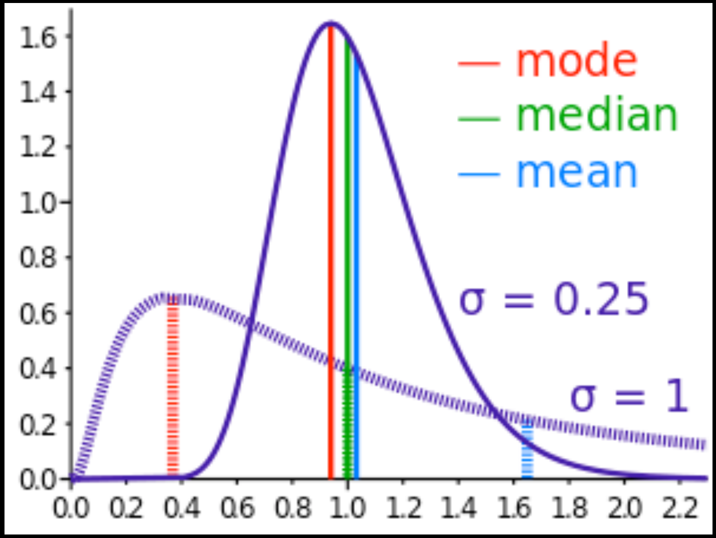
More on statistical distributions

The Log Normal Distribution

- If a variable X is log normally distributed,
  - Y = log(X) is normally distributed
  - Multiplicative product of many (positive) random variables
- In ecology, the Preston curve is log normal
  - Relative frequency of different species
  - The Black-Scholes model in finance assumes that changes in the *logarithm* of exchange rates, price indices, and stock market indices are normal



So: Wikimedia

$$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}$$

So: www.etsy.com

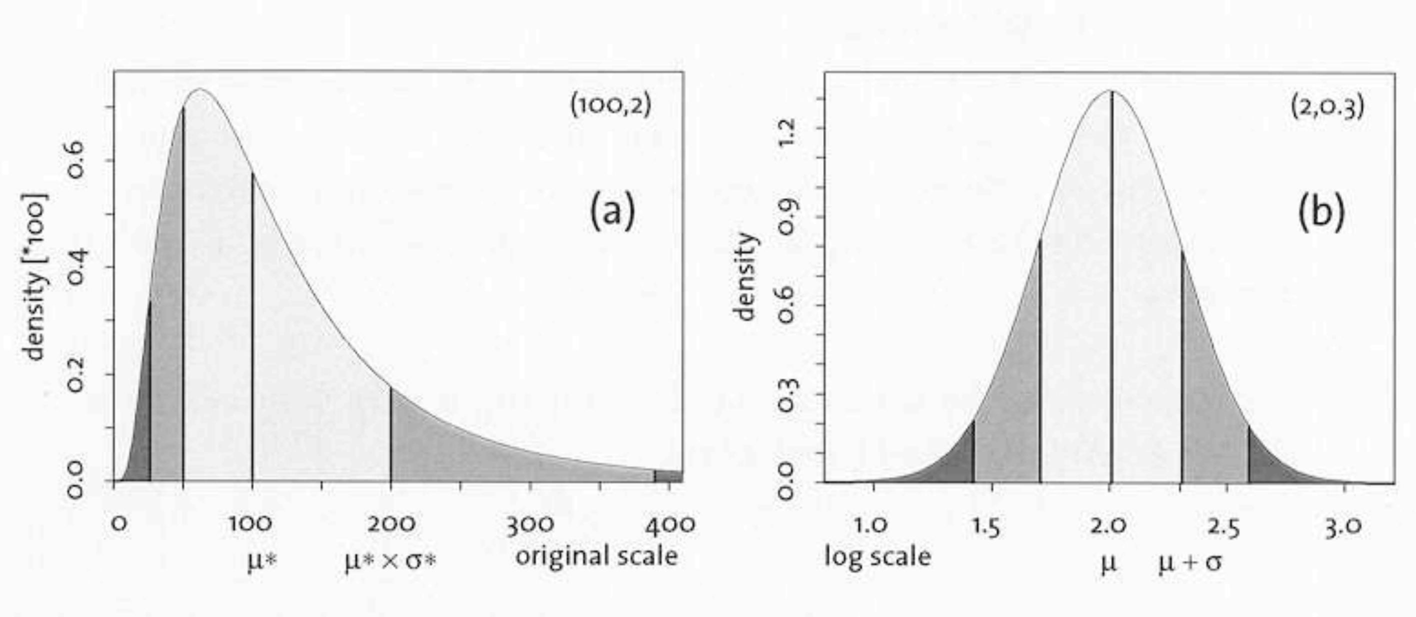
Log Normal vs Normal

$$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}$$

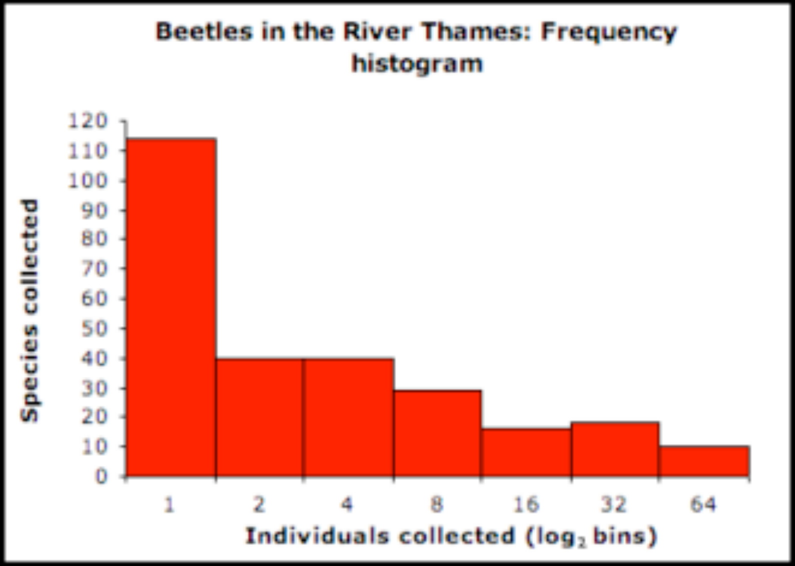
Log normal PDF

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal PDF



Relative Species Abundance
Preston Curves



So: Wikipedia

- **Frequency histogram (Preston Curve)**:
  - ***x*-axis**: logarithm of abundance bins (usually $\log_2$ (because this was historically a simple way to approximate the natural log))
  - ***y*-axis**: number of species at given abundance

Testing for distributions: Review

- Two basic strategies
  - Plot data, fit curves, measure goodness of fit
  - Maximum Likelihood calculations
- Normal
  - Plot on semilogy (look for a quadratic curve)*
  - Mean gives MLE
- Log normal
  - Log data, then treat like a normal distribution
- Exponential
  - Plot on semilogy and look for linear fit
  - MLE is given in the assignment
- Power law
  - Plot on log-log scale and look for lines (++)

* Don't try this at home

How to decide which distribution best fits your data?

1. Estimate the Maximum Likelihood Estimate (MLE) parameters for each distribution
   1. Power law, log normal, exponential
2. Compute the loglikelihood for each distribution using the MLE parameter estimates
3. Compare the loglikelihoods

What is a Likelihood function?

- *Likelihood* (L) is a function of how likely an event is
  - Fundamental concept in modern statistics
- We are given data: X = [x_1, x_2, … x_n]
  - The pdf(x_i) is the probability of observing x_i
  - But, can't compute the pdf without knowing its parameters, e.g., mean, std. deviation, etc.
  - How do we find the parameters?
- Assume some pdf (e.g., normal, power law, etc.), then

$$L = \prod_{i=1}^{n} pdf(x_i) = pdf(x_1) \times pdf(x_2)...pdf(x_n)$$

What’s the problem with this formula?

$$L = \prod_{i=1}^{n} pdf(x_i) = pdf(x_1) \times pdf(x_2) ... pdf(x_n)$$

- Hint: Probabilities are always in [0,1] interval

What’s the problem with that formula?

$$L = \prod_{i=1}^{n} pdf(x_i) = pdf(x_1) \times pdf(x_2)...pdf(x_n)$$

- Probabilities are always in [0,1] interval
- You will have almost 1200 observations in your data set
- Many small numbers multiplied together → Underflow!!!

Log Likelihood Function

$$\mathcal{L} = log[\prod_{i=1}^{n} pdf(x_i)]$$

$$= \sum_{i=1}^{n} log(pdf(x_i))$$

This quantity will always be negative!

