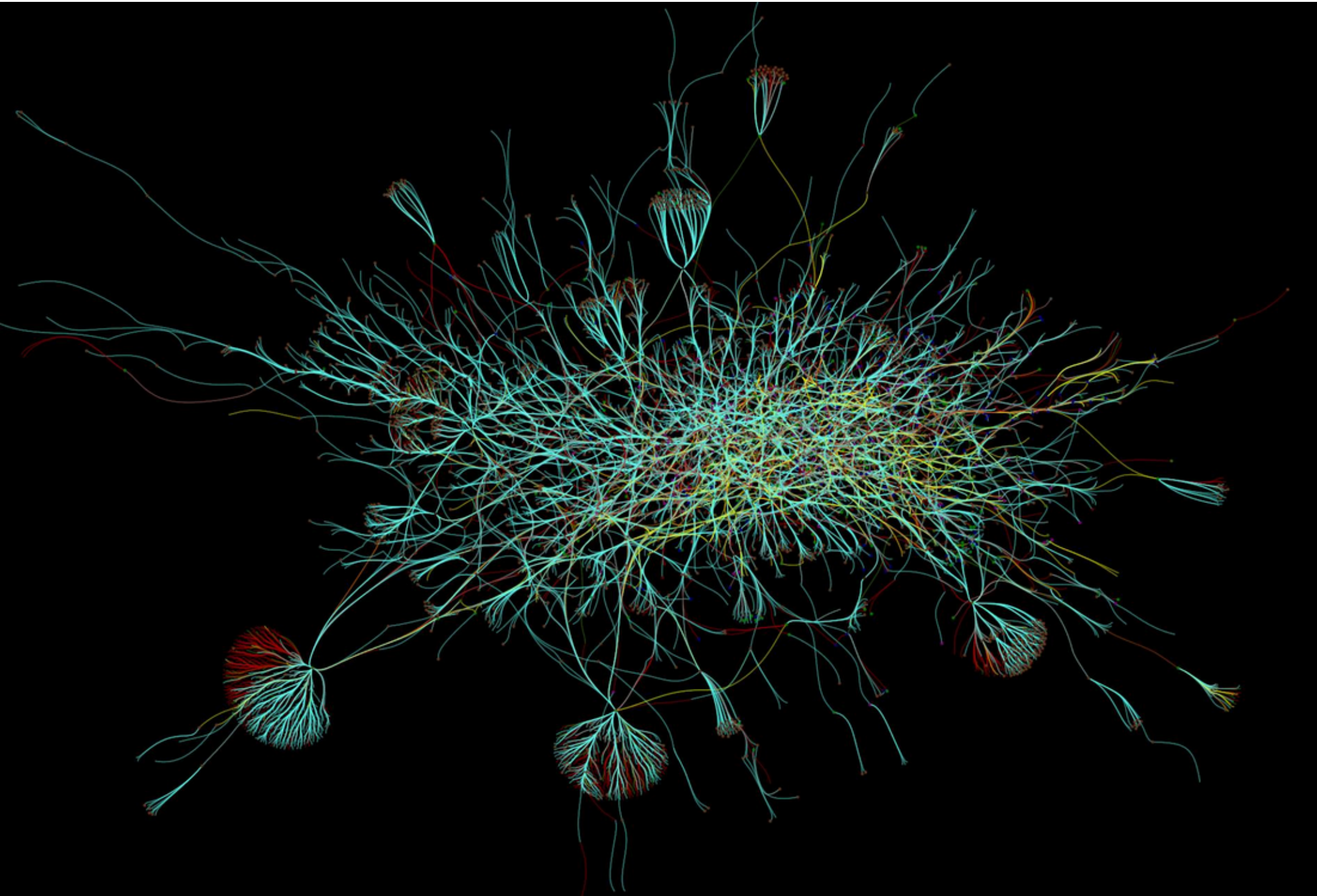
Networks

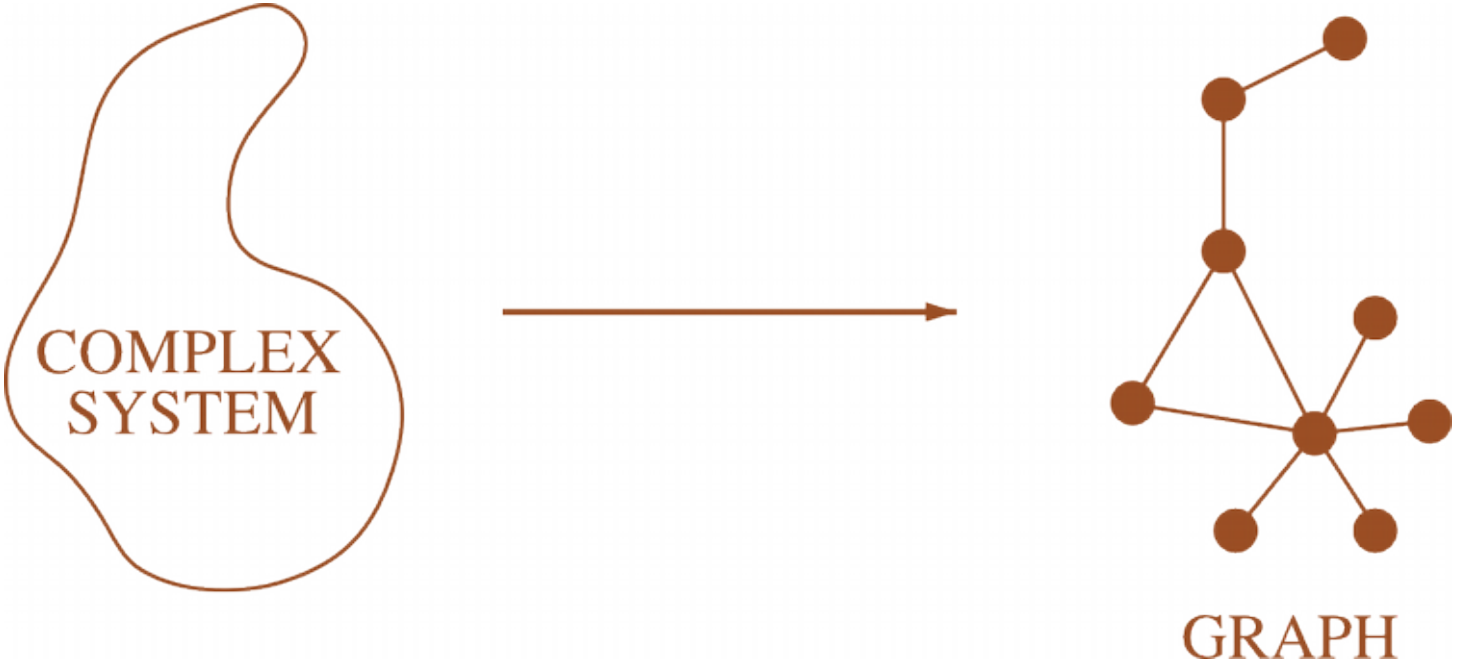


Graph theory vs. network science vs. systems science vs. ….

- Interdisciplinary area
- Difference is in approach & focus

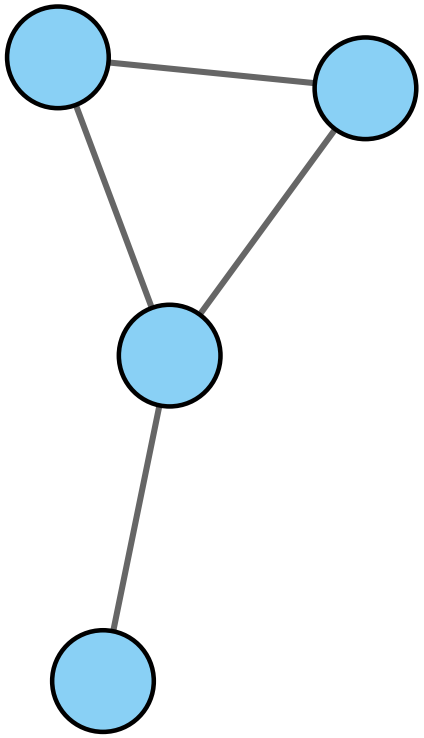
- Math-centric viewpoint: graph theory
  - existence theorems (eg. Szemerédy regularity lemma)
  - worst-case / pathological cases
- Physicist viewpoint: network science
  - Mean field approximations
  - “typical / average cases”

Why networks?



