CSCI E-82
Advanced Machine Learning, Data Mining & Artificial Intelligence
Lecture 12

Peter V. Henstock
Fall 2018

Administrivia: Final Project

- If you submitted proposal to Canvas, it's graded with feedback provided
- Final project presentations: Dec 20
- 23 projects in total
  - 11 team efforts → 8 minutes
  - 12 individual efforts → 5 minutes
  - Total = 148 minutes > 120 minutes
- Option to instead present work in progress on Dec 13

Homework

- HW4 received back from TAs Monday
  - Plan to return ~Saturday
  - Posted top students' solutions
- HW5 is being graded
- HW6 & extra credit deadlines
- Paper reviews (syllabus says 3)
  - Only 2 reviews
  - Plan to grade 1 then have you repeat with a different paper
  - Should not take much time

Remaining weeks

- Today: Network Analysis
- 12/6: Recommender Systems
- 12/13: [Little] Reinforcement Learning
- 12/20: Final presentations

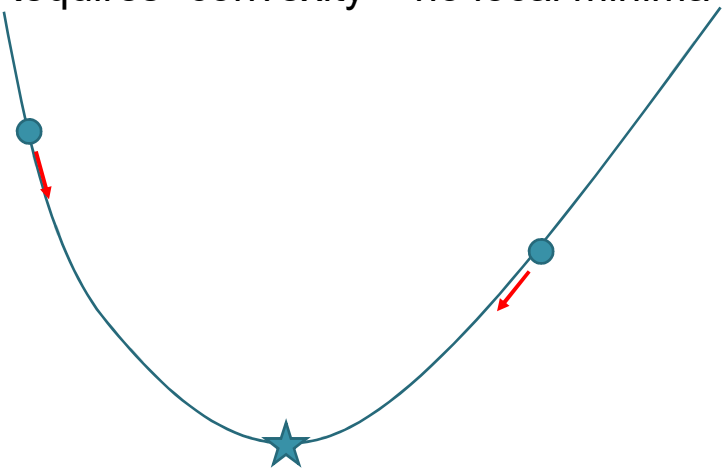
Questions

- Does anyone want to present Dec 13?
- Start early or run late on Dec 20?
- Should the final projects be made available to all students in this course?

Simulated Annealing

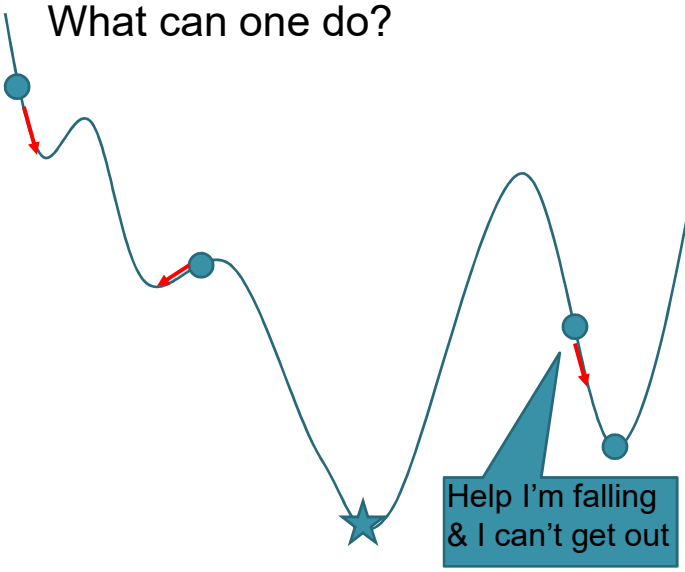
Gradient Descent

- Works in 2D all the way to ND
- "Requires" convexity = no local minima

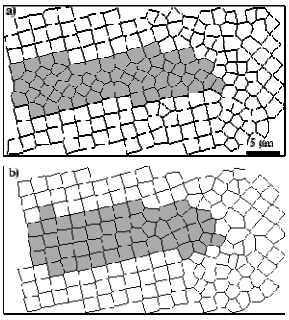


Gradient Descent

What can one do?



Annealing



- Concept: heat up materials and slowly cool
- Slow cooling will optimize crystal formation
- Rapid cooling will result in suboptimal crystals with worse properties

- http://www.preplift.com/blog/2015/06/30/metallurgy-and-materials-engineering-from-eyes-of-gate-air-3/
- http://physics.nyu.edu/grierlab/anneal7b/

Simulated Annealing Concept

- Goal of descent is to converge to globally optimal solution
- Most of the time, we want to improve our solution at each iteration
- But, search may get stuck in local minima
- Provide search with a probabilistic chance of accepting a worse solution with a certain probability
  - Provides ability to "jump out of" local minima

Simulated Annealing Aspect

- Ability to jump out is governed by energy or "temperature"
- Temperature decreases in each iteration
- Jump out early but later you can't jump out of global optimum

Simulated Annealing Algorithm

- Start with X → eval(X)
- Loop until done (or found minimum):
  - newX = move(X)
  - If eval(newX) < eval(X)
    - X ← newX
  - else if prob(X, newX, iteration) > threshold
    - X ← newX
  - else do nothing

Probability magic

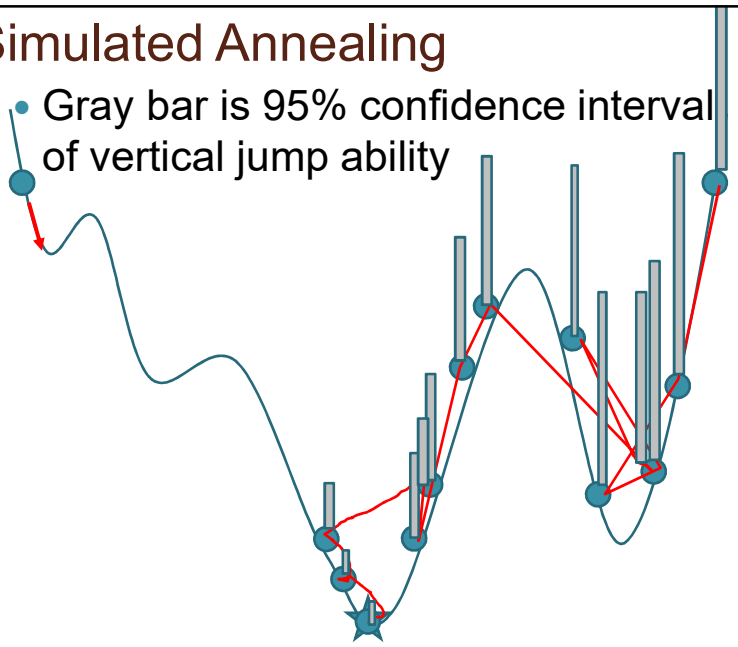
- prob(X, newX, iter) = exp[$\frac{-(eval(X)-e \quad (newX)}{T}$]

  T = temperature that decreases with iteration

- Probability distribution is called a Boltzman distribution
- T undergoes a cooling schedule such as $T_{new} \leftarrow 0.95\ T_{old}$
- Algorithm is called Metropolis algorithm after its creator in 1953
  - Sadly, no relationship to Superman

Simulated Annealing

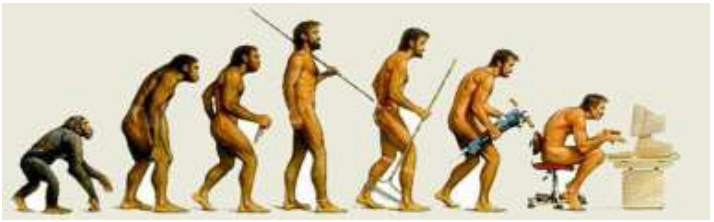
- Gray bar is 95% confidence interval of vertical jump ability



Genetic Algorithms

Genetic Algorithms

- Class of models inspired by evolution
- Previously discussed generative models
- Talk about evolutionary

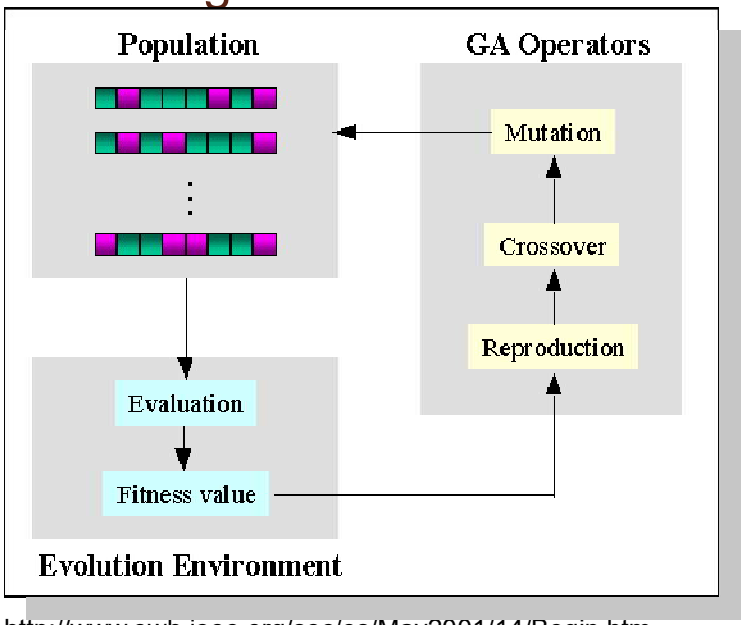


- http://www.ranchodinero.com/wp-content/uploads/2011/09/evolution.jpg
- https://encrypted-tbn3.gstatic.com/images?q=tbn:ANd9GcQ3cmTuxEFo12UYmR2xBbu40d6YvWG6srSbNiBykxD1cEhLvF1ZUx1nYQ

Evolution Ideas

- Encode information in a genome
- Phenotype ~ characteristics associated with a set of genes
- Population-based model with families carrying on lineage
  - "Begatting," mutation, cross-over
- Survival of the fittest → fitness function