Log Likelihood Function

$$\mathcal{L} = log[\prod_{i=1}^{n} pdf(x_i)]$$

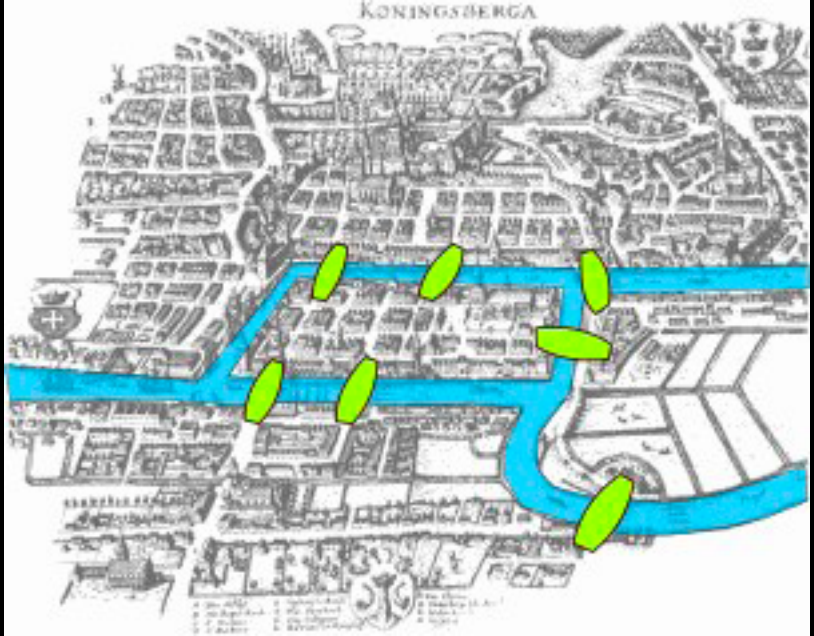
Complex Networks Introduction

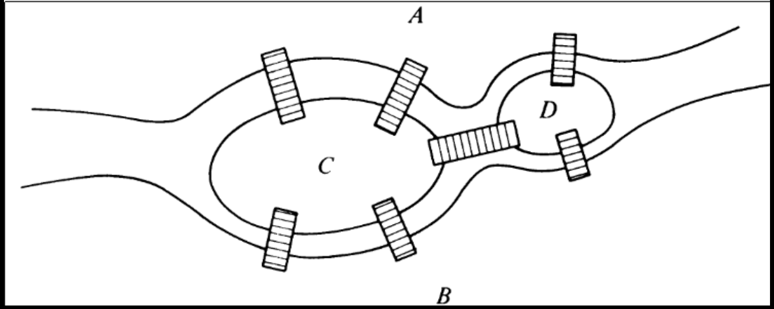
Complex Networks

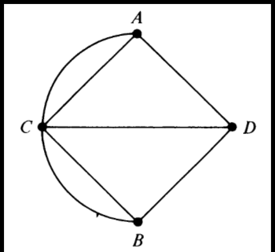
- Related to graph theory
  - Random graphs
- Small worlds

Graph Theory

- Konigsberg bridge problem
- Leonard Euler (1707 – 1783)

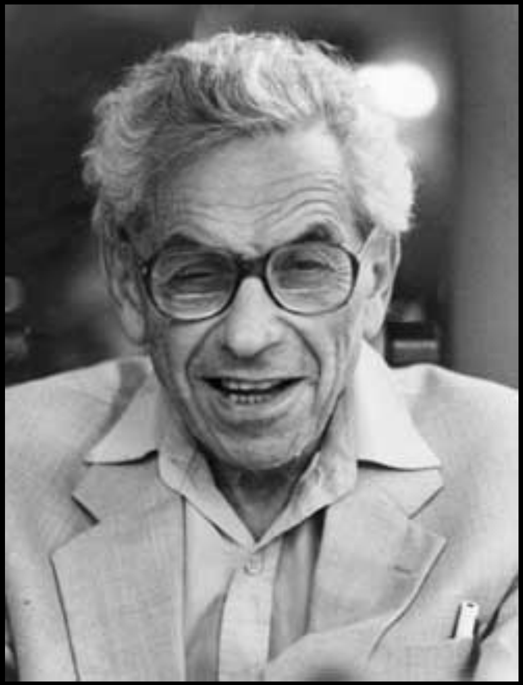






Theory of Random Graphs

- Erdos and Renyi (1960)
- Studied the evolution of random graphs as mean degree is increased
- Properties in random graphs emerge not gradually, but suddenly (phase transitions)
  - e.g., the giant component



Paul Erdos (1913 – 1996)

Small World Experiment

- Travers and Milgram (1969)
- Sent a letter to individuals asking them to forward it to someone that might know a target person
- 296 people from Omaha, Nebraska and Boston were recruited
- Target person lived in Sharon, MA

Instructions

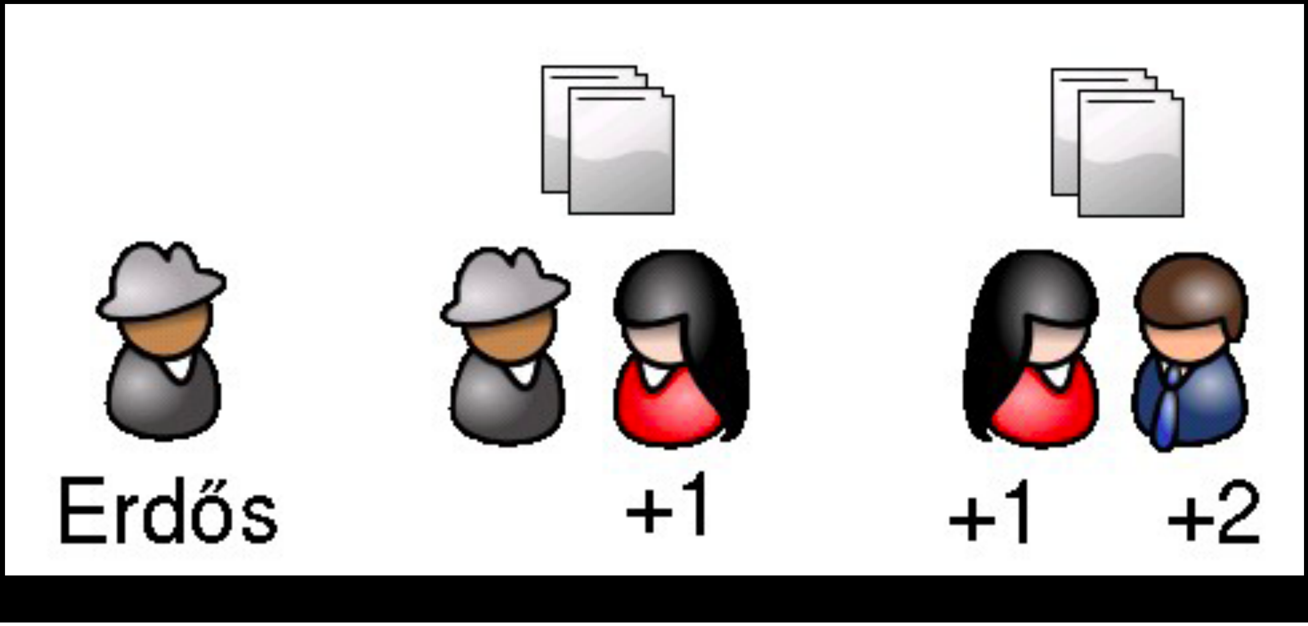
- Add your name ot the roster at the bottom of this sheet
  - Next person who receives the letter will know who it came from
- Detach one postcard, fill it out and return to Harvard University
  - Keep track of progress of folder as it moves towards the target
- If you know the target person on a personal basis, mail the folder directly to him/her, but only if you hae previously met the target person and know each other on a first name basis
- If you do not know the target person, do not try to contact him/her directly. Instead mail the folder to a personal acquaintance who is more likely than you to know the target person

Milgram Experiment

- 29% of letters reached the target
- The number of intermediate acquaintances varied from 1 to 11.
  - Median was 5.2
  - 6 degrees of separation
- Criticisms
  - Most letters did not reach the target
  - No guarantee that letters followed the shortest path

The Erdos Numbers

- Co-authorship network of scientific papers
  - Erods published more than 1500 articles with 500 co-authors
  - Erdos has Erdos number 0