Defining a network

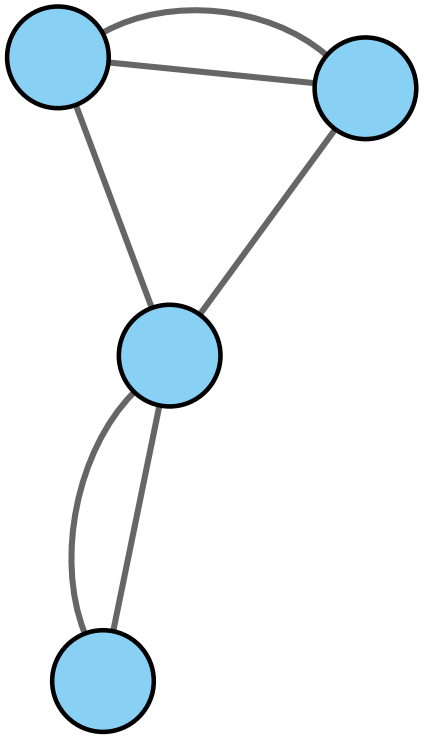
Defining a network: graphs



Simple graph:

- Nodes
- Edges

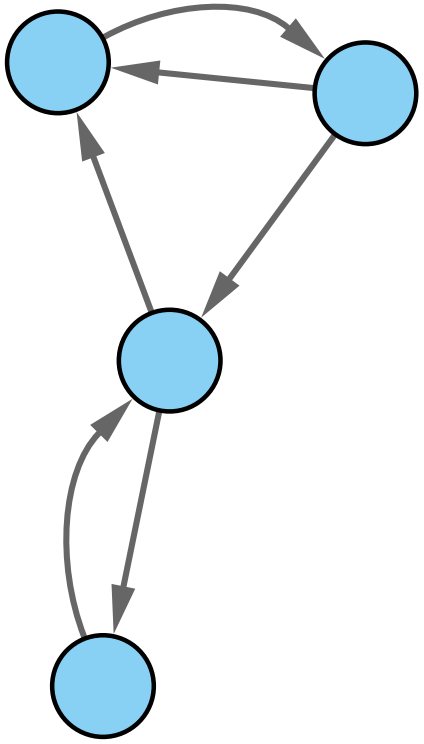
Defining a network: graphs



Multigraph:

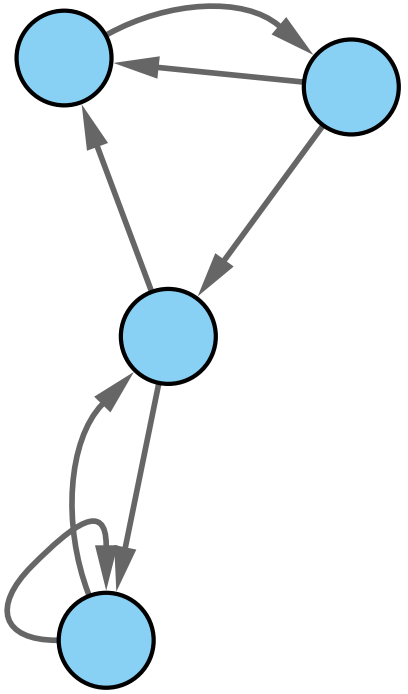
- Several, parallel edges

Defining a network: graphs



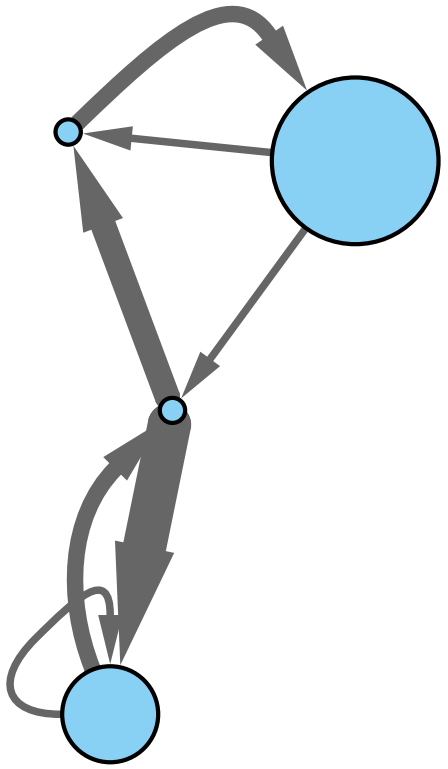
Directed graph

Defining a network: graphs



Self-edges

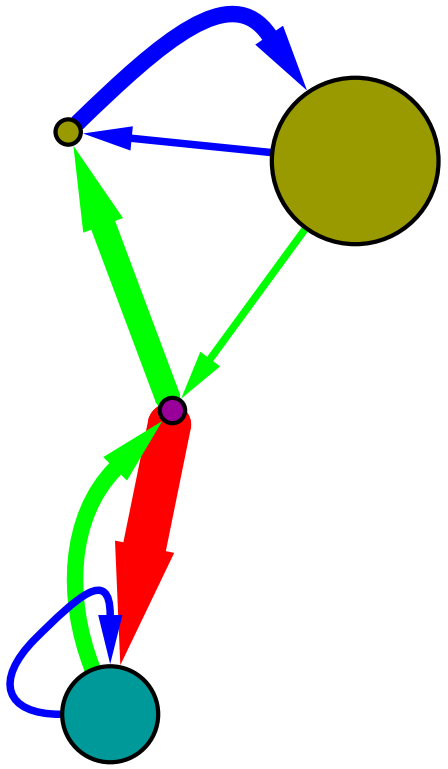
Defining a network: attributes



Additional data:

- Size / importance

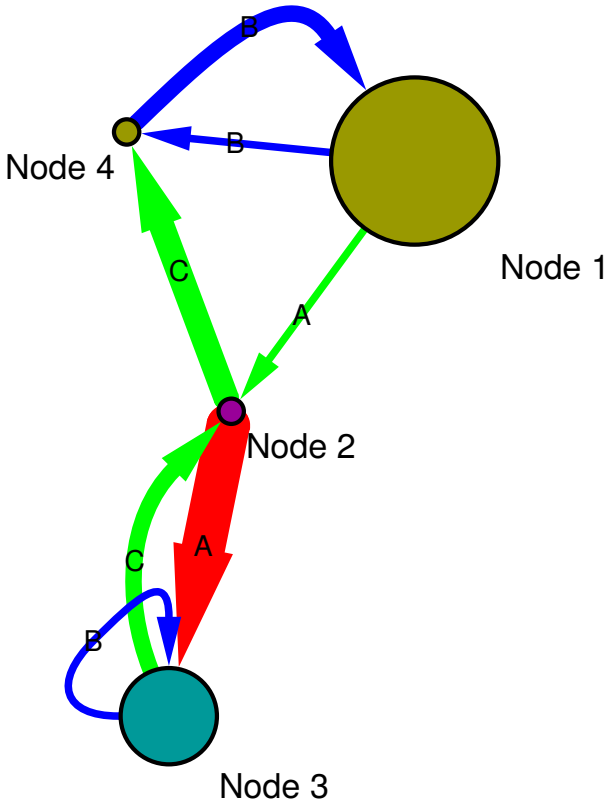
Defining a network: attributes



Extra data:

- Size / importance
- Type

Defining a network: attributes



Extra data:

- Size / importance
- Type
- Anything else

Bipartite network

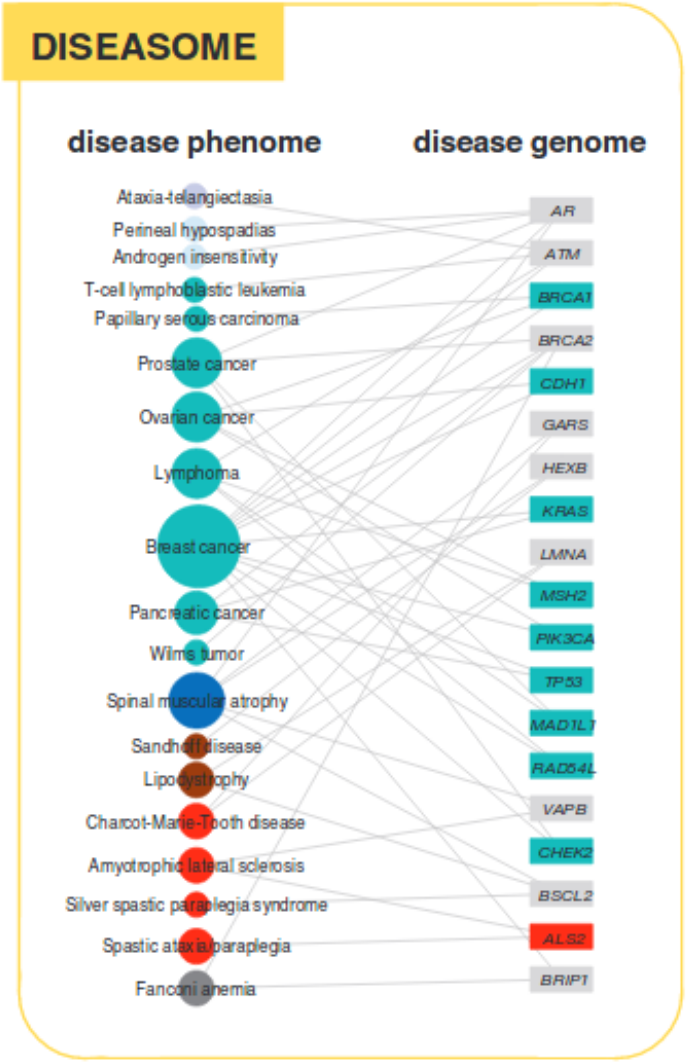


Image source: Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. PNAS (2007)

Bipartite network and its projections

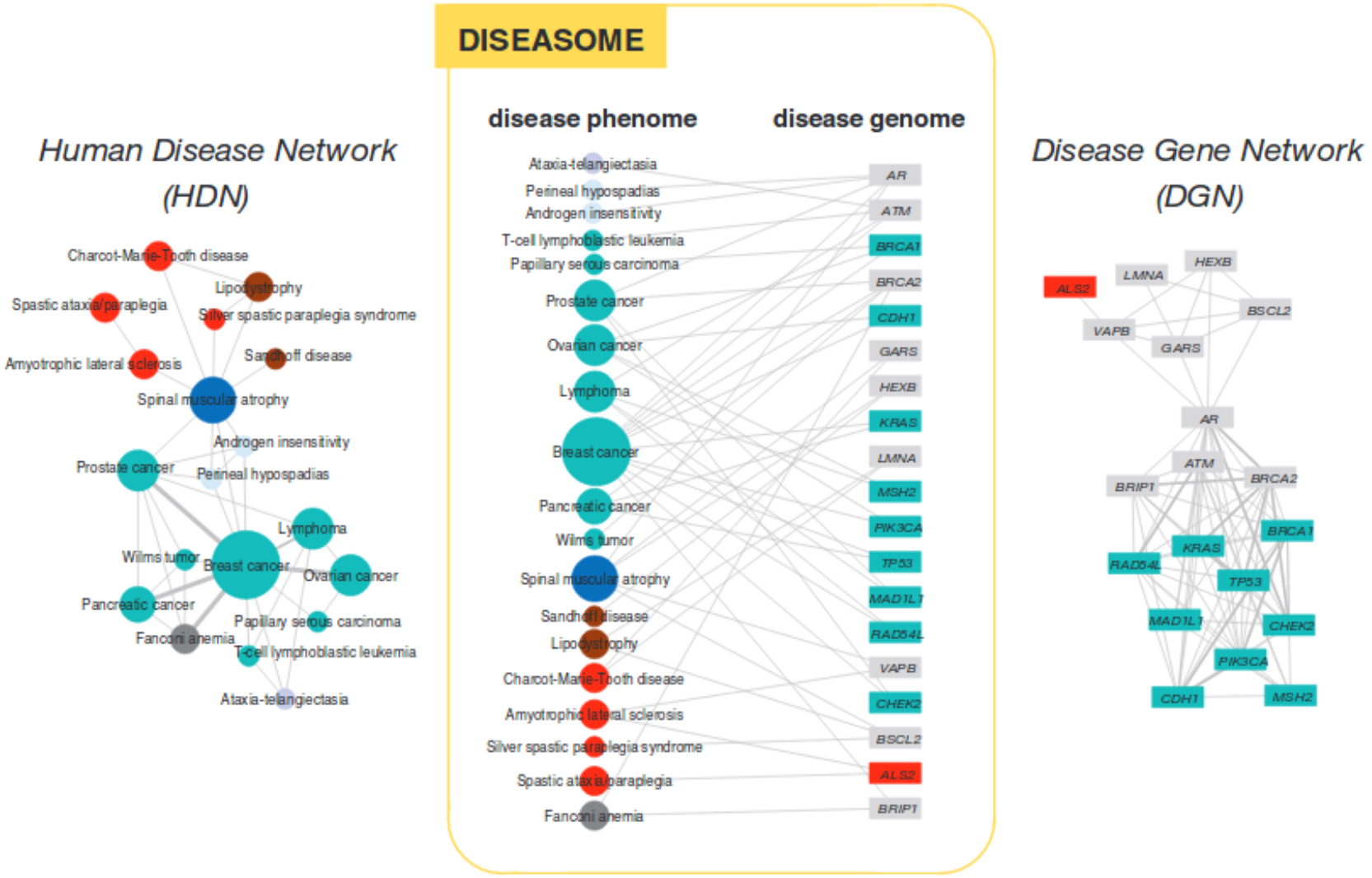
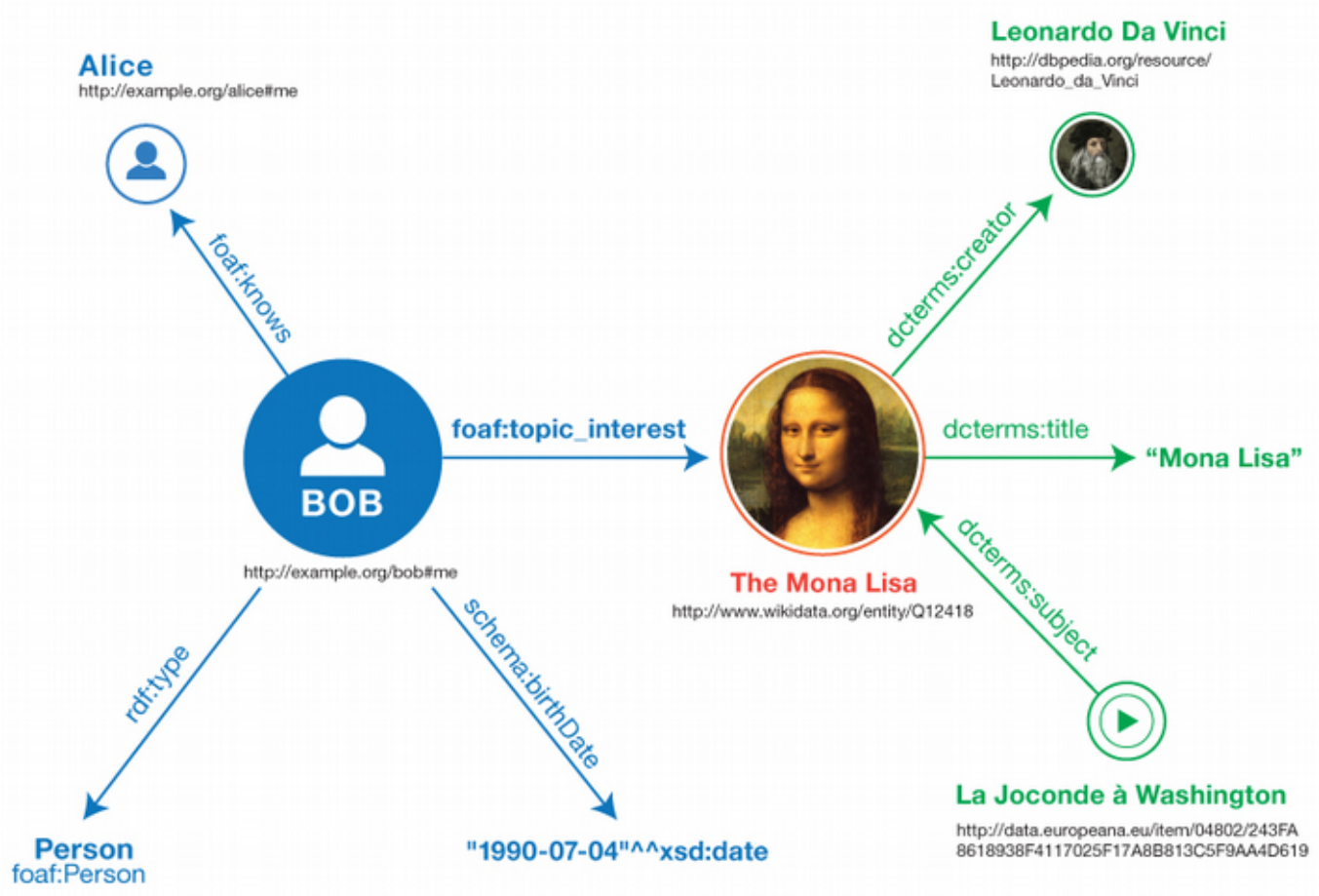
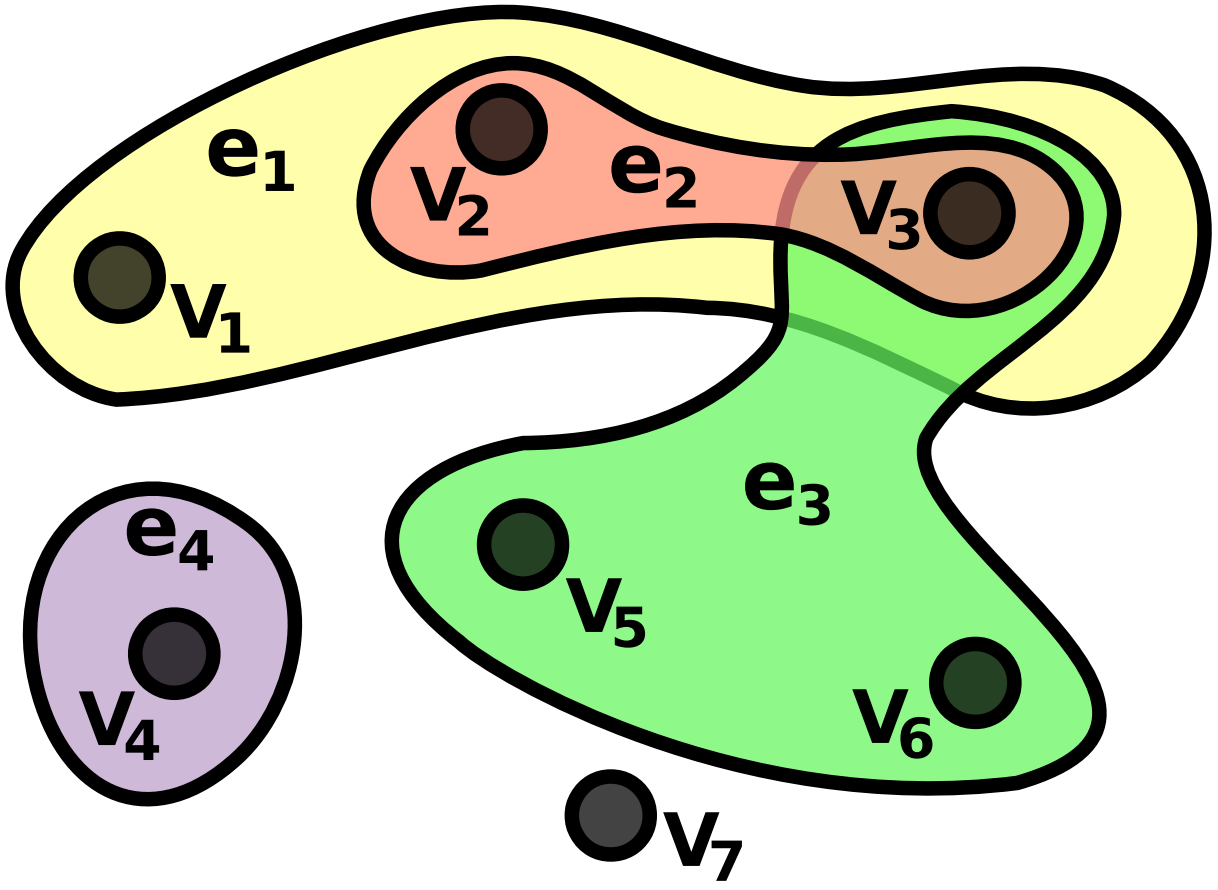