Genetic Algorithms Visual



- http://www.ewh.ieee.org/soc/es/May2001/14/Begin.htm

Encoding

- Messiest part of GA
- Encode information in bit strings
  - Great for integers
  - Great for categories (One Hot Encoding)
  - Miserable for floats (IEEE standards)
- Often use Gray coding
  - Change 1 bit at a time
  - Proven to work better

Comparison of Binary-Coded and Gray-Coded Integers

Integer	Binary	Gray
0	000	000
1	001	001
2	010	011
3	011	010
4	100	110
5	101	111
6	110	101
7	111	100

Coding & Evaluation

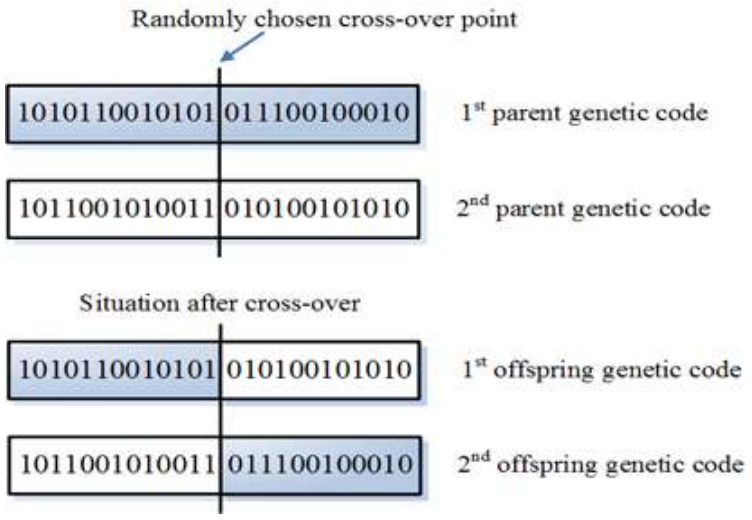
- Most of Machine Learning is based on optimization of some function
  - Least squares for regression
  - Min dist to center for K-means
  - AUC optimization
  - Neural network gradient descent
- Feat1, Feat2, … FeatN
  - [00110101101101010101011011010110]
  - f(feat1,feat2,…featN) = f([bit string])

Generating variation

- Frequent part of machine learning and statistical approaches as well

- Bagging
- Random forest row/feature sampling
- K-mean++ → sample from probabilities
- Leader randomly reorder
- Diversity in recommender engines

Begatting

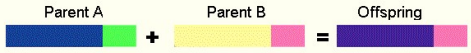


- http://www.ro.feri.uni-mb.si/predmeti/int_reg/Predavanja/Eng/3.Genetic%20algorithm/images/pic005.png

Other types of Begatting

- http://www.obitko.com/tutorials/genetic-algorithms/crossover-mutation.php

Single point crossover - one crossover point is selected, binary string from beginning of chromosome to the crossover point is copied from one parent, the rest is copied from the second parent



11001011+11011111 = 11001111

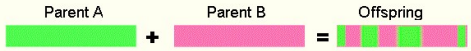
Purple should be blue?

Two point crossover - two crossover point are selected, binary string from beginning of chromosome to the first crossover point is copied from one parent, the part from the first to the second crossover point is copied from the second parent and the rest is copied from the first parent

11001011 + 11011111 = 11011111

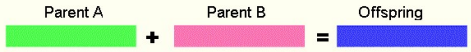
Purple should be blue?

Uniform crossover - bits are randomly copied from the first or from the second parent



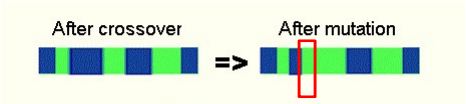
11001011 + 11011101 = 11011111

Arithmetic crossover - some arithmetic operation is performed to make a new offspring



Humans: 1-2 crossovers per chromosome

Mutation = bit inversion



11001001 => 10001001

- Typically fairly low rate of mutation
- Need to tune this parameter

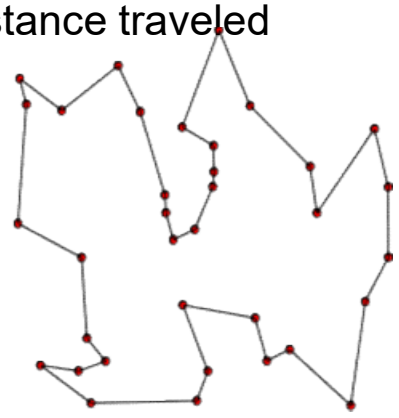
- Humans: 25/billion per generation
- Humans: 3 billion base pairs
  - 20k-25k human protein-coding genes
  - Gene lengths are ~10-15 kB each

"Genetic Algorithm" Algorithm

- Initialization: sample variations
- Loop until done:
  - Compute fitness function for current set
  - Keep the best solutions
  - "Begat" among the best solutions

Traveling Salesman Problem (TSP)

- A salesman must visit all cities at least once and return home
- Goal is to find the city order that will minimize the distance traveled
- NP-complete

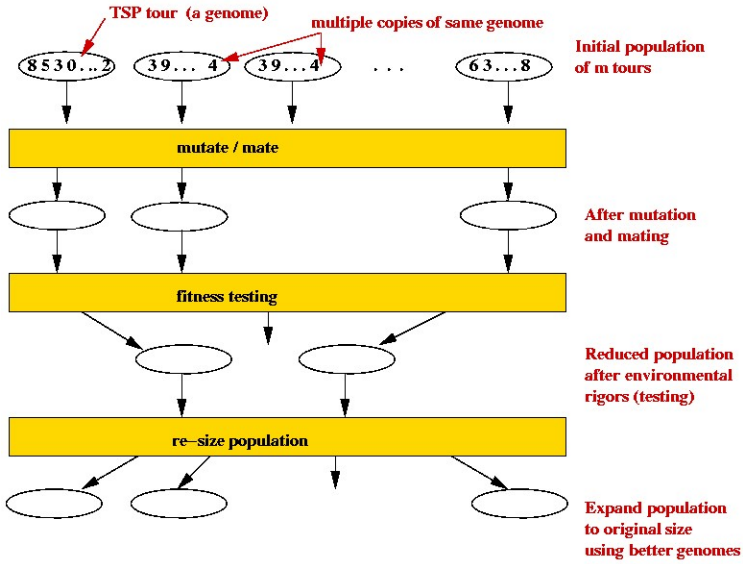


How do we set it up?

- Let's say we have 8 cities

Traveling Salesman Example

- http://www.seas.gwu.edu/~simhaweb/cs177/bioalgorithmlecture/



TSP Setup

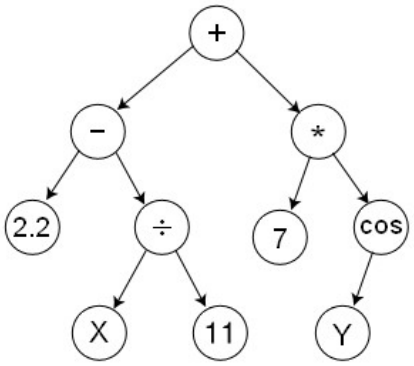
- Encoding:
  - 8 cities all of which must be included
  - Represent each city with 3 bits
  - Chromosome = 3x8 = 24 bits
- Scoring function:
  - Sum of distances in the order
- Cross-over modification
  - 12345678 + 16237854 Left same
  - 12346785 + 16234578 Right from other
- Mutation modification
  - Can't flip bits so switch 2
  - 12345678 → 12745638

Genetic Programming

- Goal is to extend genetic algorithms using the same general structure but…
  - Want to encode an equation
  - Example: classification boundary
  - Example: orbital mechanics from data

Genetic Programming

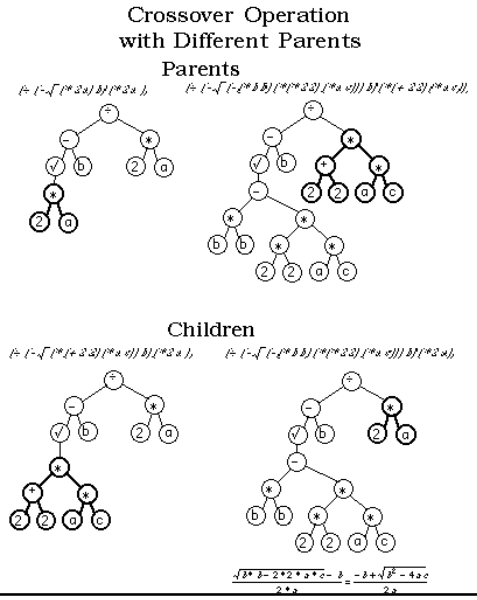
- Represent numbers with bits
- Represent numeric functions with bits



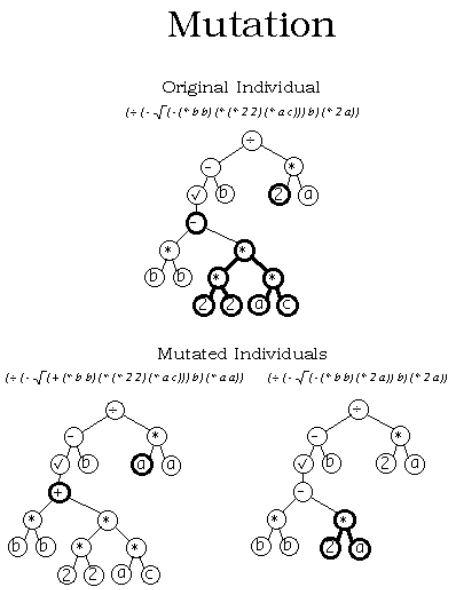
$$\left(2.2-\left(\frac{X}{11}\right)\right)+\left(7*\cos(Y)\right)$$

Genetic Programming Crossover

- http://www.geneticprogramming.com/Tutorial/



Genetic Programming Mutation



Genetic Programming Issues

- How could you prevent overtraining?