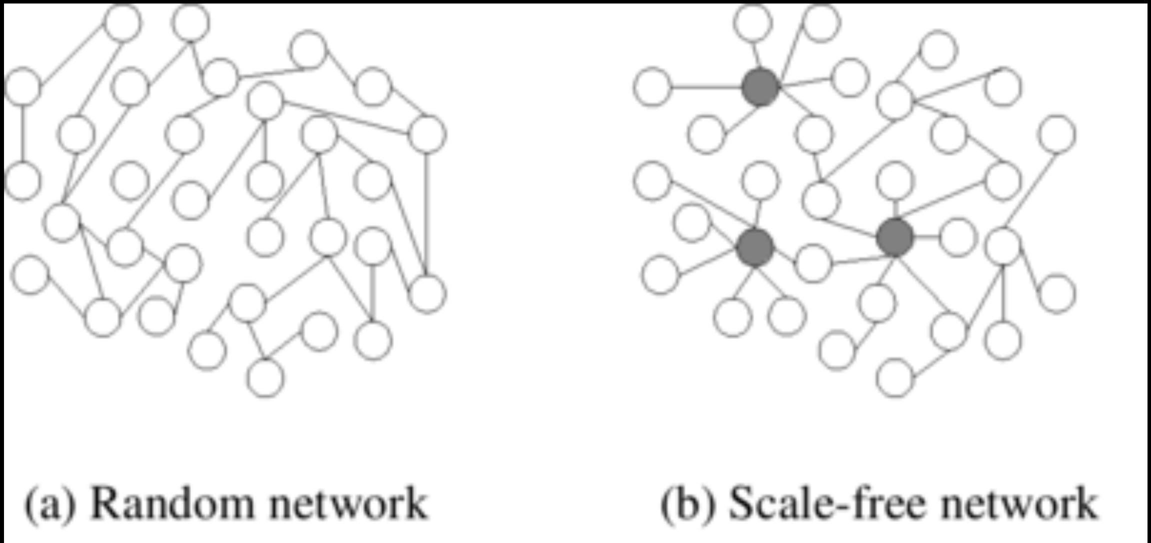


Distribution of Erdos numbers

Erdös number 0 --- 1 person
Erdös number 1 --- 504 people
Erdös number 2 --- 6593 people
Erdös number 3 --- 33605 people
Erdös number 4 --- 83642 people
Erdös number 5 --- 87760 people
Erdös number 6 --- 40014 people
Erdös number 7 --- 11591 people
Erdös number 8 --- 3146 people
Erdös number 9 --- 819 people
Erdös number 10 --- 244 people
Erdös number 11 --- 68 people
Erdös number 12 --- 23 people
Erdös number 13 --- 5 people

Small-world networks

- What is a small-world network?
  - – Small mean geodesic distance (diameter)
  - – Skewed distribution of node degree
  - – Erodos is a hub in the scientific world

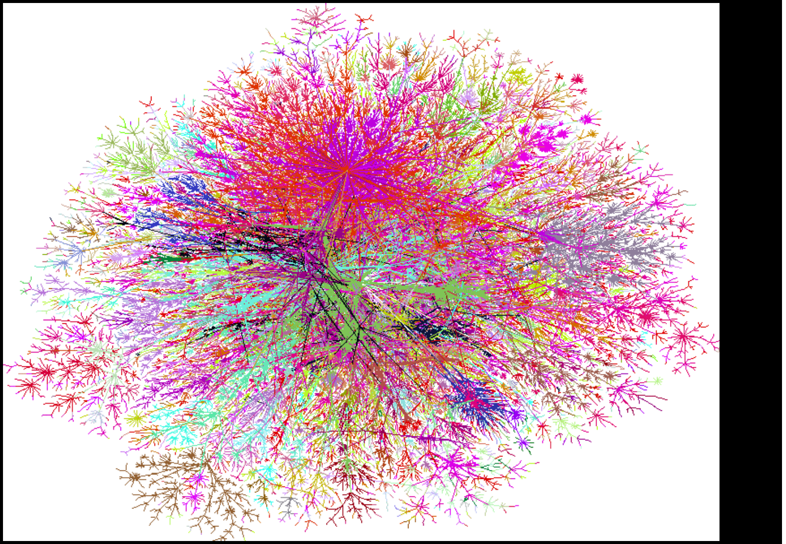


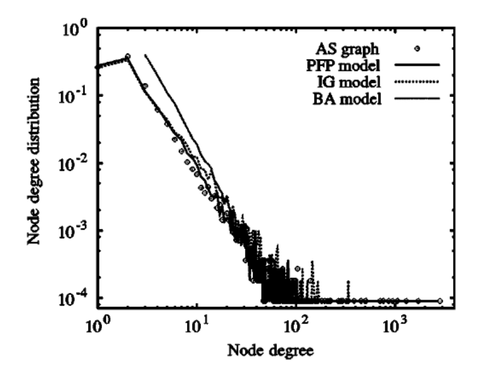
Kevin Bacon Game

- Think of an actor or actress
- If he or she has ever been in a film with Kevin Bacon, then s/he is assigned a Bacon number of 1
- If s/he has never been in a film with Kevin Bacon, but has been in a film with someone else who has, then Kevin Bacon number is 2. etc.
- Highest finite Bacon number (worldwide) is 8

Scale-Free Networks

- Focus on node degree
  - Incoming or outgoing edges
- Degree distribution
  - For each node in network
  - Count the number of incoming (outgoing) edges
  - Make a histogram based on degree
- For many networks, it is power-law distributed, or something similar
  - A few high-connectivity nodes





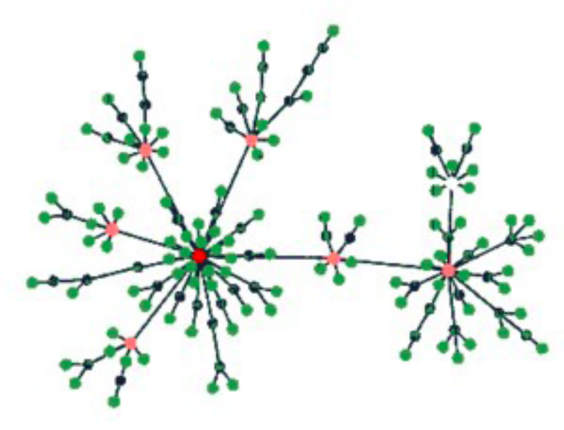
Complex Networks Examples

- Social networks
- Airline connections
- Scientific collaborations
- Email connections
- Metabolic networks
- Actors
- Semantic networks
- Neuronal networks
- Gene regulatory networks
- Terrorist networks
- Software call networks
- Food webs

What mechanisms produce power laws?

Generating Power Laws
Preferential Attachment

- Barabasi-Albert (BA) model (1999)
  - Dates back to H. Smon (1953)
- Grow a network with a rule:
  - Begin with 2 nodes, 1 link
  - New link is added to an existing node with probability proportional to its degree.
  - “Rich get richer”
- Properties:
  - Robust against failure of random nodes
  - Vulnerable to non-random attacks. The network quickly disintegrates when nodes are removed according to their degree.
  - Short average path length



$$L \approx \log N / \log\log k$$

Error- and Attack-Tolerant Networks

- Single-structure networks are homgeneous
  - Each node has approximately the same number of links
  - Diameter increases monotonically with random removal of nodes
  - Diameter increases monotonically under preferential attack
- Scale-free networks are inhomogeneous
  - Highly connected nodes occur with statistical signifcance
  - Generated with preferential attachment models
  - Diameter remains unchanged with random removal of nodes (as many as 50% of nodes can be removed with no effect)
  - Diameter increases dramatically under targeted attack
- Claim:
  - Many communication networks are scale-free or close approximations
- Conclude: Error tolerance in communication networks comes at expense of survivability

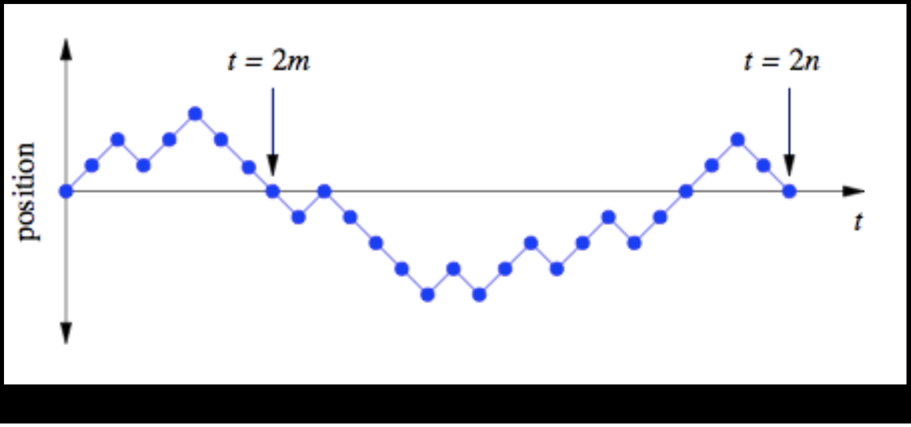
Generating Power Law Distributions cont.