Image source: Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. PNAS (2007)

Tripartite network: RDF

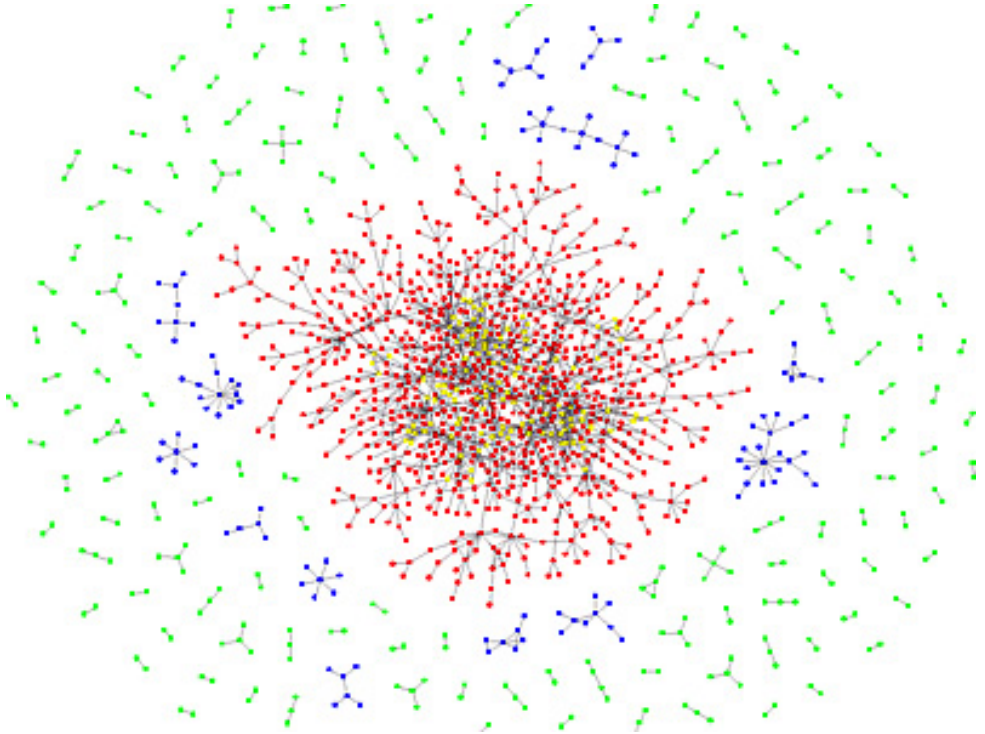


Hypergraph



Common basic properties

Components



- Isolated parts
- One large connected component

→ complications for:

- shortest paths
- diameter
- any quantity defined for a component

Image: an online community network, from:
http://www.visualcomplexity.com/vc/project_details.cfm?id=139&index=139&domain