Network Analysis

Network applications

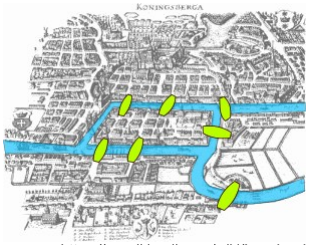
- Search the web
- Leverage Facebook or LinkedIn
- Understand human disease
- Understand biology
- Interpret neuroscience
- Advertising

More & Larger Networks Everywhere

- Social networks (Facebook, LinkedIn)
- Economic networks
- Biological pathway networks
- Journal authorship networks
- Road/river/plane transportation paths
- Sexual partner networks
- Epidemiology disease spread networks
- Power networks (electricity)
- World Wide Web

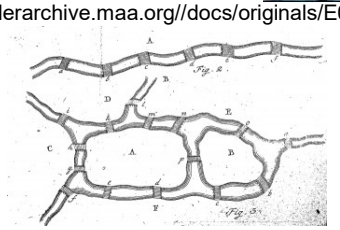
Network Science

- Graph theory: Euler's first paper 1736
- Seven bridges of Konigsberg



https://en.wikipedia.org/wiki/Leonhard_Euler

- http://eulerarchive.maa.org//docs/originals/E053.pdf



Network Science

- Graph Theory
- Ecological Networks
  - 1880 Food web graph by Camerano

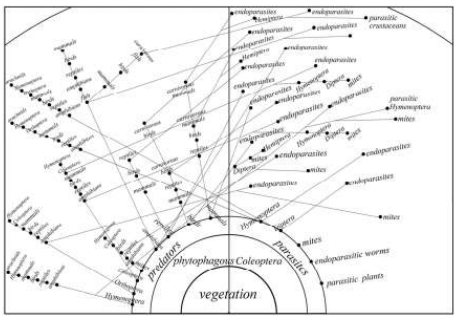
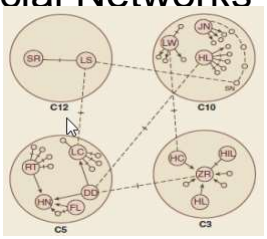


Figure 2. The first food-web graph reported in the literature, from Camerano (1880) (17). The graph was redrawn to improve readability. The structure is similar but species names were translated from Italian to English. (Adapted from Camerano, 1880) (17).

  - History of the study of ecological networks
    - Bersier 2007

Network Science

- Graph Theory
- Ecological Networks
  - 1880 Food web graph by Camerano
- Social Networks 1930s Moreno

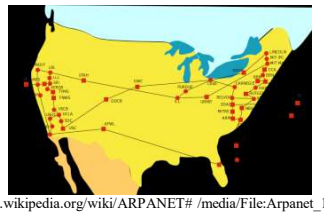


http://science.sciencemag.org/content/323/5916/892.full

- Internet 1960s, then 1980s
- Ecological Networks 1979

Network Science

- Graph Theory
- Ecological Networks
- Social Networks 1930s Moreno
  - Children at play: small networks
- Communications networks 1960s
  - Early internet (ARPANET)



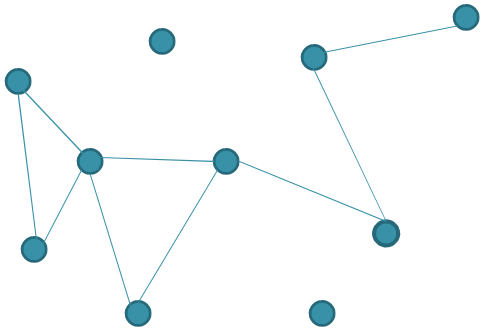
https://en.wikipedia.org/wiki/ARPANET# /media/File:Arpanet_19

- Explosion of WWW
- Facebook & social networking

Market Capitalization of Networks

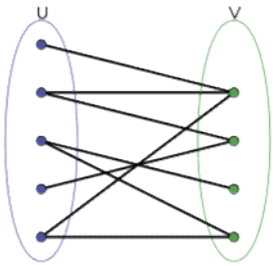
- Cisco: $168 billion
- Facebook $519 billion
- LinkedIn purchased by Microsoft
  $26 billion in 2016
- Netanomics network analysis company
  - NetMapper: extracts networks from texts

Network



Bipartite Graph (Bigraph)

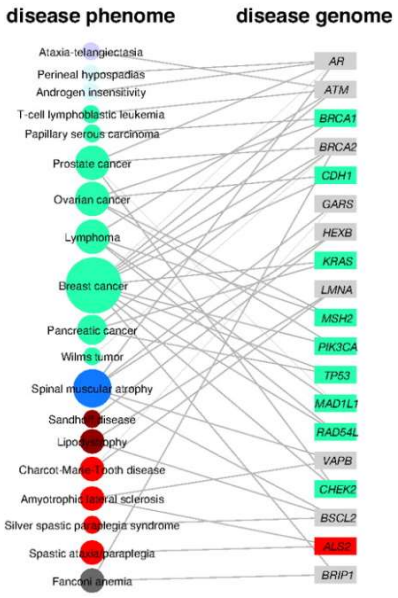
- Graph that can be divided into two disjoint sets U and V such that every node in U is connected to one in V



Wikipedia

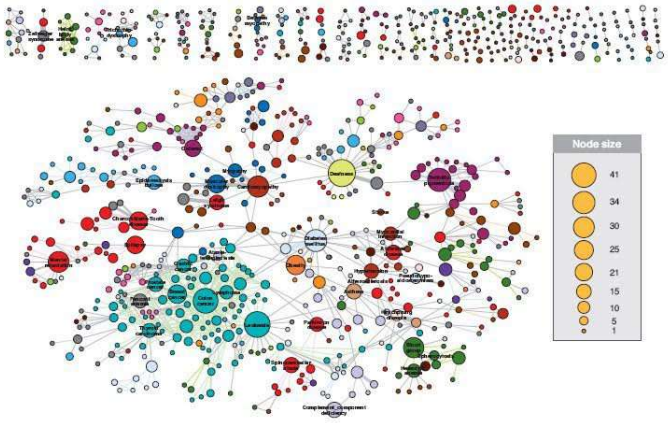
- Uber drivers – Passengers
- Ingredients – Dishes

Bipartite Graph



http://www.pnas.org/content/104/21/8685.full

Genome-Wide Association Studies and Human Disease Networks



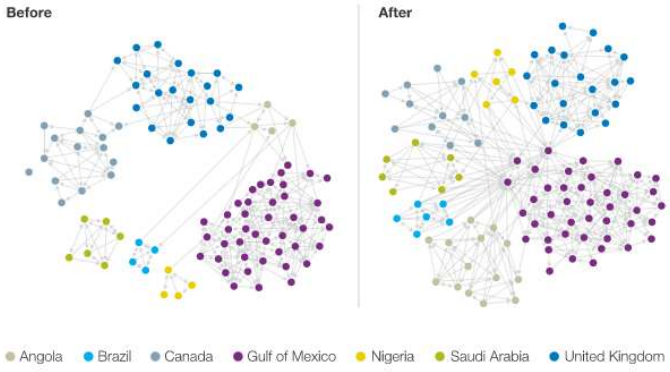
In the HDN, each node corresponds to a distinct disorder, colored based on the disorder class to which it belongs, the name of the 22 disorder classes being shown on the right. A link between disorders in the same disorder class is colored with the corresponding dimmer color and links connecting different disorder classes are gray. The size of each node is proportional to the number of genes participating in the corresponding disorder (see key), and the link thickness is proportional to the number of genes shared by the disorders it connects. We indicate the name of disorders with 10 associated genes, as well as those mentioned in the text.

-- Leslie Pray 2008

Corporate Structure Analysis

One oil company used a social-network analysis to target improved communication between field workers and technical experts.

Social-network analysis at a major oil and gas company

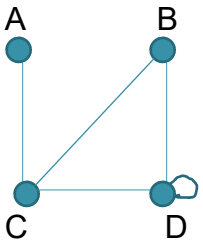


- https://www.mckinsey.com/business-functions/organization/our-insights/organizing-for-an-emerging-world

Network

- Node
  - Can have values or attributes
- Edge = link = tie
  - Could be present or absent
  - Could have weights
  - Could be directed or undirected
  - Represent edges using adjacency matrix
- Can represent graphically or with an adjacency matrix

Representing Network



Connections of nodes

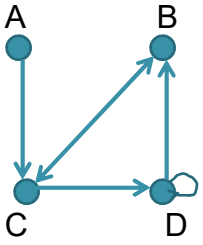
	A	B	C	D
A	0	0	1	0
B	0	0	1	1
C	1	1	0	1
D	0	1	1	1

Adjacency Matrix

0	0	1	0
0	0	1	1
1	1	0	1
0	1	1	1

{AC, BC, BD, CD, DD}

Representing Directed Network



into

out

	A	B	C	D
A	0	0	1	0
B	0	0	1	0
C	0	1	0	1
D	0	1	0	1

Adjacency Matrix

0	0	1	0
0	0	1	0
0	1	0	1
0	1	0	1

Notation: Out-In
{AC, BC, CB, CD, DB, DD}

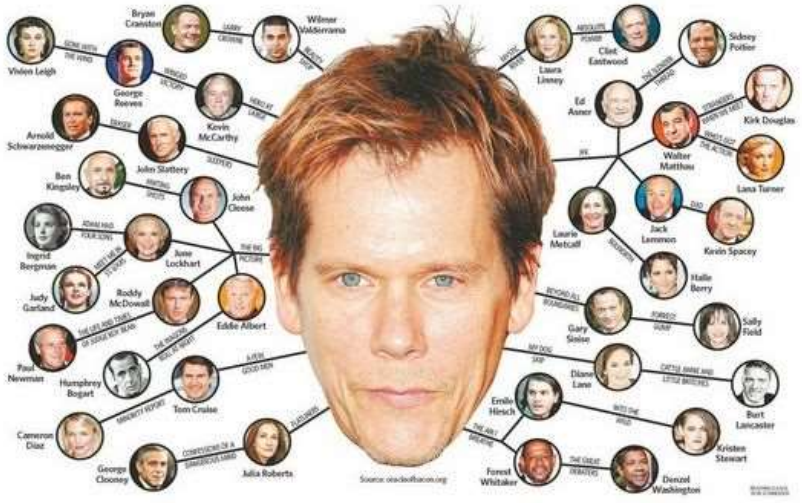
Network Properties

- N = #nodes
- E = #edges
- Degree L (or “k”) = #edges from node
- Degree distribution: P(x)
  - Normalized histogram as f(degree)
- Diameter = d = greatest distance between any pair of nodes in network
  - = max shortest path between nodes