- Preferential attachment (previous slides)
- Combinations of exponentials: $p(y) \approx e^{ay} \qquad x \approx e^{by}$
  - Used to explain power-law dists. of word frequencies
    - Monkey with a typewriter
    - Two distributions: Type letters randomly, P(hit space bar)
  - Example: Dish of reproducing bacteria (exponential population growth) combined with random stopping time (say, to stop the experiment)
  - Then, distribution of X is:

$$p(x) = p(y)\frac{dy}{dx} \approx \frac{e^{ay}}{be^{by}} = \frac{x^{-1+a/b}}{b}$$

Generating Power Law Distributions cont.

- Preferential attachment (previous slides)
- Combinations of exponentials
- Random walks:
  - Many properties of random walks are power-law distributed
  - A walker takes a single step randomly to the right or left along a line, each unit of time. After t steps, what is the probability of returning to the initial position? The distribution of first-return times is power-law distributed:



Newman, 2006

Random Walks cont.

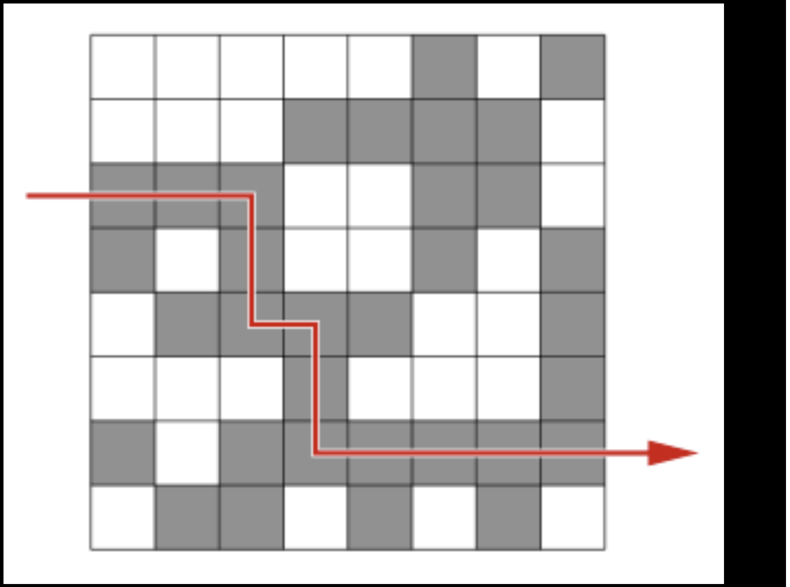
- Random walk models of lifetime of biological taxa (groups of species)
  - Assume that taxon gains and loses species at random over evolutionary time
  - First return time (when taxon goes extinct) is lifespan of the taxon
  - “Gambler’s Ruin”
    - Could explain why lifetimes of genera in the fossil record follow a power law

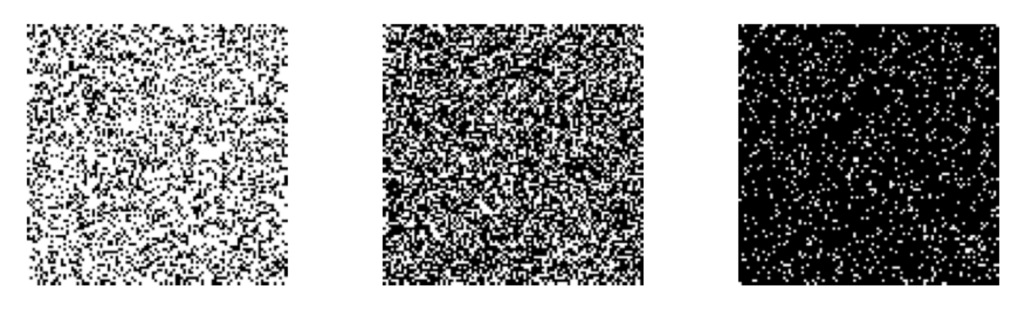
Generating Power Law Distributions cont.

- Preferential attachment (previous slides)
- Combinations of exponentials
- Random walks
- Phase transitions (next slide)

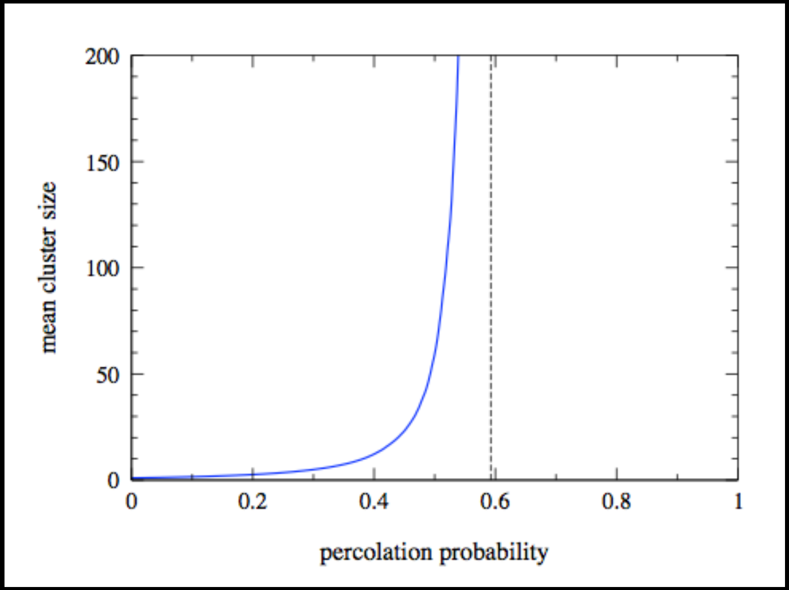
Phase Transitions, Criticality, and Power Laws
(taken from Newman, 2006)

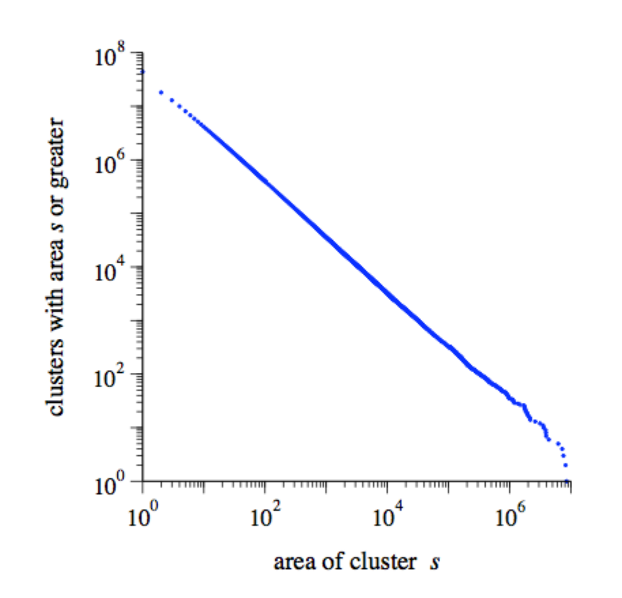
- Example: The percolation threshold
  - Is there a continuous path through colored squres in the lattice?
  - Depends on p = P(a square is black)
  - Cluster size = N (connected colored squares)





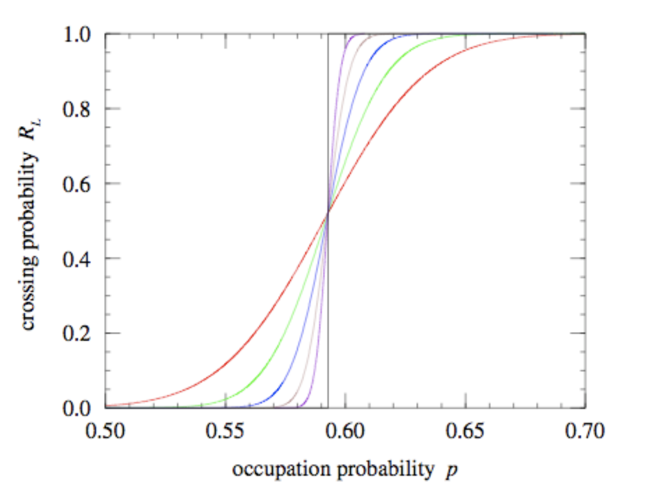
Cluster Sizes and Percolation

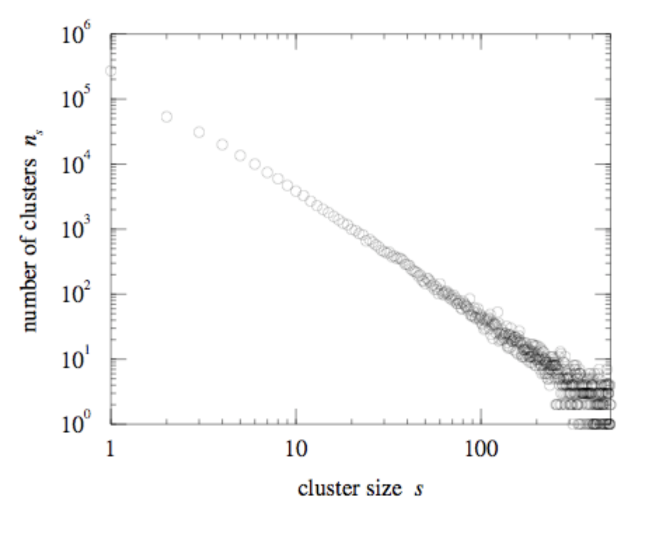




Phase Transitions and Critical Points

- Critical point
  - As size of grid increases, transition becomes increasingly sharp
  - Phase transition
- What is the distribution of cluster sizes in the grid?
  - Power law at the critical point
  - Hallmark of a phase transition
  - Must be exactly at the critical point to observe this
  - Self-organized criticality (SOC) is when a system drives itself to the critical point