Sparseness

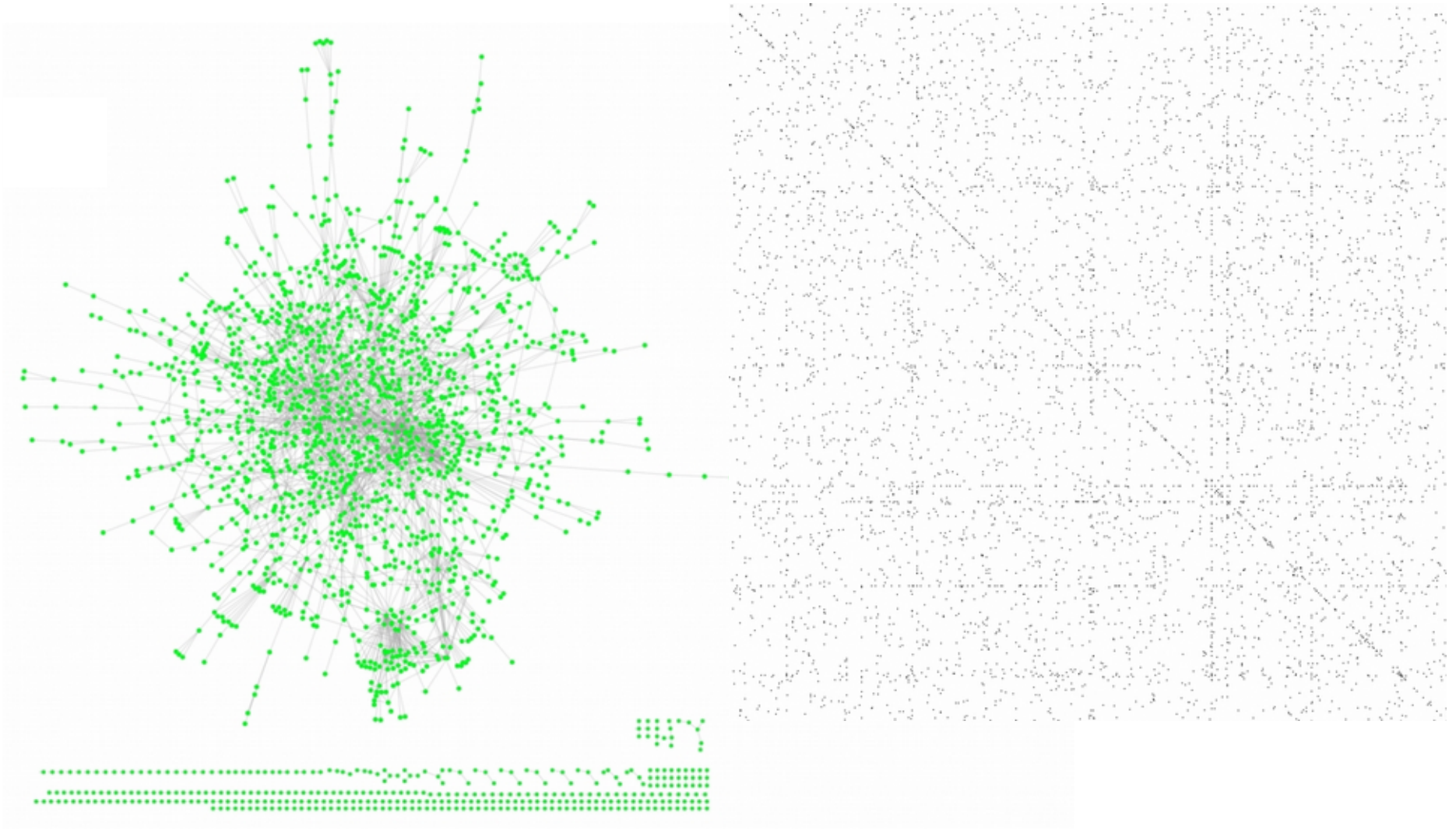
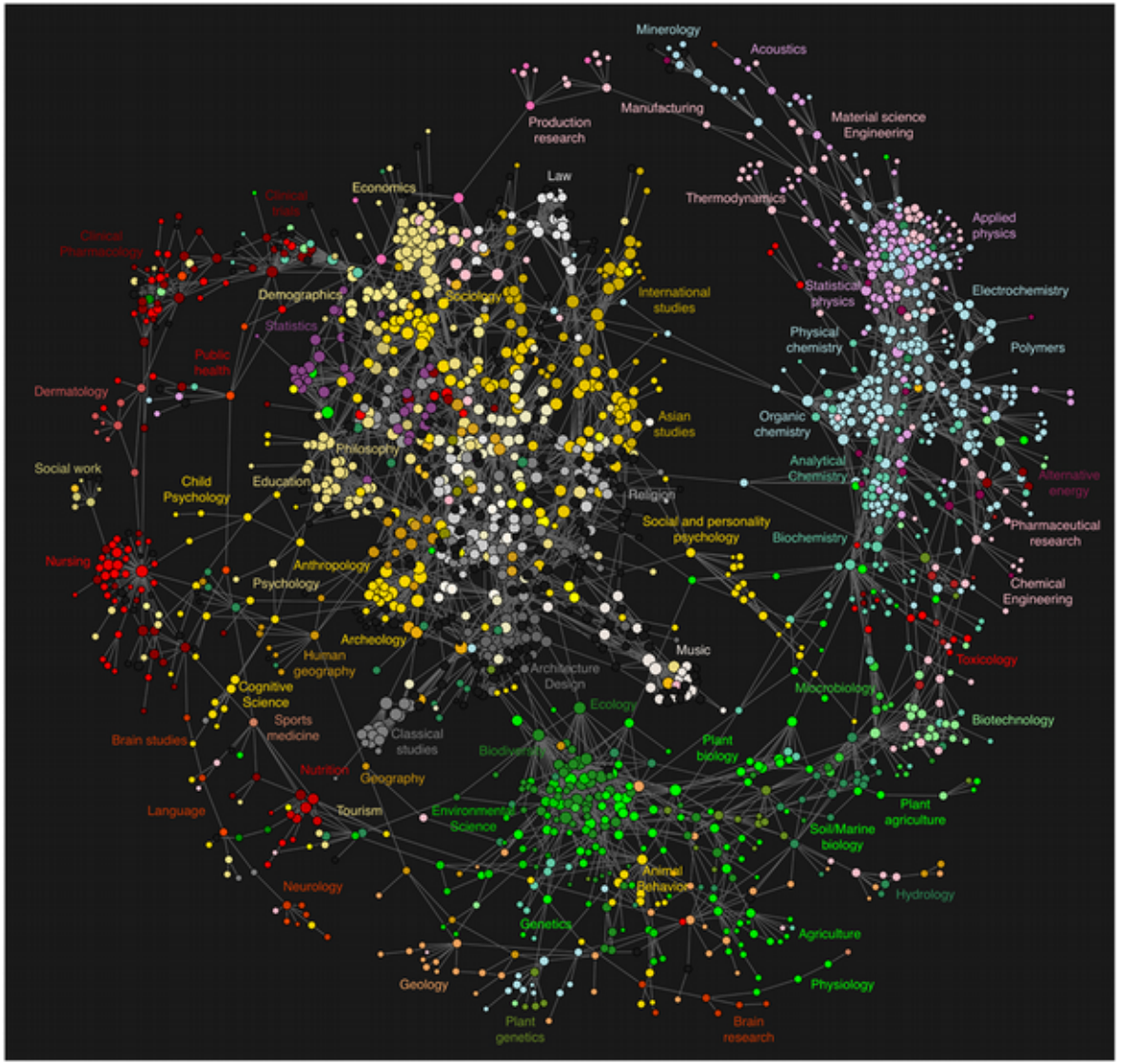


