthew:
  "For to every one that hath shall be given...
  (Wait! There's more....)
  but from him that hath not, that also which he seemeth to have shall be taken away.
  And cast the worthless servant into the outer darkness; there men will weep and gnash their teeth."

Evolution of catch phrases



Merton was a catchphrase machine:

1. Self-fulfilling prophecy
2. Role model
3. Unintended (or unanticipated) consequences
4. Focused interview → focus group

Bonus achievement:

Robert C. Merton won the Nobel Prize for Economics in 1997.

Evolution of catch phrases—Scale-free networks:

- Barabási and Albert [3]—thinking about the Web
- Independent reinvention of a version of Simon and Price's theory for networks
- Another term: "Preferential Attachment"
- Basic idea: a new node arrives every discrete time step and connects to an existing node i with probability $\propto k_i$.
- Connection:
  Groups of a single flavor $\sim$ edges of a node
- Small hitch: selection mechanism is now non-random
- Solution: Connect to a random node (easy)
- + Randomly connect to the node's friends (also easy)
- Scale-free networks = food on the table for physicists

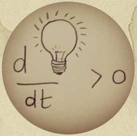
References I

[1] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman.
Search in power-law networks.
Phys. Rev. E, 64:046135, 2001. pdf (⊞)

[2] P. W. Anderson.
More is different.
Science, 177(4047):393–396, 1972. pdf (⊞)

[3] A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
Science, 286:509–511, 1999. pdf (⊞)

[4] P. M. Blau and J. E. Schwartz.
Crosscutting Social Circles.
Academic Press, Orlando, FL, 1984.



References II

[5] R. L. Breiger.
The duality of persons and groups.
Social Forces, 53(2):181–190, 1974. pdf (⊞)

[6] D. Brockmann, L. Hufnagel, and T. Geisel.
The scaling laws of human travel.
Nature, pages 462–465, 2006. pdf (⊞)

[7] J. T. Cacioppo, J. H. Fowler, and N. A. Christakis.
Alone in the crowd: The structure and spread of loneliness in a large social network.
Journal of Personality and Social Psychology, 97:977–991, 2009. pdf (⊞)

References III

[8] N. A. Christakis and J. H. Fowler.
The spread of obesity in a large social network over 32 years.
New England Journal of Medicine, 357:370–379, 2007. pdf (⊞)

[9] N. A. Christakis and J. H. Fowler.
The collective dynamics of smoking in a large social network.
New England Journal of Medicine, 358:2249–2258, 2008. pdf (⊞)

[10] V. Colizza, A. Barrat, M. Barthelmey, A.-J. Valleron, and A. Vespignani.
Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions.
PLoS Medicine, 4:e13, 2011. pdf (⊞)

References IV

[11] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. players.
Predicting protein structures with a multiplayer online game.
Nature, 466:756–760, 466. pdf (⊞)

[12] P. S. Dodds, R. Muhamad, and D. J. Watts.
An experimental study of search in global social networks.
Science, 301:827–829, 2003. pdf (⊞)

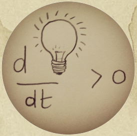
[13] J. H. Fowler and N. A. Christakis.
Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study.
BMJ, 337:article #2338, 2008. pdf (⊞)

References V

[14] M. C. González, C. A. Hidalgo, and A.-L. Barabási.
Understanding individual human mobility patterns.
Nature, 453:779–782, 2008. pdf (⊞)

[15] R. Guimerà, B. Uzzi, J. Spiro, and L. A. N. Amaral.
Team assembly mechanisms determine collaboration network structure and team performance.
Science, 308:697–702, 2005. pdf (⊞)

[16] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausman.
The product space conditions the development of nations.
Science, 317:482–487, 2007. pdf (⊞)



References VI

[17] J. Kleinberg.
Navigation in a small world.
Nature, 406:845, 2000. pdf (⊞)

[18] A. J. Lotka.
The frequency distribution of scientific productivity.
Journal of the Washington Academy of Science, 16:317–323, 1926.

[19] B. B. Mandelbrot.
An informational theory of the statistical structure of languages.
In W. Jackson, editor, Communication Theory, pages 486–502. Butterworth, Woburn, MA, 1953. pdf (⊞)

References VII

[20] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. A. Lieberman.
Quantitative analysis of culture using millions of digitized books.
Science Magazine, 331:176–182, 2011. pdf (⊞)

[21] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland.
Time-critical social mobilization.
Science, 334:509–512, 2011. pdf (⊞)

[22] D. J. d. S. Price.
Networks of scientific papers.
Science, 149:510–515, 1965. pdf (⊞)

References VIII

[23] D. J. d. S. Price.
A general theory of bibliometric and other cumulative advantage processes.
J. Amer. Soc. Inform. Sci., 27:292–306, 1976.

[24] L. F. Richardson.
Variation of the frequency of fatal quarrels with magnitude.
J. Amer. Stat. Assoc., 43:523–546, 1949. pdf (⊞)

[25] G. Simmel.
The number of members as determining the sociological form of the group. I.
American Journal of Sociology, 8:1–46, 1902.

[26] H. A. Simon.
On a class of skew distribution functions.
Biometrika, 42:425–440, 1955. pdf (⊞)

References IX

[27] D. W. Thompson.
On Growth and From.
Cambridge University Pres, Great Britain, 2nd edition, 1952.

[28] D. W. Thompson.
On Growth and Form — Abridged Edition.
Cambridge University Press, Great Britain, 1961.

[29] J. Travers and S. Milgram.
An experimental study of the small world problem.
Sociometry, 32:425–443, 1969. pdf (⊞)

[30] D. J. Watts, P. S. Dodds, and M. E. J. Newman.
Identity and search in social networks.
Science, 296:1302–1305, 2002. pdf (⊞)

References X

[31] D. J. Watts and S. J. Strogatz.
Collective dynamics of ‘small-world’ networks.
Nature, 393:440–442, 1998. pdf (⊞)

[32] G. U. Yule.
A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S.
Phil. Trans. B, 213:21–, 1924.

[33] G. K. Zipf.
Human Behaviour and the Principle of Least-Effort.
Addison-Wesley, Cambridge, MA, 1949.