Diameters for real graphs

- World population 7.5 billion
- 100 friends on Facebook
- ln(7.5 billion) / ln(100) = 25 / 4.6 = 5.43
- Corresponds to 6 degrees of separation

Kevin Bacon: 6 degrees



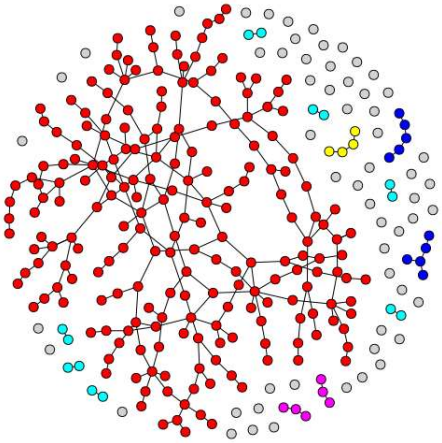
http://www.readingeagle.com/life/article/six-degress-of-kevin-bacon-a-game-changer

Network Density

- Actual links / possible links
- Density
  - Maximum links in network: N(N-1)/2
  - Complete graph has maximum links
  - Average degree k = N-1
- Sparse network: L << Lmax
  - Many common networks L ~sqrt(Lmax)

Giant component

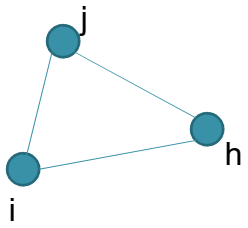
- Connected component that occupies a [large] constant fraction of the nodes



https://linbaba.wordpress.com/2010/10/15/fluid-limits-and-random-graphs/

Clustering Coefficient

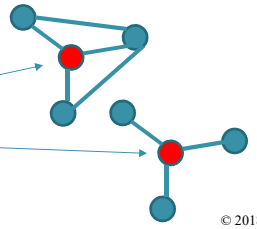
- C(i) = $\frac{\sum_{j \neq h}^{N} e_{ij} e_{jh} e_{hi}}{k_i(k_i - 1)}$



- Numerator is count of triangles that include the node *i* in question
- k_i = #neighbors
- C(g) = average Ci = $\frac{\#triangles}{\#connected\ triples}$

Clustering coefficient

- Fraction of neighbors connected
- k_i = degree of node i
- e_i = number of edges connecting the neighbors of node i
- $c_i = 2\ e_i\ /\ [\ k_i(k_i\text{-}1)\ ]$ = value in [0, 1]
- $c_i = 2*2/(3*2) = 2/3$
- $c_i = 2*0\ /\ (3*2) = 0$



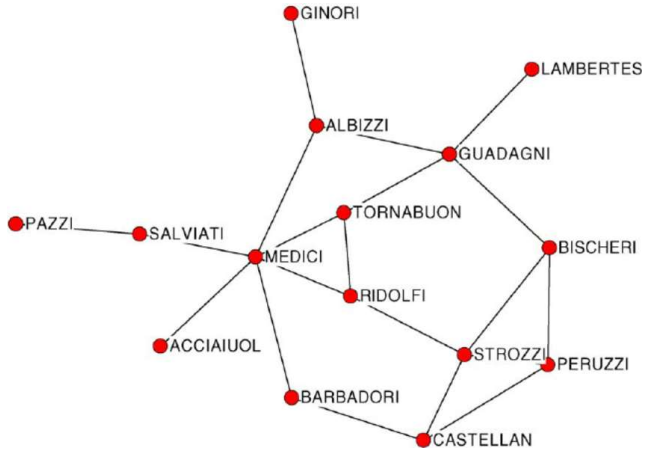
Centrality Measures

Empirical Analysis

- What if you advertise to leaders of Fortune 500 companies in New York
- What if you advertise to Boston startups
- Conduct a survey in NY & Boston areas
- Examine the various centrality measures
- Characterize the networks of information based on the flow of information

Florentine Family Marriages

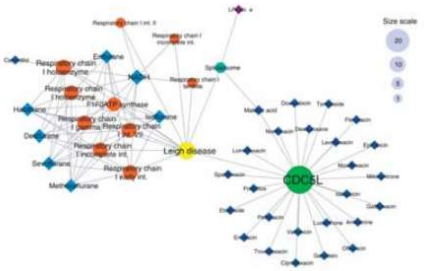
- Nodes are families during Renaissance
- Which family has most impact?



Centrality and Network Flow Jan 2005

Find Cures to Diseases

Centrality analysis is used to identify new pharmacological strategies



Source: Modularity in Protein Complex and Drug Interactions Reveals New Polypharmacological Properties, Jose C. Nacher mail, and Jean-Marc Schwartz, *PloS One, 2012*

- Yellow node is the disease nodes, protein complexes are circles, and diamond nodes are drugs.
- Links between the disease node and protein complexes represent associations between genes involved in these complexes and the named disease, as specified by the Disease Ontology.
- A drug is connected to a protein complex if at least one protein target of the drug is also a subunit of the protein complex.

https://www.cs.purdue.edu/homes/neville/courses/NetworkSampling-KDD13-final.pdf

Centrality

- Addresses which nodes are most important within a graph
- Which people are most influential

- Multiple measures based on various definitions of 'important'
  - Some measure information flow
  - Some measure cohesion or closeness

Node Centrality

- Degree Centrality: Degree / (n-1)
- Closeness of node i: (n-1) / Σ_j lsp(i,j)
  - lsp=len shortest path between i and j
  - Proportional to distance
  - What if a node is disconnected?
- Decay centrality $C_i^d(g) = \sum_{j \neq i} \delta^{lsp(i,j)}$
  - Exponential decay across distance
  - Delta near 1 (no decay) → component size
  - Delta near 0 (full decay) → degree
  - Normalized version: $C_i^d(g) = \frac{\sum_{j \neq i} \delta^{lsp(i,j)}}{(n-1)\delta}$

Node Centrality

- Centrality (Freeman) of node k
  - $Betweenness = \frac{\sum_{i,j \neq k} \frac{P_k(i,j)}{P(i,j)}}{\frac{1}{2}(n-1)(n-2)}$
  - P(i,j) = #shortest paths between i and j
  - P_k(i,j) = P(i,j) that contain k (i,j ≠ k)
- Random-walk Betweenness
  - Choose 2 nodes m and n at random
  - Start at node m and randomly walk to node n
  - Betweeness(i,j) = p(i,j)
  - = prob of crossing ij given all choices of paths

Eigenvector Centrality

- Centrality = function(neighbors' centrality)
- $\text{Centrality}_i = a\Sigma_j g_{ij} C_j$
  - Write this as a matrix since gij are weights
  - C = agC is a self-referential equation
  - Use the eigenvector with max eigenvalue
- Approach related to Google's PageRank
  - Links to other pages were the measure
  - Characterized used eigenvector centrality

Bonacich-Katz Centrality

- Evolving weighted sum of neighbors
  - Include weighted sum of their neighbors
  - Include weighted sum of their neighbors
- $C = ag + kgag + k^2g^2ag + \ldots$
- k = small weighting term for distance
- $\text{Centrality} = \Sigma_i k^i g^i ag$

TOM: Topological Overlap Matrix

- Based on the adjacency matrix A
- $TOM_{ij} = \frac{\sum_{k \neq i,j} A_{ik} A_{kj} + A_{ij}}{\min(\sum_{k \neq i,j} A_{ik}, \sum_{k \neq i,j} A_{kj}) - A_{ij} + 1}$
- Measure of overlap of neighbors between nodes i and j

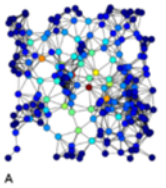
Network Analysis with the Enron Email Corpus, Hardin, Sarkis, URC 2015

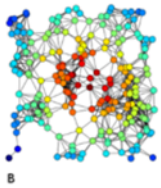
Fabrikant's centrality measures

- Average #hops to other nodes
- Max #hops from other nodes
- #hops from the network center

Centrality Measures Visualized

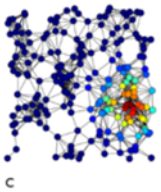
Betweenness

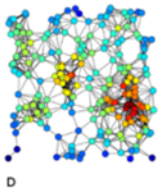




Closeness

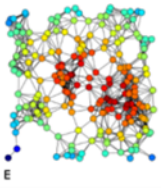
Eigenvector

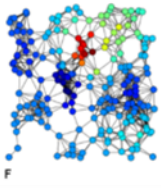




Degree

Harmonic

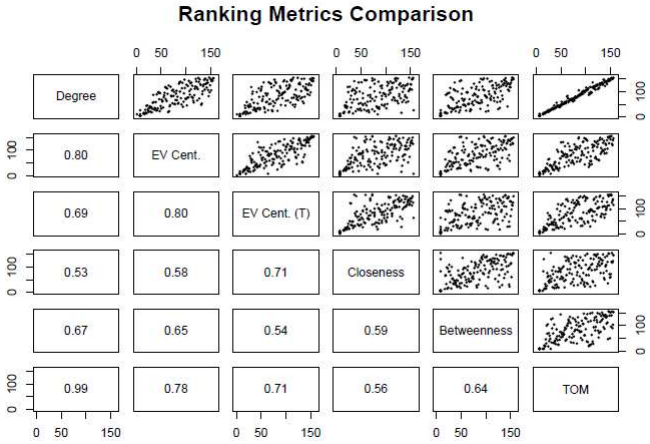




Katz

- https://en.wikipedia.org/wiki/Centrality

Comparing Metrics (Enron dataset)