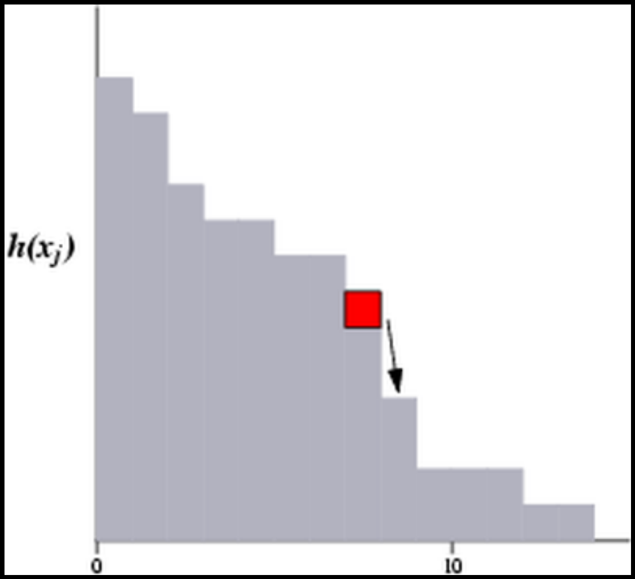




Self-Organized Criticality (SOC)

- Canonical example: Sandpiles (Bak, Tang, and Wiesenfield, 1987)
  - Linear lattice of L sites
  - Grains of sand are distributed to the sites one at a time
  - Height of the pile at location x_j is denoted $h(x_j)$
  - Grains of sand are allowed to accumulate as long as the height difference: $h(x_j) – h(x_j+1) <= 2$
  - Such a situation is unstable and is resolved by "tumbling" grains from x_j to x_{j+1} recursively, until all sites are stable
- Possible effects of adding a single grain of sand:
  - Nothing: $h(x_j) = j(x_j + 1)$
  - One grain tumbles, 2 grains tumble, … n grains tumble ➔
  - An avalanche

Idealized Sandpile (from Jensen, 1998)



- Distribution of avalanche sizes is a power law
  - The sandpile responds to perturbation with events of any size (scale free)
  - Phenomena at many scales contribute to the overall state
- After an avalanche, the sandpile will eventually return to a critical state as more grains of sand are added
  - Self-organized criticality
  - Can generalize to d dimensions

Review: Forest Fire Model

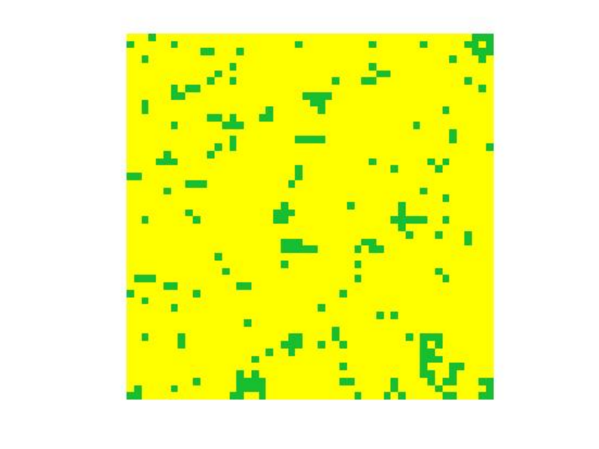
- Imagine a 2-dimensional lattice, where each site can be in one of 3 states:
  - E (empty)
  - G (containing a green tree)
  - B (containing a burning tree)
- Model dynamics:
  - A site occupied by a burning tree becomes an empty site in the succeeding time step.
  - A green tree becomes a burning tree if one or more of its nearest neighbors contains a burning tree.
  - An empty site becomes occupied by a green tree with probability *p* (the growth rate) in each time step.
  - A green tree that is not a neighbor to burning sites catches fire spontaneously with probability *f* (the lightning rate) in each time step.
- Assume periodic boundary conditions and random initial configuration.

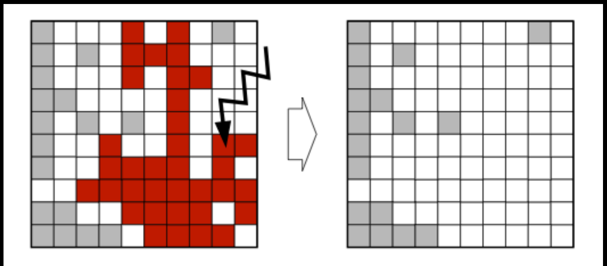
Example: SOC in Forest Fires
Drossel and Schwabl, 1992

- Using the CA model of forest fires discussed earlier,
  - The probability P(S) that a cluster of green trees contains s trees exhibits power law behavior:

$$P(s) \sim \frac{1}{s^{\tau}}$$

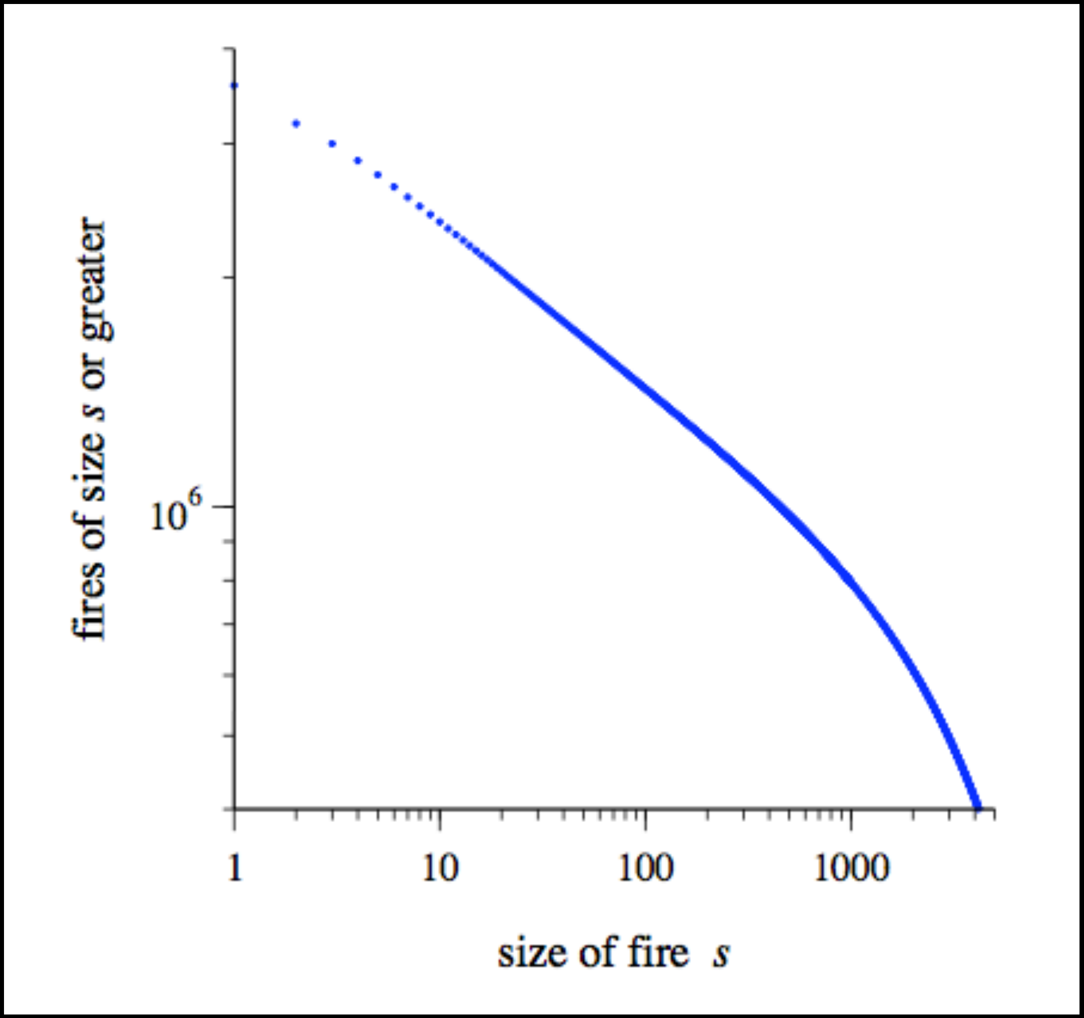
  - The clusters of green trees are fractal objects if the model is extended beyond 2 dimensions
- Model has been extended to study incidence of Measles in isolated environments (Rhodes and Anderson, 1996)





Newman, 2006

CCDF of Fire Size



Newman, 2006

Highly Optimized Tolerance (HOT)

Jean Carlson and John Doyle (1999)

- Claim: Designed sandpiles and percolation models produce power law distributions by a different mechanism than criticality.
- Robustness tradeoffs are an essential feature of complex systems
- Claim:
  - Evolved (natural) or designed (artificial) systems produce rare, structured states which are,
    - Robust to perturbations they were designed to handle
    - Fragile to design flaws and unanticipated perturbations
- These robustness tradeoffs cause complex systems to be “robust yet fragile”:
  - Organisms and ecosystems are robust to large variations in temperature, moisture, nutrients, predation.
  - But, they can be catastrophically sensitive to tiny perturbations, such as genetic mutation, exotic species, or a novel virus

HOT cont.