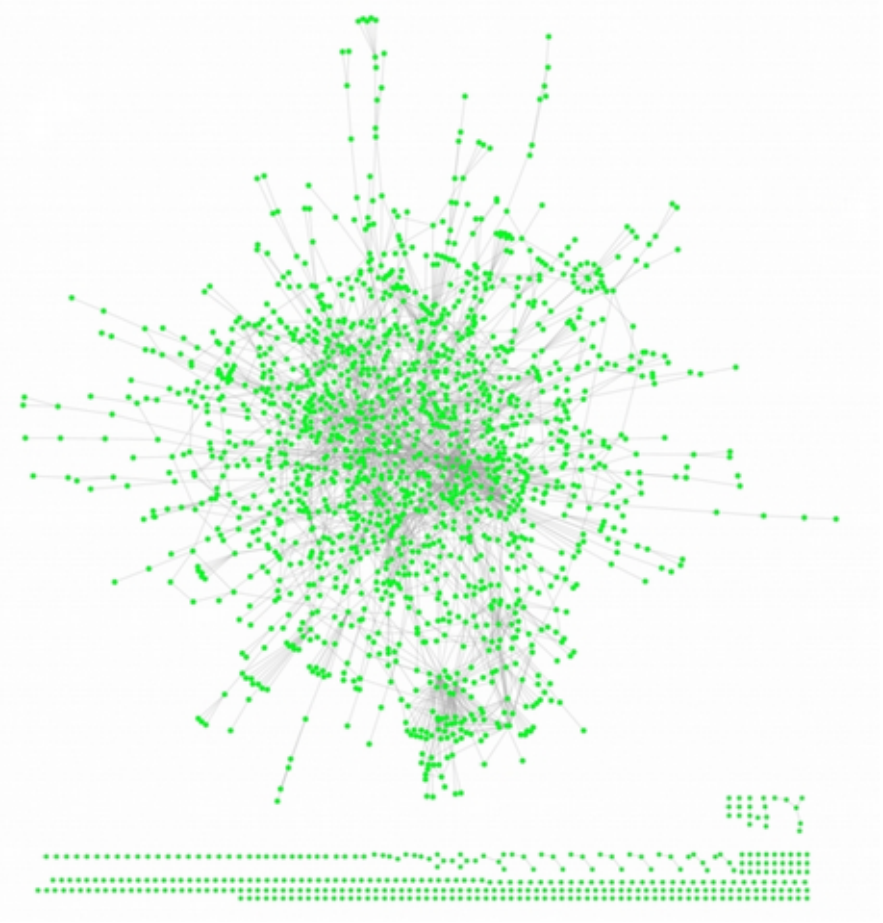
Image 2.4 and 2.7 from Barabasi's book

Globally sparse, locally dense



Structure of science based on readership (clickstream) data of scientific articles
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004803

Small world property



Compared to grid

advantages and disadvantages

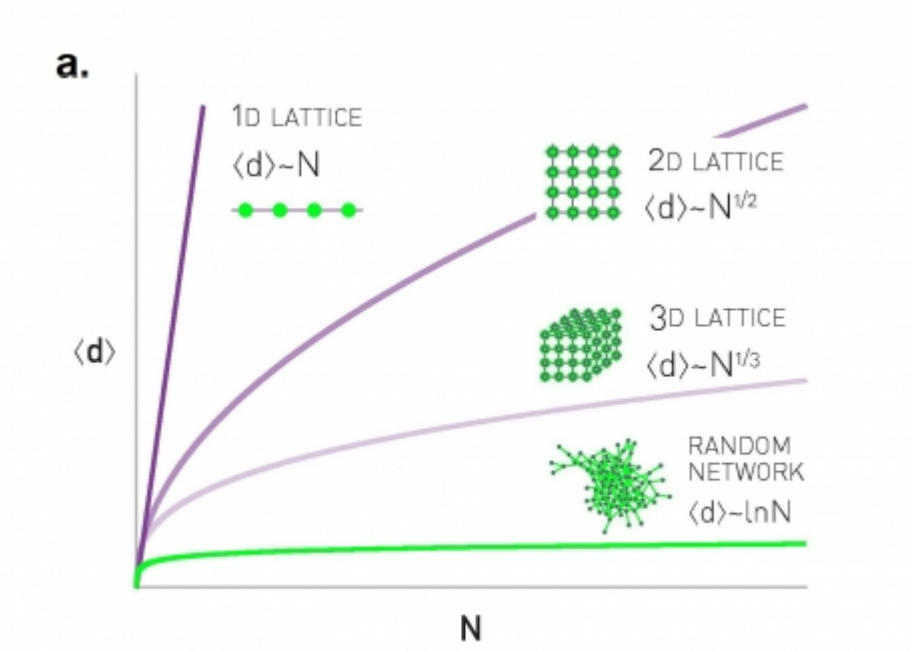


Image 3.11 from Barabasi's book

Degree distribution

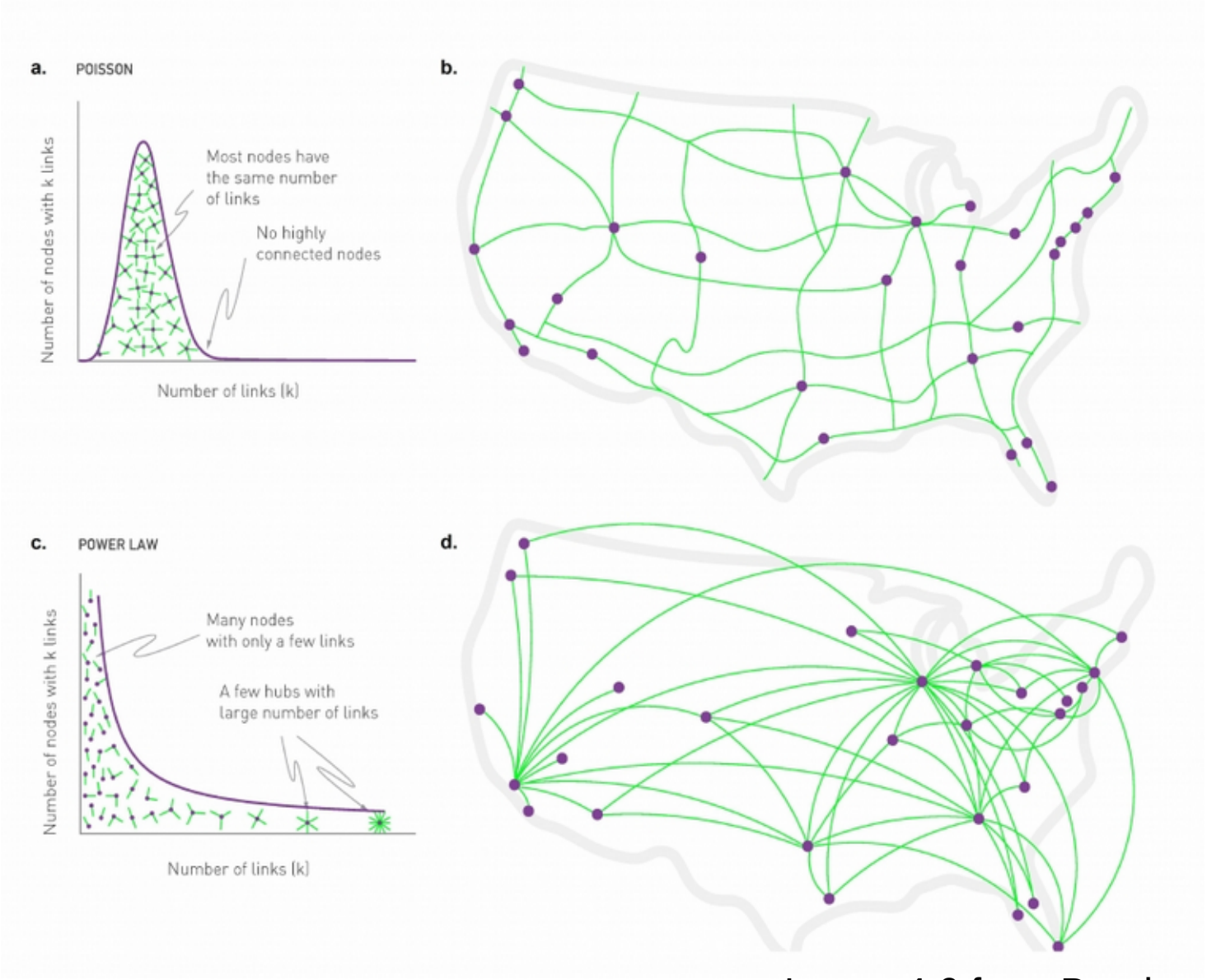
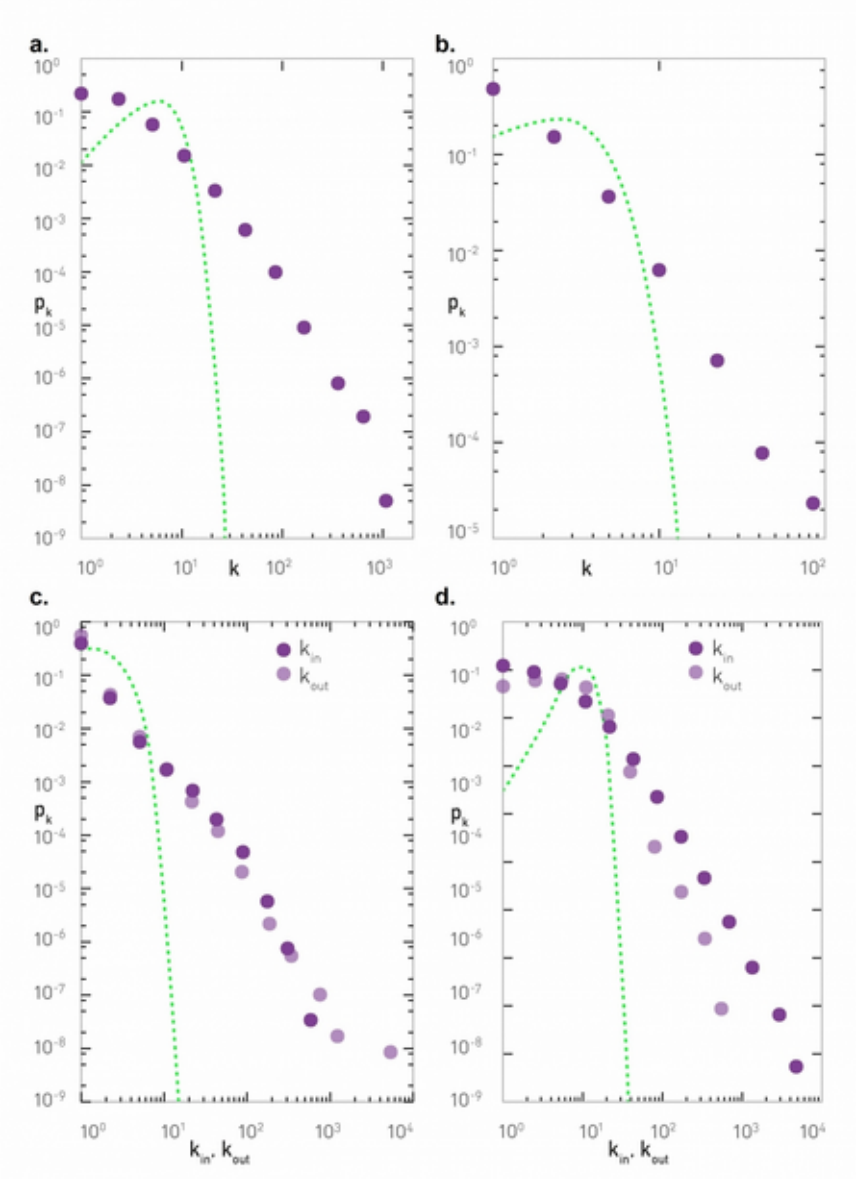