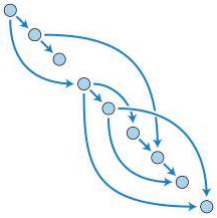
- Clearly don't all measure the same thing
- https://ww2.amstat.org/publications/jse/v23n2/hardin/EnronTutorial.pdf

Properties of Networks

- G(N) = generator of networks
  - Random networks produce sets of networks
- A(N) = property of network
  - Subset of networks G(N) that have property
  - Example: A(N): average degree > 3

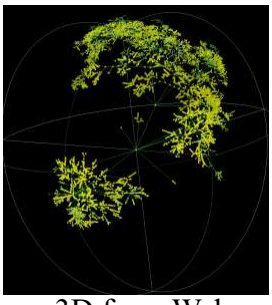
- Monotone property of A(N)
  - Property that is satisfied for g
  - Property is satisfied if add extra links to g

Visualization Aspects

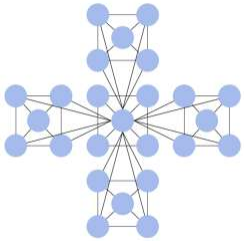


https://en.wikipedia.org/wiki/Directed_acyclic

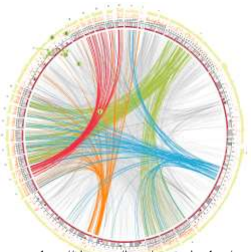
Time-oriented



3D from Walrus

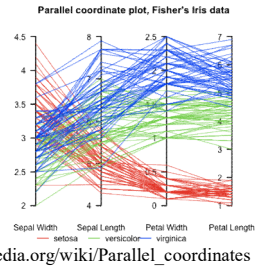


http://www.conceptdraw.com/How-To-Guide/hierarchical-network-topology



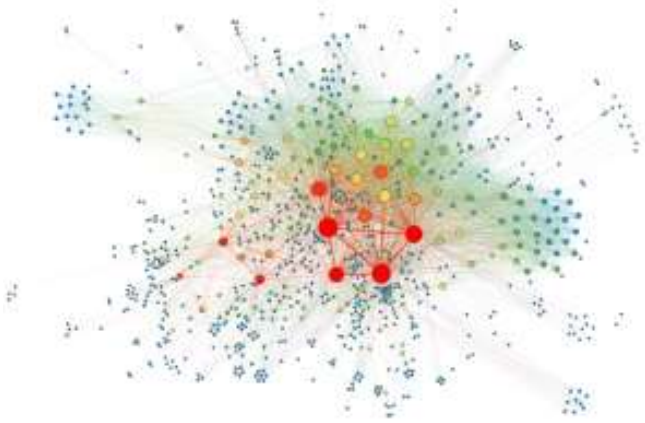
http://circos.ca/intro/genomic_data/

Circos Plot



https://en.wikipedia.org/wiki/Parallel_coordinates

Force-Directed Graph



https://bl.ocks.org/mbostock/4062045

Software Packages

Free Software

Name	Description
Krackplot 2.0.	Graph drawing software by David Krackhardt
UCINET IV/X for DOS	Software for the analysis of social network data. A manual is included in WordPerfect format, but MS Word users should be able to read it.
Anthropac 3.22	Cultural domain analysis software. Does not run on fast machines.
KeyPlayer 1.0	Identifies sets of key nodes that one would either want to remove to maximally disconnect a network, or contact to learn the most about a group or persuade in order to maximally influence others. Zip file.
NetDraw	Draws networks. Reads UCINET datasets, and can import Pajek files as well.
EICENT	Generalizes centrality measures to combine with attributes of nodes -- e.g., betweenness between specified groups

- Pajek is one of my favorites
- Graph-tool library in python
- NetworkX in python
- http://www.analytictech.com/free_software.htm

Network Models

Network Models

- Why?
  - Might want to do A-B testing on your UI
  - Assume everyone is equivalent – not true
- Want to understand network structure or at least make a model of it
  - Will enable you to effectively sample it
- Want to understand how it will grow

Network Models

- Erdos-Renyi random graph $G_{n,p}$
  - n vertices with pairs connected iid prob p
- Stochastic block model
  - Generalized random graph
  - k communities each with probability p_α
  - Connect nodes i, j with probability $B_{\alpha i \alpha j}$
- Power-law
  - Long tail with average degree → n^α

$$d_i = n^\alpha \left(\frac{n}{i}\right)^\beta \rightarrow \Pr\{d_i \geq k n^\alpha\} = k^{-1/\beta},\ 0 < \beta < 1$$

- Preferential attachment

Random Networks

Stochastic network

- Network with probability of each edge
- Probabilities occur on each of edges
- "Row stochastic" = Every row sums to 1

- Stochastic network is not one network
- Statistical ensemble of networks

Erdos-Renyi Random Graphs

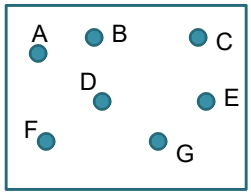
- Developed in 1960

http://netsci2015.net/index.php/erdoes

- G(N,p) model
  - Create N nodes
  - For each pair,
    - Create a connection with probability p
    - Flip a coin essentially for each node pair
- G(N,L) variation:
  - Specify L links placed randomly

- Standard model for networks
- Works well for large networks

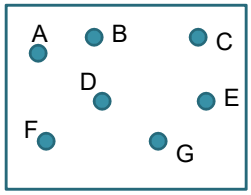
Erdos-Renyi G(N,p)



P = 30%

1st	2nd	Link?
A	B	
A	C	
A	D	
A	E	
A	F	
A	G	
B	C	
B	D	
B	E	
B	F	
B	G	
…		
F	G	

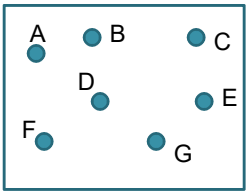
Erdos-Renyi G(N,p)



P = 30%

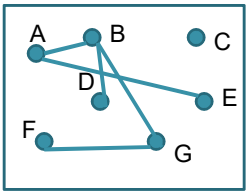
1st	2nd	Link?
A	B	yes
A	C	
A	D	
A	E	yes
A	F	
A	G	
B	C	
B	D	yes
B	E	
B	F	
B	G	yes
…		
F	G	yes

Erdos-Renyi G(N,p)



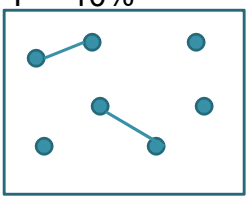
P = 30%

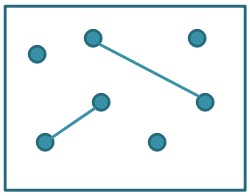
1st	2nd	Link?
A	B	yes
A	C	
A	D	
A	E	yes
A	F	
A	G	
B	C	
B	D	yes
B	E	
B	F	
B	G	yes
…		
F	G	yes

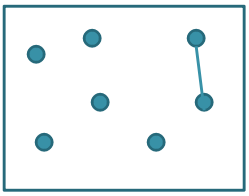


Erdos-Renyi G(N,p)

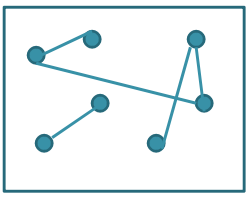
P = 10%

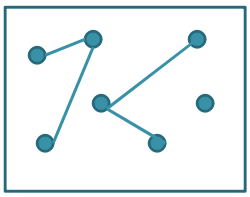


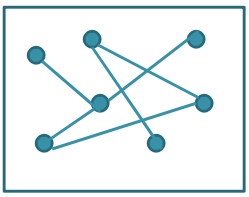




P = 25%

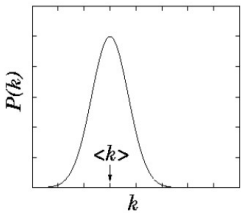






Properties of Erdos-Renyi Graphs

- Avg #edges: pN(N-1)/2
- Avg degree: p(N-1) or ~pN
- Degree distribution ~ Poisson
  - P(k) = $\frac{(pN)^k}{k!}\exp(-pN)$



- Cluster coefficient:
  - = #triangles/#connected triples
  - Generally no clustering
  - More like a tree network

Poisson Random Network

- Clustering for this network = p
- We know the probability of a given link is defined as p from the definition
- Poisson Random networks are good approximations of Erdos-Renyl random networks for large n, small p
- In reality, cluster values >> p so not really random graphs

Erdos-Renyi Properties

- k(n) > (1+ε)log(n) for ε > 0
  - Ensures probabilistic chance of connectivity
- k(n)/N → 0 bound so not over-connected
- If hold, then for large n:
  - Average path length: log(N)/log(avg k)
  - Average diameter: log(N)/log(avg k)
- Homogeneous network

Real Network Properties

			Degree k	Path length d		
NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,439	8.08	5.35	15	4.81
Actor Network	702,388	29,397,908	83.71	3.91	14	3.04
Citation Network	449,673	4,707,958	10.43	11.21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Given the huge differences in scope, size, and average degree, the agreement is excellent.

http://irt.enseeiht.fr/dhaou/NetworkScience/Network_Science.pdf

Giant Component

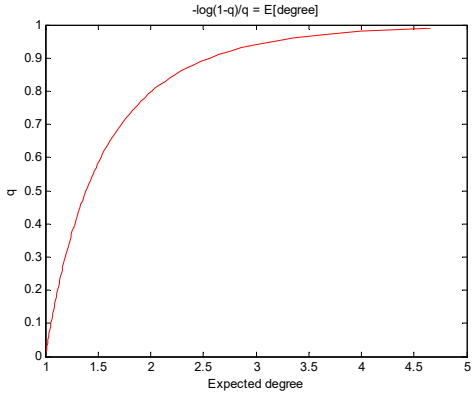
- Connected component that contains constant fraction of network nodes
- Erdos-Renyi network f(p)
  - If p < 1/n, then many isolated nodes
  - If p > log(n)/n then mostly connected
  - Else: giant component is interesting
- How to calculate size of giant component?