- SOC requires criticality to get power laws, but in HOT, power law distributions arise under many noncritical conditions
- SOC contains no element of design or planning, but HOT does.
- Example: forest fires
  - Planting forests optimally to reduce fire threat
  - Plant trees sparsely to reduce risk of fire
    - Expect a percolating cluster at critical density, leading to catastrophic fires
    - Engineered solution: Plant trees densely with fire breaks, isolating different regions
  - HOT corresponds to a planting that produces both high yields and protection against catastrophic fires
  - HOT is thus robust to anticipated failures (fires) but susceptible to unanticipated failures (design flaws).

References

- P. Bak, C. Tang, and K. Wiesenfeld. “Self-organized criticality: An explanation of 1/f noise.” Physical Review Letters (1987).
- How Nature Works: The Science of Self-Organized Criticality by P. Bak. Springer-Verlag (1996).
- Self-Organized Criticality by H. Jensen. Cambridge University Press (1998).
- J. Doyle and J. Carlson “Highly Optimized Tolerance: Robustness and Design in Complex Systems.” Physical Review E (1999).
- Small Worlds by Duncan Watts. Princeton University Press (1999).
- R. Albert, H. Jeong, and A. Barabasi “Error and attack tolerance of complex networks” Nature (2000).

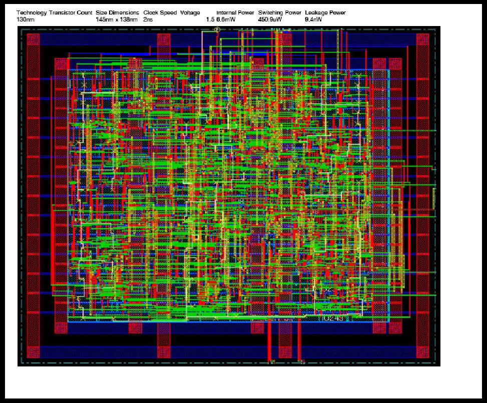
Summary

Basic Mechanisms that Produce Power Laws

- Preferential attachment (previous slides)
- Combinations of exponentials
- Random walks
- Phase transitions (next slide)
  - Critical points
- Optimizations
  - Evolution
  - Engineering

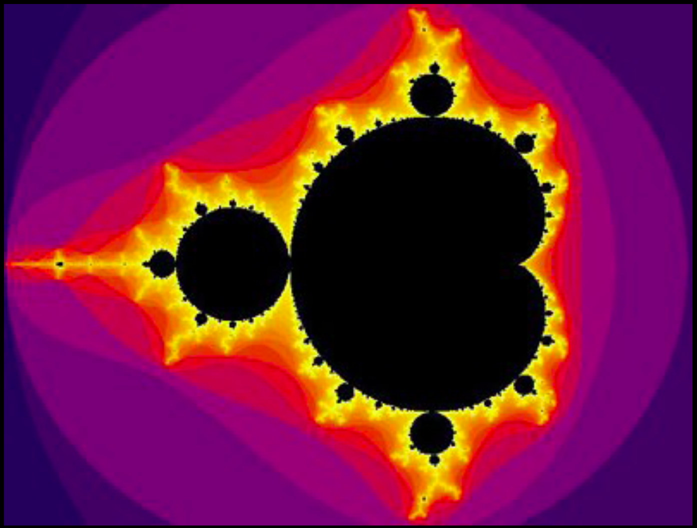
Physical and Geometric constraints determine network architecture and growth

- Network capacity limits performance as systems scale
- Metabolism, response times, power consumption
- Are universal patterns in system behavior predictable from the scaling properties of distribution networks?

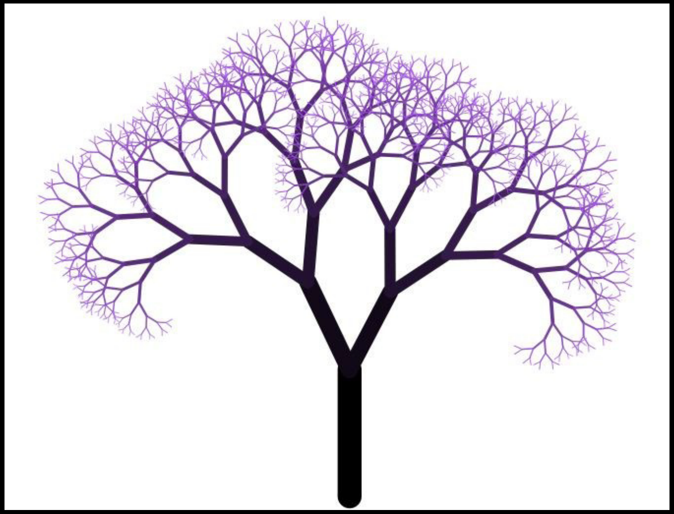


Technology	Transistor Count	Size Dimensions	Clock Speed	Voltage	Internal Power	Switching Power	Leakage Power
130nm		145nm x 138nm	2ns	1.5	6.6mW	450.9uW	9.4nW