Image 4.6 from Barabasi's book

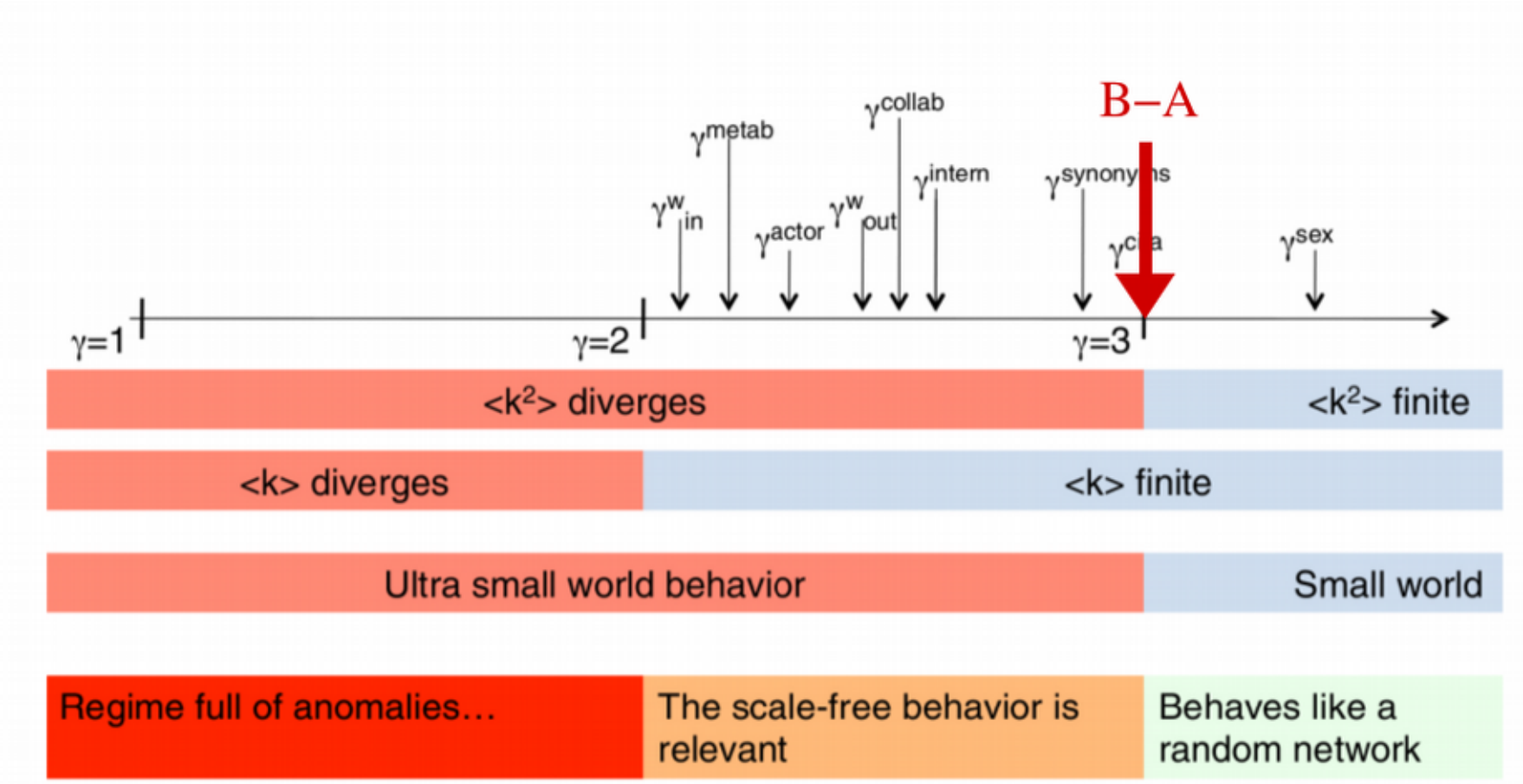
Degree distribution



- Four networks:
  - Internet routers
  - Protein interactions
  - Email network
  - Citation network of scientific articles
- Purple dots: network
- Green line: Poisson-distribution with same <k>

Image 4.10 from Barabasi's book

Consequences of the distribution



From Barabási's slides

A warning

- Power law
- Scale-free
- Fat-tailed

	name	distribution $p(x) = Cf(x)$ $f(x)$	C
continuous	power law	$x^{-\alpha}$	$(\alpha-1)x_{\min}^{\alpha-1}$
	power law with cutoff	$x^{-\alpha}e^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$
	exponential	$e^{-\lambda x}$	$\lambda e^{\lambda x_{\min}}$
	stretched exponential	$x^{\beta-1}e^{-\lambda x^{\beta}}$	$\beta\lambda e^{\lambda x_{\min}^{\beta}}$
	log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\operatorname{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$
discrete	power law	$x^{-\alpha}$	$1/\zeta(\alpha, x_{\min})$
	Yule distribution	$\frac{\Gamma(x)}{\Gamma(x+\alpha)}$	$(\alpha-1)\frac{\Gamma(x_{\min}+\alpha-1)}{\Gamma(x_{\min})}$
	exponential	$e^{-\lambda x}$	$(1-e^{-\lambda})e^{\lambda x_{\min}}$
	Poisson	$\mu^x/x!$	$\left[e^{\mu}-\sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$