Calculation of Size of Giant Comp.

- Assume GC contains q% of nodes
- Prob(node not in GC) = 1-q
- Must state that neighbors not in GC
  - $(1-q)^{degree}$
  - $1-q = \Sigma_{neighbors} (1-q)^{degree} P(degree)$
  - P(degree) = chance of degree neighbors
    - Can apply Poisson or other distribution

Poisson Model

- E[degree] = -log(1-q)/q



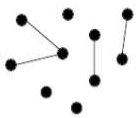
- If average degree = 3, pretty connected
- If average degree = 1, then not GC
- Rapid transition

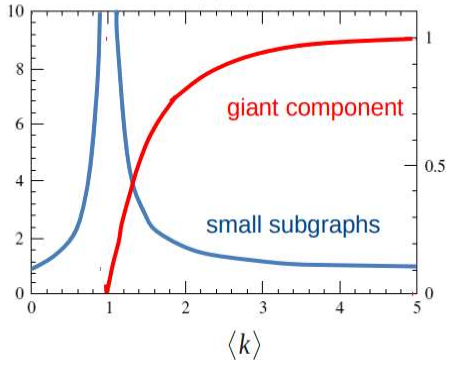
Connected Components

- Giant component avg degree d = 2E / N

$$N \to \infty \;\; s.t. \;\; \langle k \rangle = const = average\ degree$$

mean component size





relative giant component size

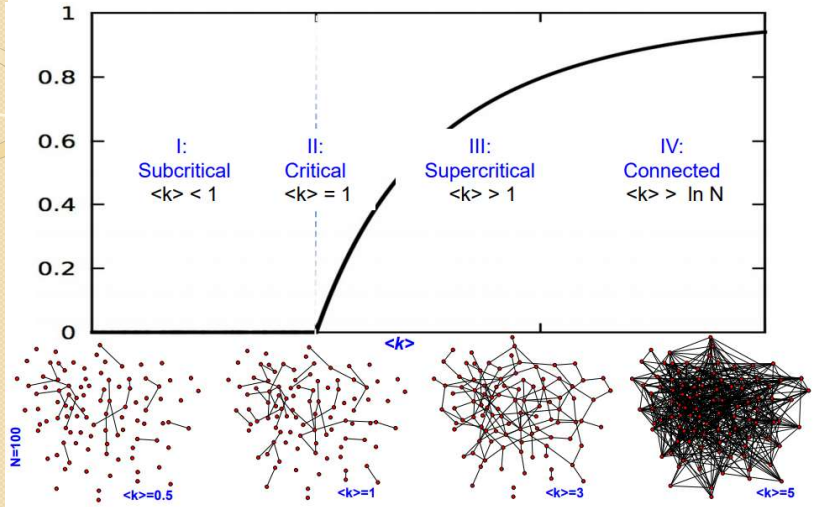
$\langle k \rangle < 1$: many small subgraphs

$\langle k \rangle >> 1$: giant component + small subgraphs

$\langle k \rangle = 1$: phase transition (percolation)

- Complex Network Analysis: Clustering Methods Nefedov 2013

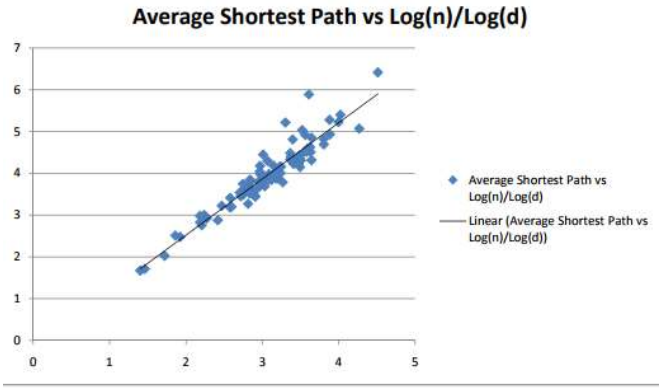
Random Network Transitions



- http://irt.enseeiht.fr/dhaou/NetworkScience/Network_Science.pdf

Health data set

- Famous data set of 84 high schools
- Recorded friends
- Avg shortest path :: log(N)/log(degree d)

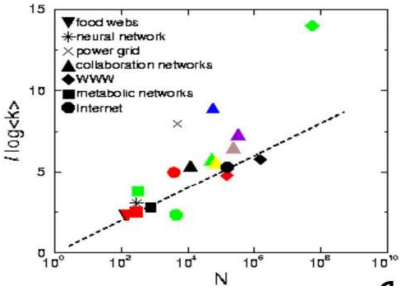


http://www.people.fas.harvard.edu/~bgolub/presentations/homophily.pdf

Random vs. Real Networks

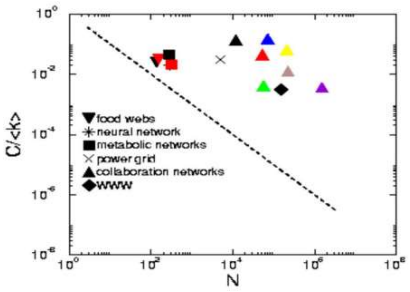
Degree vs. N

Log(<k>) = logN/AvgPathLen



Cluster Coeff vs. N

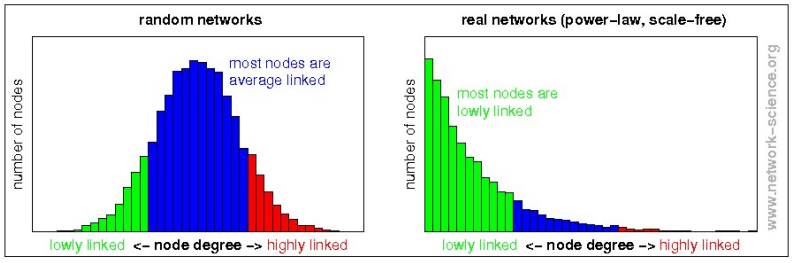
C = <k>/N = 2<L>/[k(k-1)]



- http://irt.enseeiht.fr/dhaou/NetworkScience/Network_Science.pdf

Random vs. Real Networks

Parameters govern the Poisson distribution but mean = peak

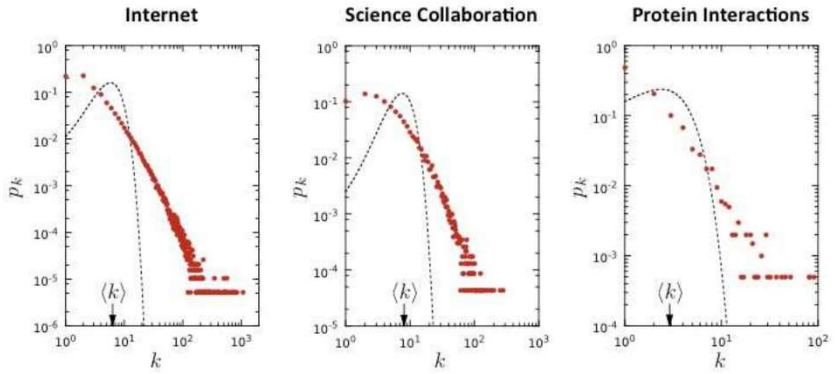


- http://www.network-science.org/powerlaw_scalefree_node_degree_distribution.html

Random vs. Real Networks

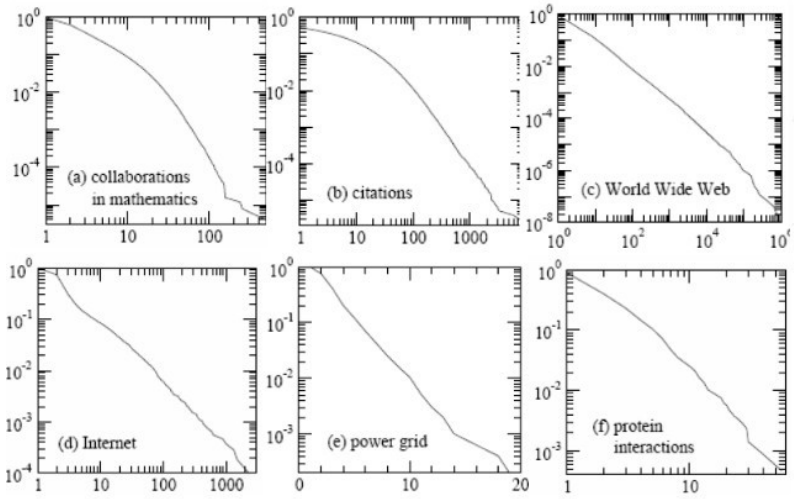
- Degree Distribution
- Poisson (black) vs. Real World

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$



- http://irt.enseeiht.fr/dhaou/NetworkScience/Network_Science.pdf

Standard Real Degree Distributions



- http://1.bp.blogspot.com/_kdPlXSbamSU/RujLKctJaBI/AAAAAAAAAE4/oOdFKUDXjJc/s1600-h/powerlaw.jpg

Small World Model

It's a Small World After all

- https://ohmy.disney.com/insider/2016/08/26/the-world-behind-its-a-small-world/

Small World

- Milgram 1967 (pre-email)
  - Took random people in Nebraska & Kansas
  - Asked them to get letters to targeted people in Boston
  - Restricted mail destinations
  - Only send to people they knew on a first-name
  - Passed instructions on