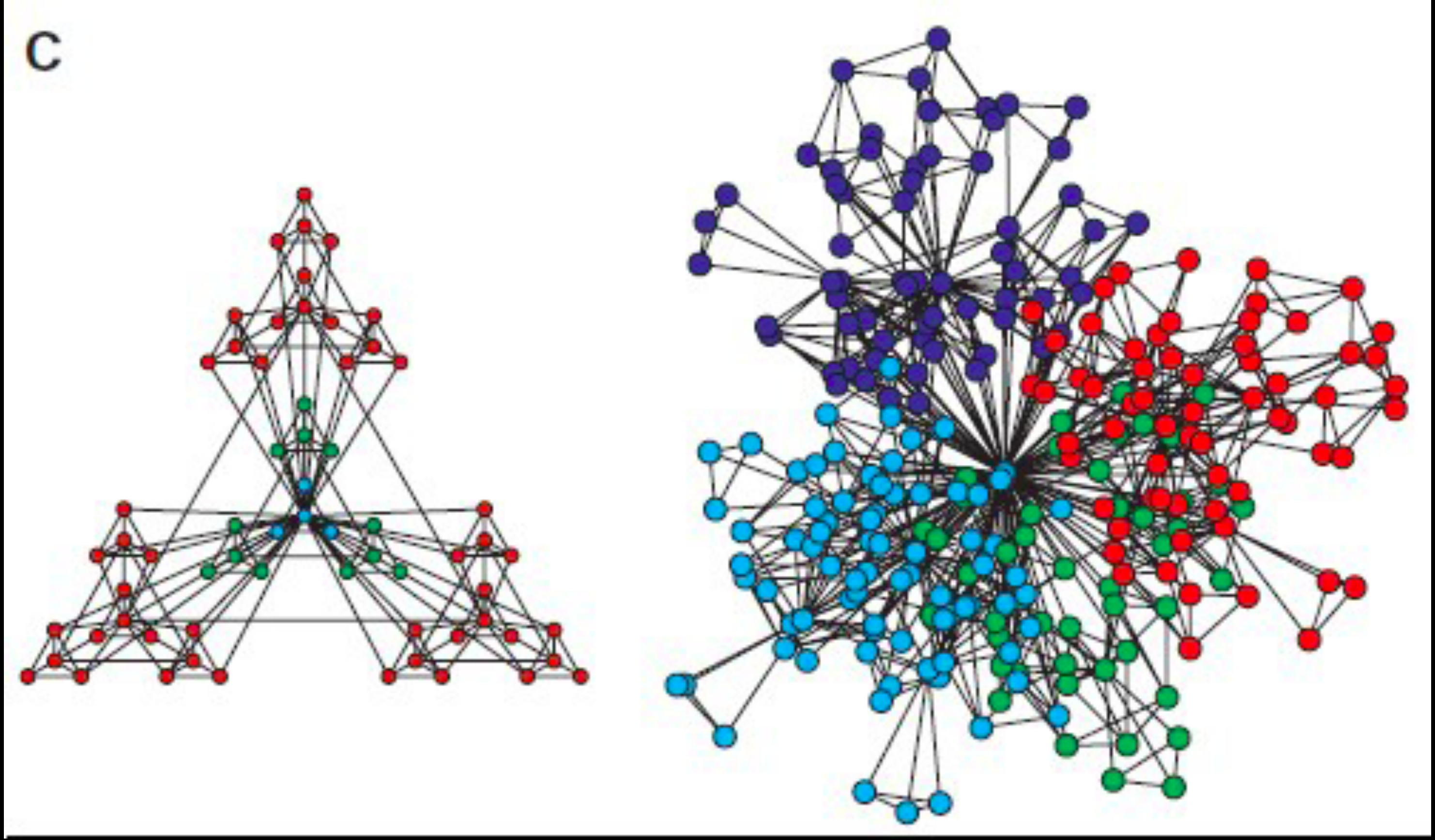
Fractal Networks



More Fractals



Hierarchical Modularity



Hierarchical Modularity

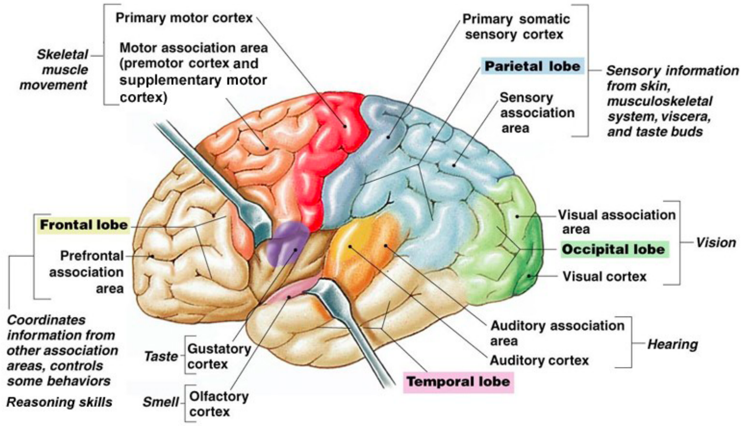
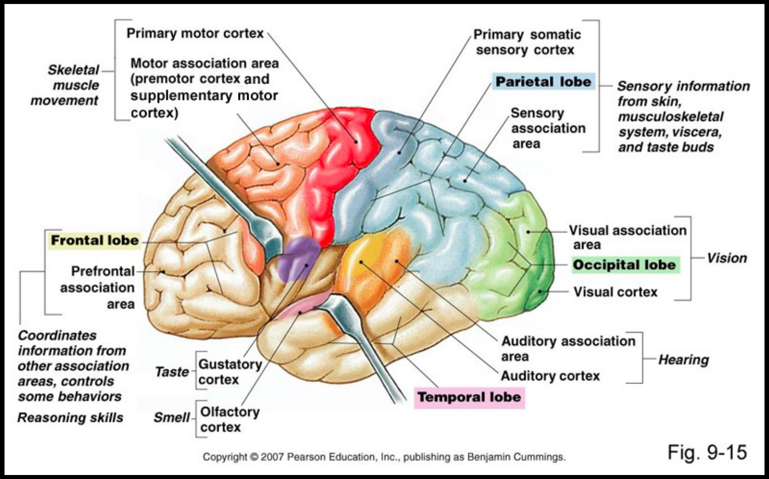

Primary motor cortex
Skeletal muscle movement
Motor association area (premotor cortex and supplementary motor cortex)
Primary somatic sensory cortex
Parietal lobe
Sensory association area
Sensory information from skin, musculoskeletal system, viscera, and taste buds
Frontal lobe
Prefrontal association area
Coordinates information from other association areas, controls some behaviors
Reasoning skills
Visual association area
Occipital lobe
Visual cortex
Vision
Taste
Gustatory cortex
Smell
Olfactory cortex
Auditory association area
Auditory cortex
Hearing
Temporal lobe

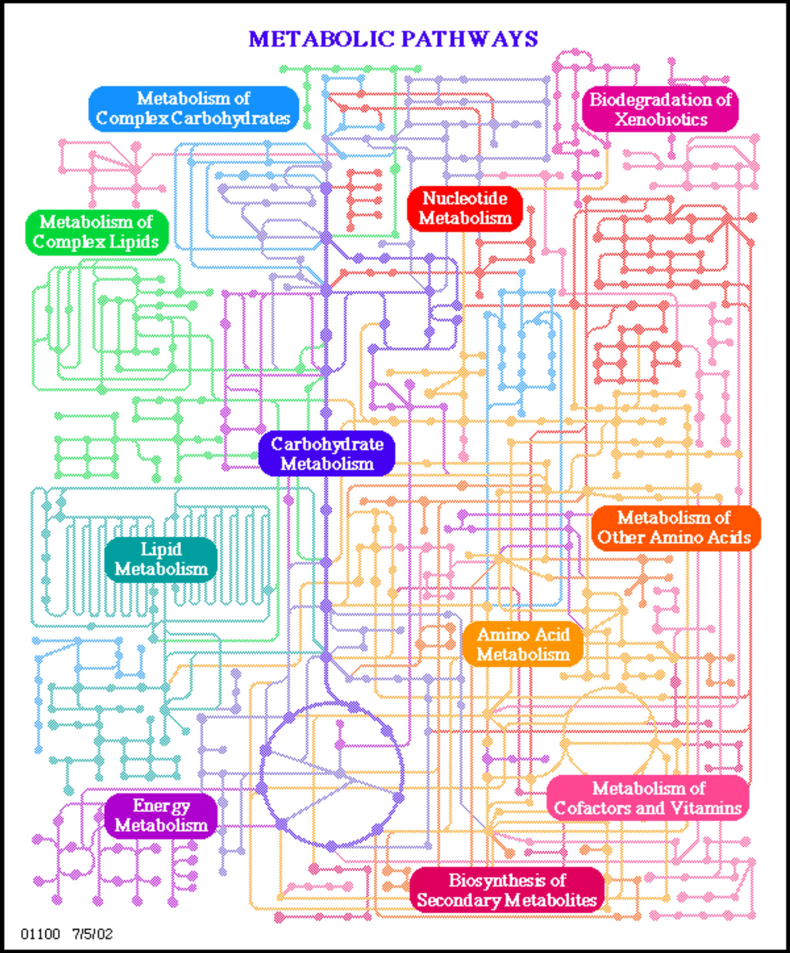
Fig. 9-15

Hierarchical Modularity

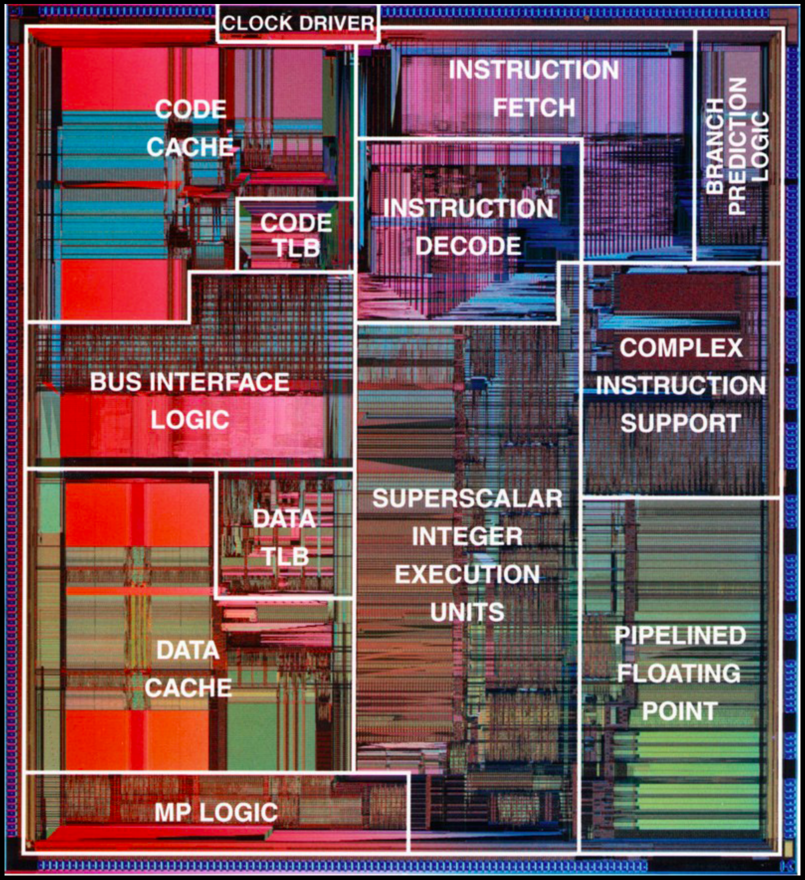

Primary motor cortex
Motor association area (premotor cortex and supplementary motor cortex)
Skeletal muscle movement
Primary somatic sensory cortex
Parietal lobe
Sensory association area
Sensory information from skin, musculoskeletal system, viscera, and taste buds
Frontal lobe
Prefrontal association area
Coordinates information from other association areas, controls some behaviors
Reasoning skills
Visual association area
Occipital lobe
Visual cortex
Vision
Auditory association area
Auditory cortex
Hearing
Taste
Gustatory cortex
Smell
Olfactory cortex
Temporal lobe

Copyright © 2007 Pearson Education, Inc., publishing as Benjamin Cummings.