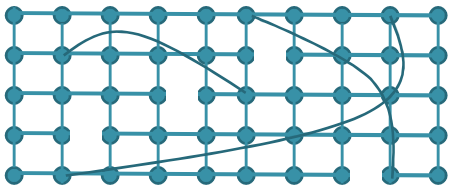
- 25% of letters reached the goal
- Letters reached goal in 6 steps
- Proved short paths exist
- Proved people can locate such paths

Navigating network

- Geographically greedy algorithm
  - Each node moves to neighbors closest to the destination node
- Delivery time is O(log(n)) if distant connections are added at rate p ~ $dist^2$
- If extend to large network for Milgram experiment, get 13% success rate vs. the 18% success rate for Milgram

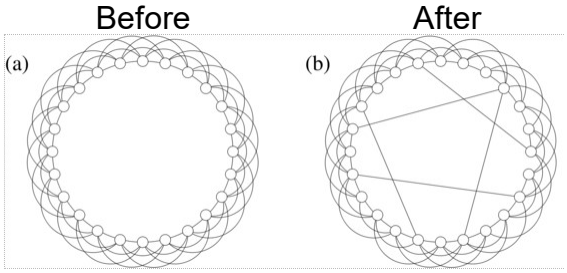
Small World Model

- Applicable for geographically constrained networks
- Hybrid between normal and random
  - Start with a low-dimensional network
  - Randomly add/remove links
  - Connect distant regions of network



Small World Network Model

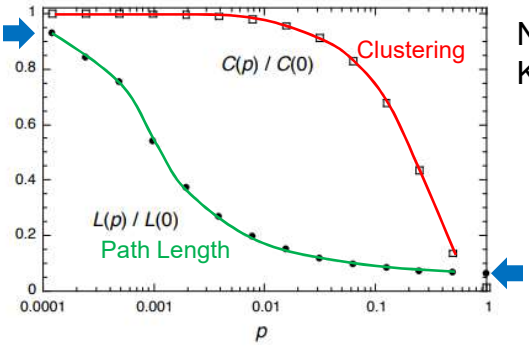
- Collective dynamics of 'small-world' networks. Watts & Strogatz 1998
- Take regular clustered network
- Rewire endpoint of each link to new location with probability p



http://www.scholarpedia.org/article/Small-world_network

Path Length L & Cluster Coefficient C

Regular Network



N=1000
K=10

Random Network

- Small-World is half way between regular and random networks depending on p

Small Worlds

- Low costs for local links
- Clustering will occur
- Few distant links
  - High cost
  - High value
  - Low diameter

Average Degree of real graphs

HS Friendships (CJP 09)	6.5
Romances (BMS 03)	0.8
Borrowing (BCDJ 12)	3.2
Co-authors (Newman 01, GLM 06)	
Bio	15.5
Econ	1.7
Math	3.9
Physics	9.3
Facebook (Marlow 09)	120

https://www.coursera.org/learn/social-economic-networks/lecture/XVc9H/1-7-diameters-in-the-world

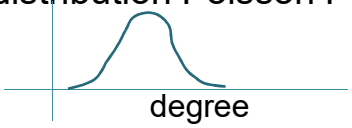
Degree Distributions

- Characterize network but the number of edges coming out of each node as a whole
- Degree distributions:
  - Binomial distribution G(n,p)
  - Poisson distribution for large n, small p
    - Approximation to a binomial distribution
    - Good for p values < 0.05 and 50+ n

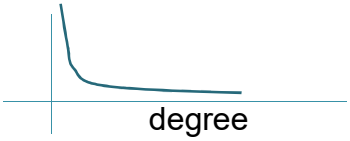
Power Law

Poisson vs. Power Law Network

- Poisson (Erdos-Renyl)
  - Random structure with tree structure
  - Degree distribution Poisson $P(k)=\frac{(pN)^k}{k!}\exp(-pN)$



- Power Law (scale free)
  - Many hubs
  - Degree distribution is power-law $P(k) = ck^{-\gamma}$



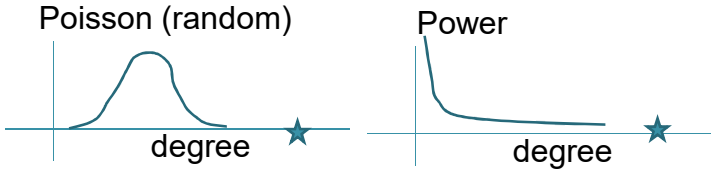
Problem with Random Networks

https://www.pinterest.com/Shakytail/african-fat-tail-gecko/

Problem with Random Networks

https://www.pinterest.com/Shakytail/african-fat-tail-gecko/

- Fat tails for real world data



- What are the stars on a network?

Fat Tails of Distributions

- Real world distributions tend to have fat tails in their distributions
  - Many have large #degrees
  - Many have small #degree
  - Middle are under-represented

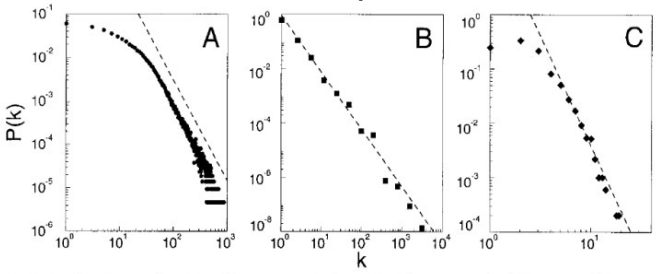


Fig. 1. The distribution function of connectivities for various large networks. **(A)** Actor collaboration graph with $N = 212{,}250$ vertices and average connectivity $\langle k \rangle = 28.78$. **(B)** WWW, $N = 325{,}729$, $\langle k \rangle = 5.46$ (6). **(C)** Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$.

Barbasi & Albert 1999

Power Law Model

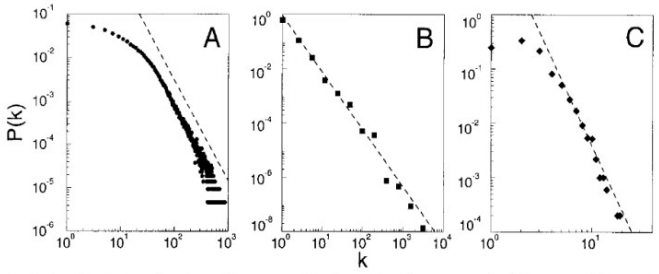
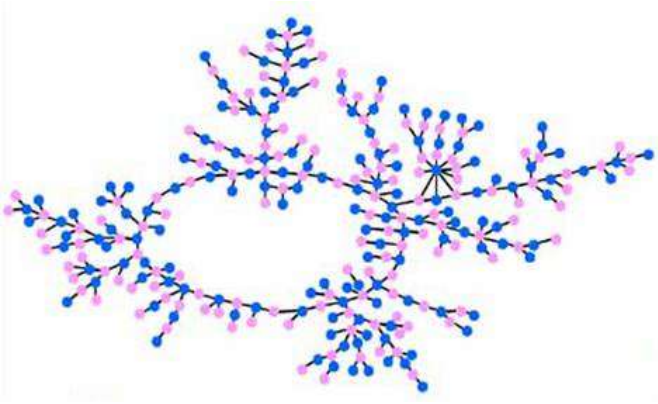


Fig. 1. The distribution function of connectivities for various large networks. **(A)** Actor collaboration graph with $N = 212{,}250$ vertices and average connectivity $\langle k \rangle = 28.78$. **(B)** WWW, $N = 325{,}729$, $\langle k \rangle = 5.46$ (6). **(C)** Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$.

- "Scale free distribution" or "power law"
  - k = degree of nodes
  - $P(k) = ck^{-\gamma}$ where γ usually in (2,3) range
  - $\log(P(d)) = \log(c) - \gamma\log(k)$ linear in log-log

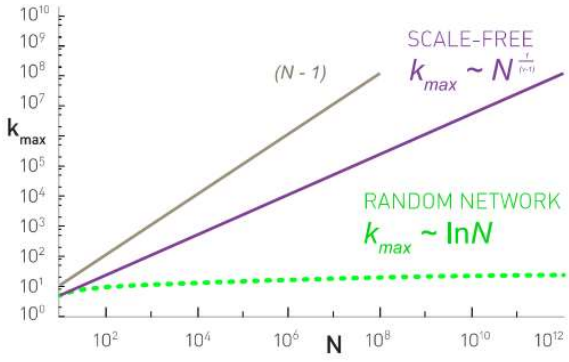
High School Romance

National Longitudinal Study of Adolescent Health 1995 in a mostly white Midwestern high school



http://www.nbcnews.com/id/6862058/ns/technology_and_science-science/t/high-school-romances-untangled/#.WZ5FVyh96Uk

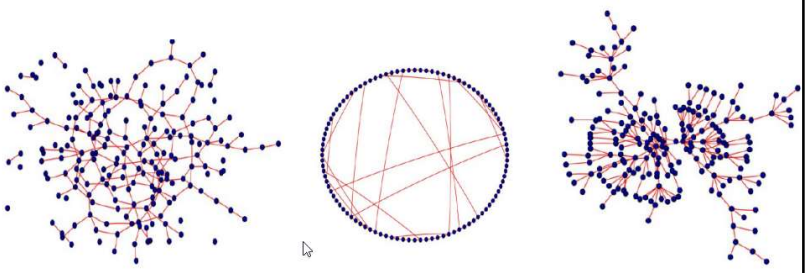
Power Law Properties



$$k_{max} = k_{min} N^{\frac{1}{\gamma-1}}$$

- http://irt.enseeiht.fr/dhaou/NetworkScience/Network_Science.pdf

Topology of Networks



Random $p = 0.02$ | Small world $p = 0.1$ | Scale free $\langle k \rangle = 2$

- http://people.ee.ethz.ch/~nnefedov/CNA_lecture_01.pdf

Network Property Summary

Erdös-Renyi model

- short path lengths
- Poisson distribution (no hubs)
- no clustering

Barabási-Albert scale-free model

- short path lengths
- power-law distribution for degrees
- robustness
- no clustering (may be fixed)