Fig. 9-15


METABOLIC PATHWAYS
Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Lipid Metabolism
Metabolism of Other Amino Acids
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolites
01100 7/5/02

Computer Designs are also Fractal
Rent's rule (1963)



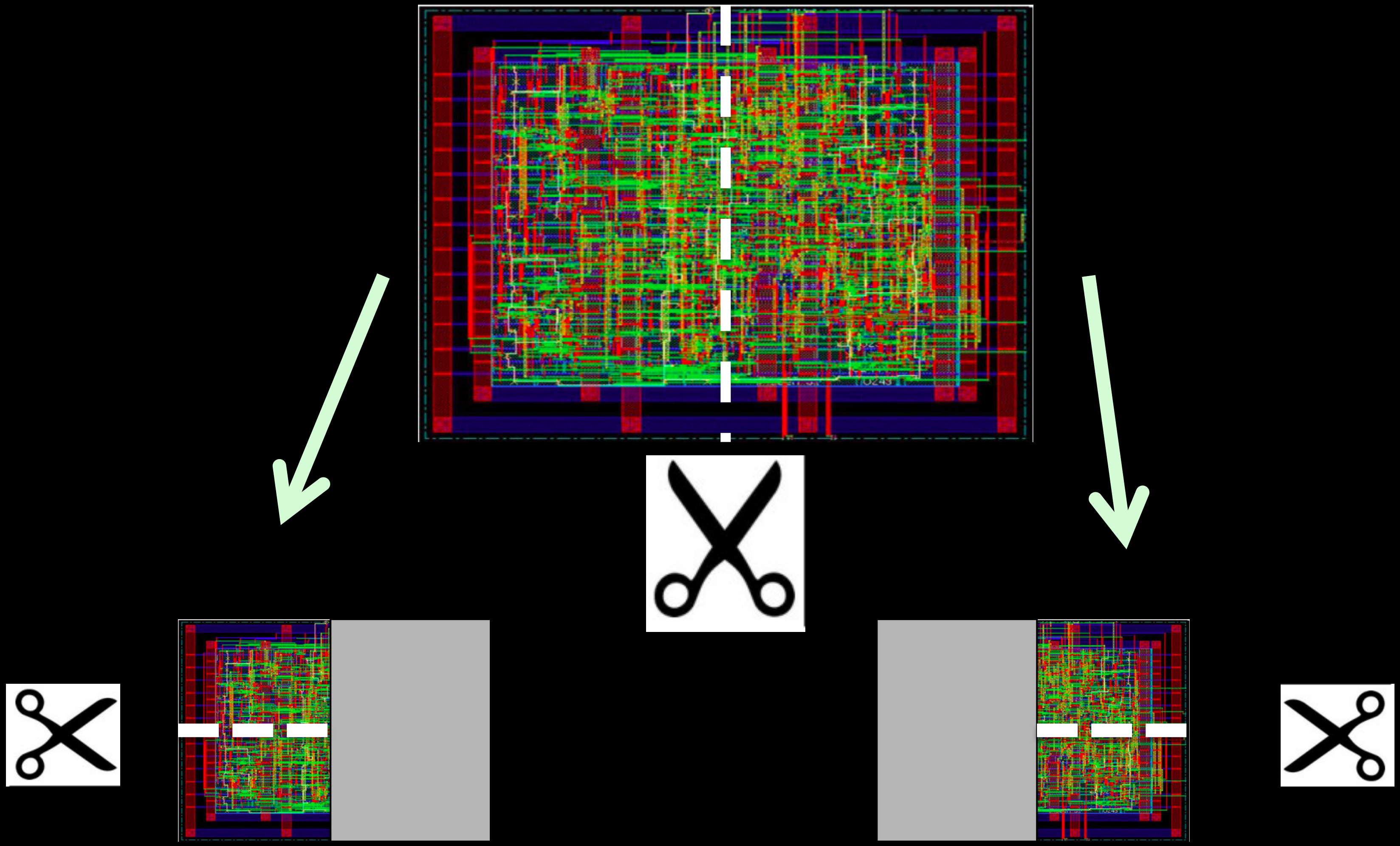
Rent’s Rule

- Relationship between external signal connections in a logic block and number of logic gates in the block
- A scaling relationship for VLSI circuits

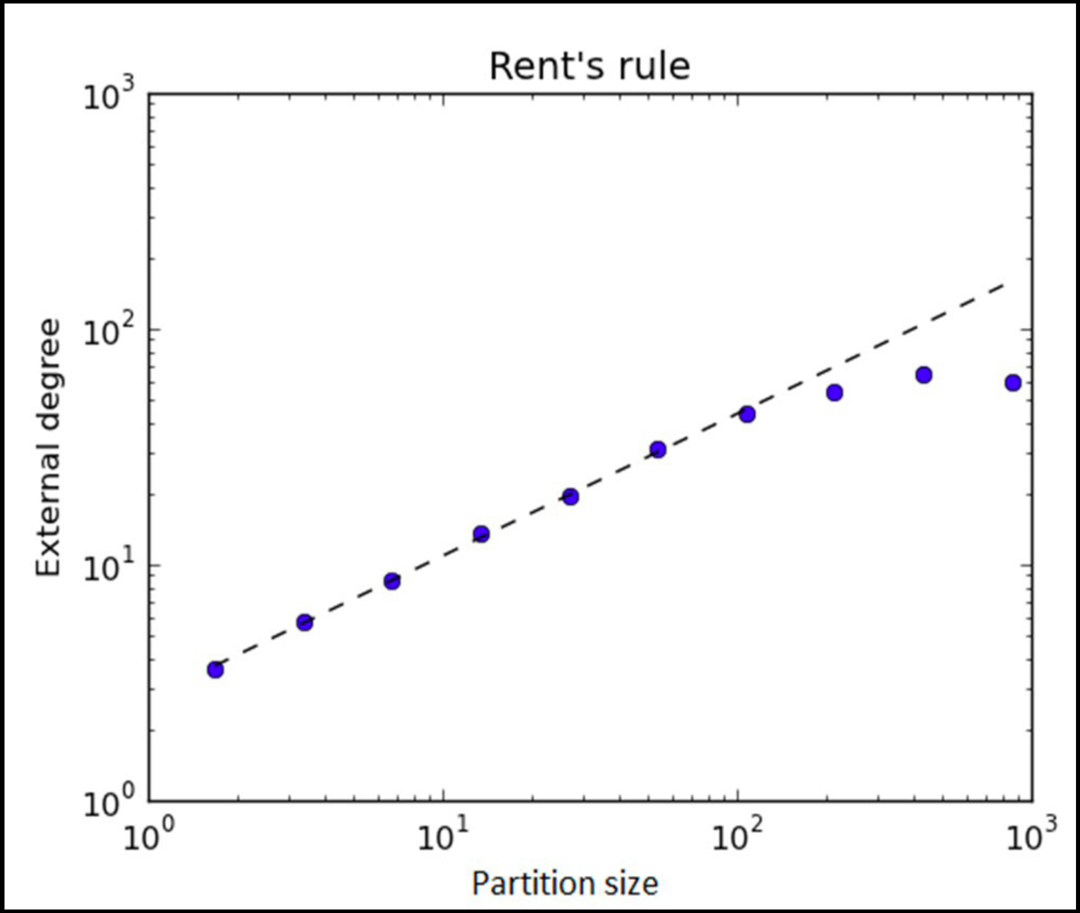
$$C \propto N^p$$

- A power law
  - C = communication
  - N = circuit size
  - P = Rent’s exponent, in [0,1]

Hierarchical Partitioning



Log-log plot of C vs. N



Rent’s rule for benchmark circuit c3540