Watts-Strogatz Small World model

- short path lengths
- high clustering (*N* independent)
- almost constant degrees

Real-world networks

- short path lengths
- high clustering
- broad degree distributions, often power laws

- http://people.ee.ethz.ch/~nnefedov/CNA_lecture_01.pdf

Preferential Attachment

Preferential Attachment Algorithm

- Take an initial network with a few connected components
- Add new node of degree m
- Connect new node such that
  - Prob. of connecting to node i of degree ki
  - $= \pi_i = k_i/\Sigma k_i$

Preferential Attachment Properties

- Degree distribution $P(k) = 2m^2/k^3$
- Average shortest path lengths
  - Len proportional to ln(N)/ ln(ln(N))
- Clustering coefficient
  - C proportional to $(\log N)^2/N$

Stochastic Block Models

- Links between similar characteristics
- Essentially cluster the data
- Estimate probabilities inter or intra based on the clusters
- $\text{Log}(p_{ij}/(1-p_{ij}) = W_iX_i + W_jX_j + W_{ij}|X_i-X_j|$
  - X's represent clusters and W weights
  - Logistic regression format or log odds
  - Applicable to marriages between races

Homophily

Network Dynamics

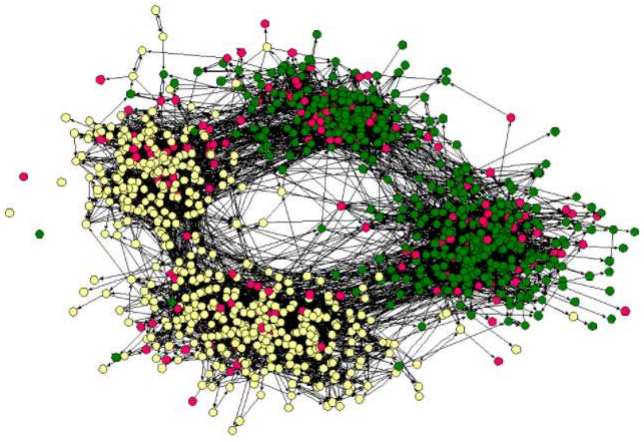
- Networks change over time
- Strength of links is not constant
- Impact of few strong vs. many weak

Homophily

- Birds of a feather flock together
- Linking is usually not random
  - Interracial marriages
  - Best friends
  - Businesses where you make purchases
- Useful for predicting links (LinkedIn)

Homophily

- Race, school integration, and friendship segregation in America American Journal of Sociology 107, 679=719 (2001)
- Color = race: Yellow=white, Green=African-American, Red=Hispanic
- Connections are best friends in a high school



http://social-dynamics.org/tag/homophily/

Types of Homophily

- Race
- Gender
- Age
- Religion
- Education, occupation, social class
- Value: way you think

- Wikipedia

Why Homophily?

- Geography
- Family/social connections
- Organizations
- Similar roles (managers)
- Cognitive: shared knowledge or culture
- Social pressure

- Wikipedia

Schelling Model

- Awarded 2005 Nobel Prize in economics
- Population has 2 groups
- Members have homophily goals:
  - Want ≥ thr neighbors of same group
- At each time point
  - Members move to unoccupied positions to satisfy homophily goal
  - Changes to network have follow-on effects
- Results in group spatial segregation

Forces of Homophily

- Fixed characteristics (race, gender) become correlated with other aspects (where to live)→ segregation
- Selection:
  - Grow connections with others who have similar characteristics
- Social influence:
  - Modify behavior to align to behavior of your neighbors

Network Stability & Assigning Link Value

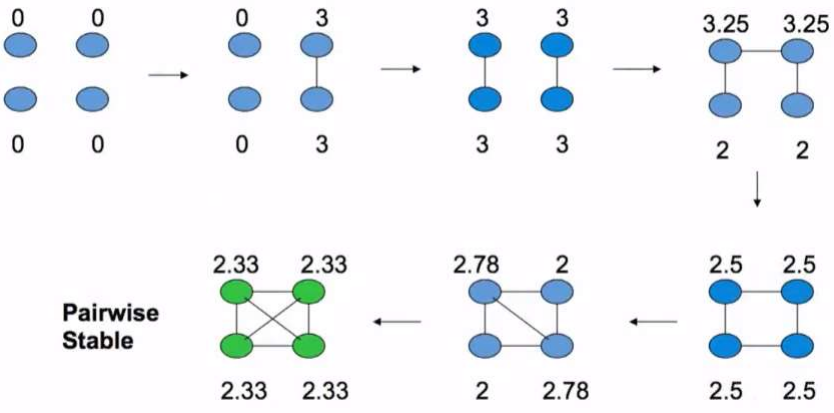
Dating Network

- Individuals alone don't choose
- Need both parties to declare intentions at the same time
- Game theory: Nash equilibrium
  - Assume players know strategies of others
  - Assume no player has anything to gain by changing only their strategy
  - Doesn't always work exactly

Pairwise Stability Model

- Alternative to Nash
- Pairwise stability instead:
  - No party gains from removing a link
    - $u_i(g) \geq u_i(g - ij)$
  - No pair both gain from adding a link
    - $u_i(g+ij) \geq u_i(g) \rightarrow u_j(g+ij) < u_j(g)$
  - Beneficial links pursued when available
  - Non-beneficial links should be removed
- Focus is only pairwise links (weak)

Jackson Example (Coursera)

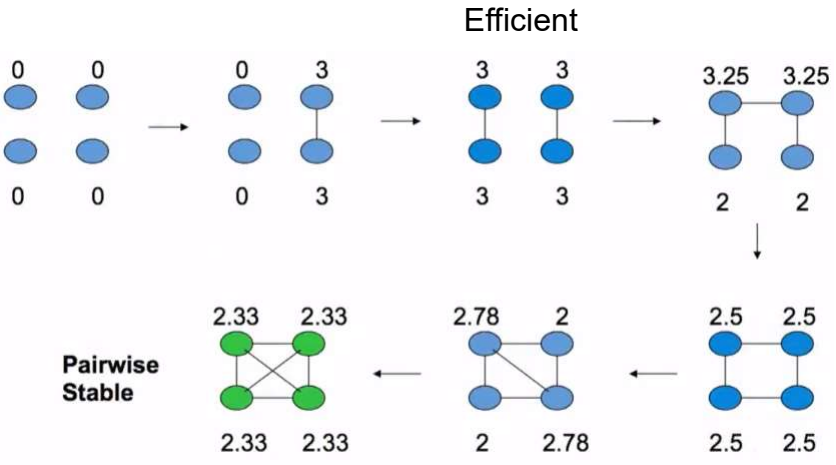


- Evolution to pairwise stable (value=10)
- Passed optimal (value=12)

Network Efficiency

- Relates to the ability to improve network
- Pareto efficient network
  - $\Sigma_i u_i(g') \geq \Sigma_i u_i(g)$ with condition $u_i(g') \geq u_i(g)$
  - Seek to optimize network globally
  - Improvements for some must be at least neutral for others (can't be worse)
- Efficient network maximizes $\Sigma_i u_i(g)$

Jackson Example (Coursera)



- Evolution to pairwise stable (value=10)
- Optimal efficiency (value=12)

Current & Future Work

Who cares?

- Networks for leveraging structured info
- MOLIERE: Automatic Biomedical Hypothesis Generation System. Sybrandt et al. KDD 2017
- Text mined medical abstracts → network
  - 24.5 million nodes, 989 million edges

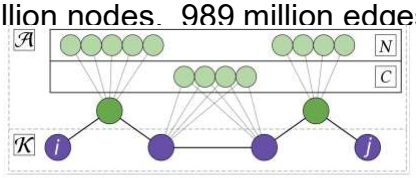
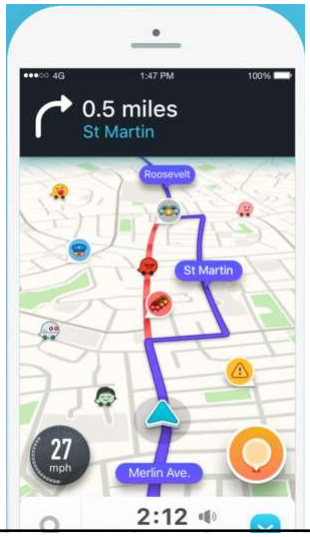


Figure 4: Process of extending a path to a cloud of abstracts.

- Created hypotheses to link terms i & j
  - N: nearest abstracts to i or j concepts
  - C: subset intersecting i and j neighborhoods
- Linked Venlafaxine & HTR1A/B

Waze

Navigating through a network with dynamically changing edges



Expert Finding

- Can you find experts in your company?
- Can you find expert advisors?
- Can you find the future experts?
- Can you find future collaborators?

Current Work

- Scaling up the network
  - Performing all the analyses
  - Characterizing the network
  - Facebook has 2 billion users (nodes)
- Working with streaming data and dynamic networks at scale
- Mapping other problems to networks
  - Classification, clustering, anomalies, etc.

References

- Outstanding Coursera course by a field leader
  - https://www.coursera.org/learn/social-economic-networks
- Tutorial slides by Barabasi (author in field)
  - http://irt.enseeiht.fr/dhaou/NetworkScience/Network_Science.pdf
- Tutorial on network analysis by Nefedov
  - http://people.ee.ethz.ch/~nnefedov/CNA_lecture_01.pdf
- YouTube on the future & cutting edge of network analyses with some math references
  - https://www.youtube.com/watch?v=9tny8lbt25A&t=3700s
- CSCI E-134 Harvard Extension: Networks
- Edx course on networks crowds and markets
  - https://www.edx.org/course/networks-crowds-markets-cornellx-info2040x-2

For Further Reading