TABLE 2.1

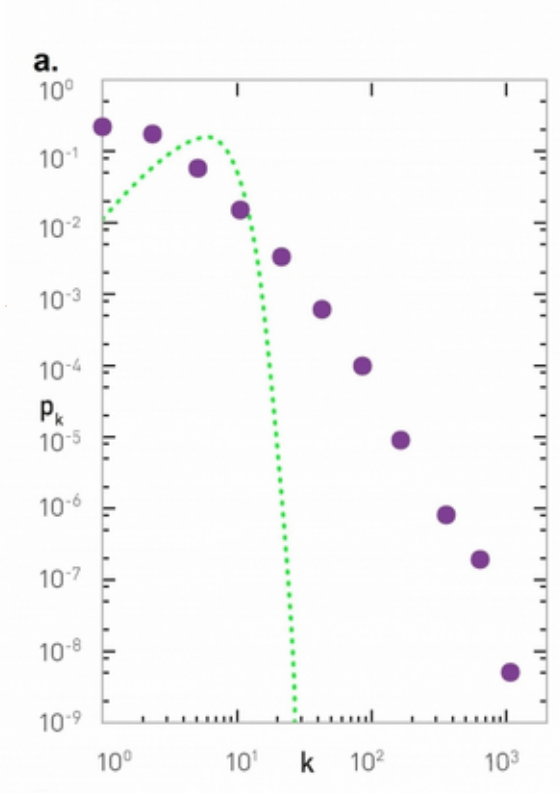


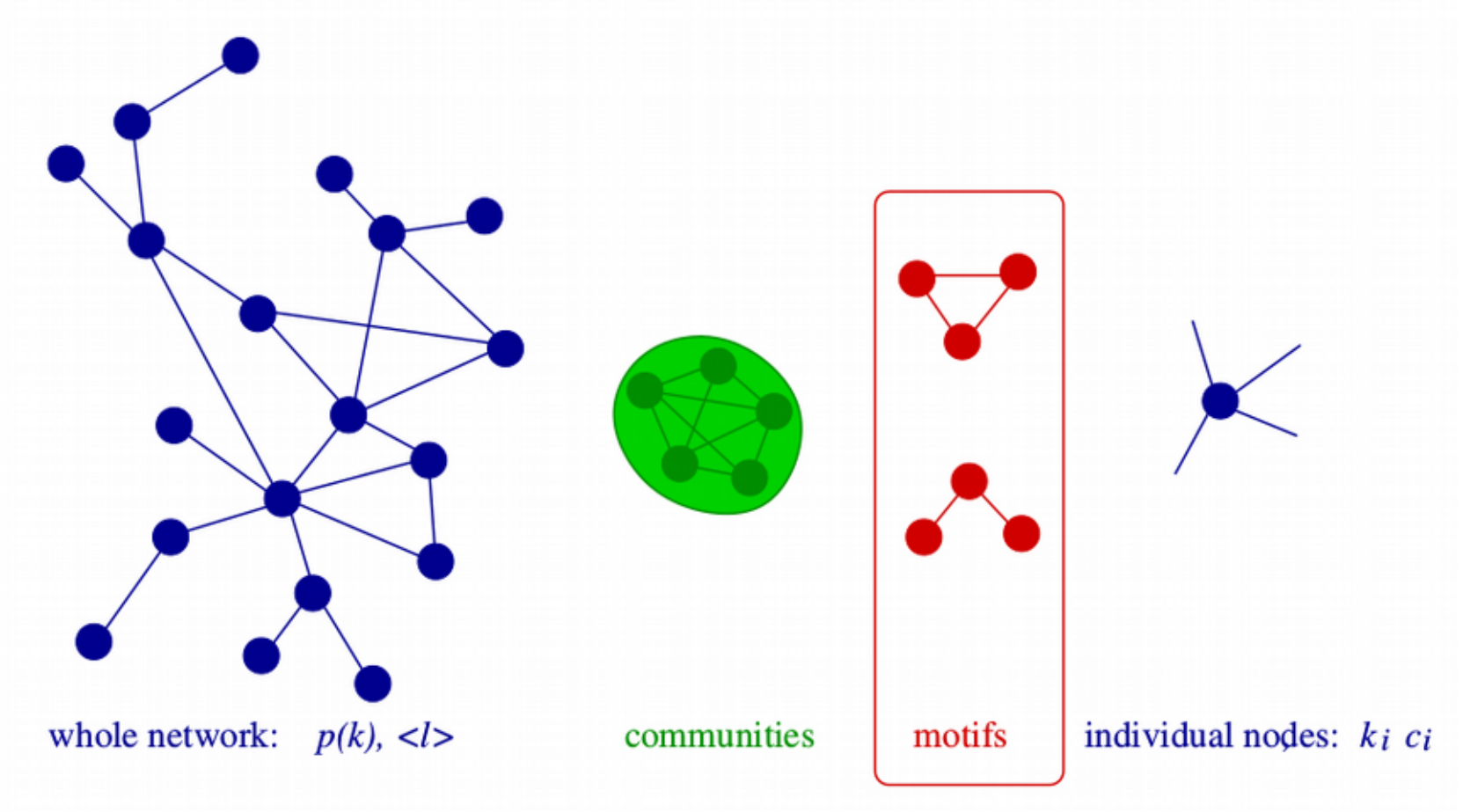
Table from:
http://tuvalu.santafe.edu/~aaronc/powerlaws/

What wasn’t visible on these slides

- Dirty data, artifacts
- “you get what you pay for”
- Network analysis is more sensitive to data cleaning than other methods
- But: for data analysis, less of a difference between “artifact” and “interesting phenomena”

Quantities

Scales

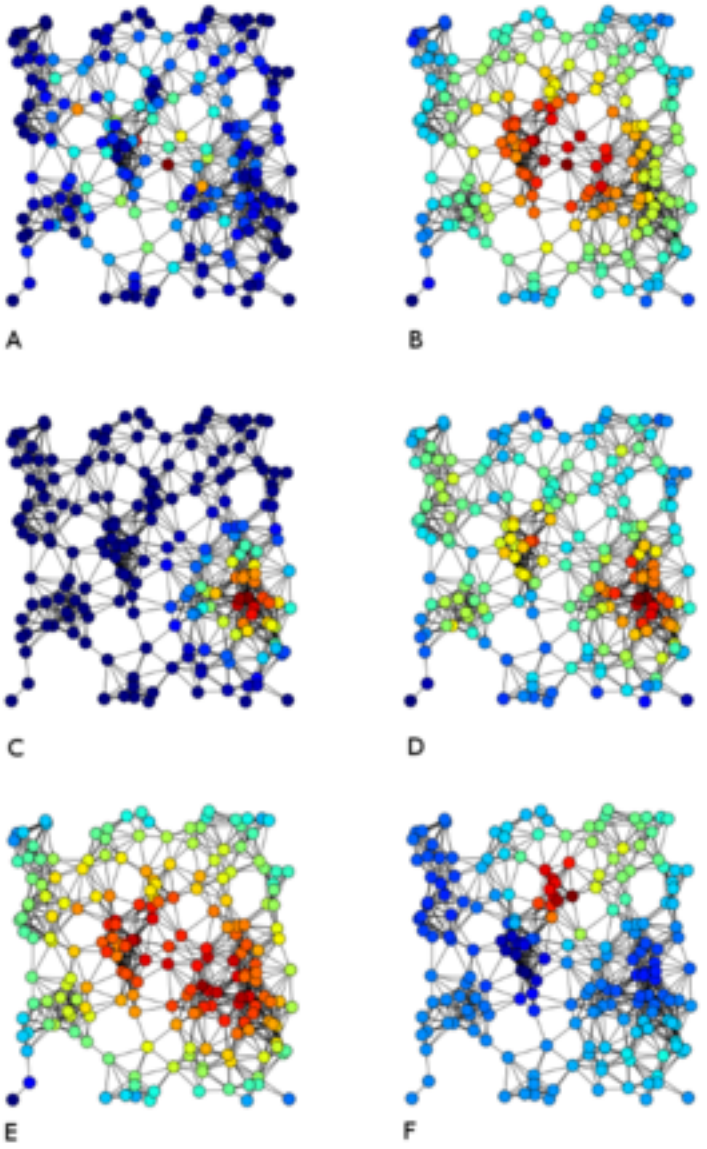

whole network: p(k), <l>
communities
motifs
individual nodes: k_i c_i

Quantities

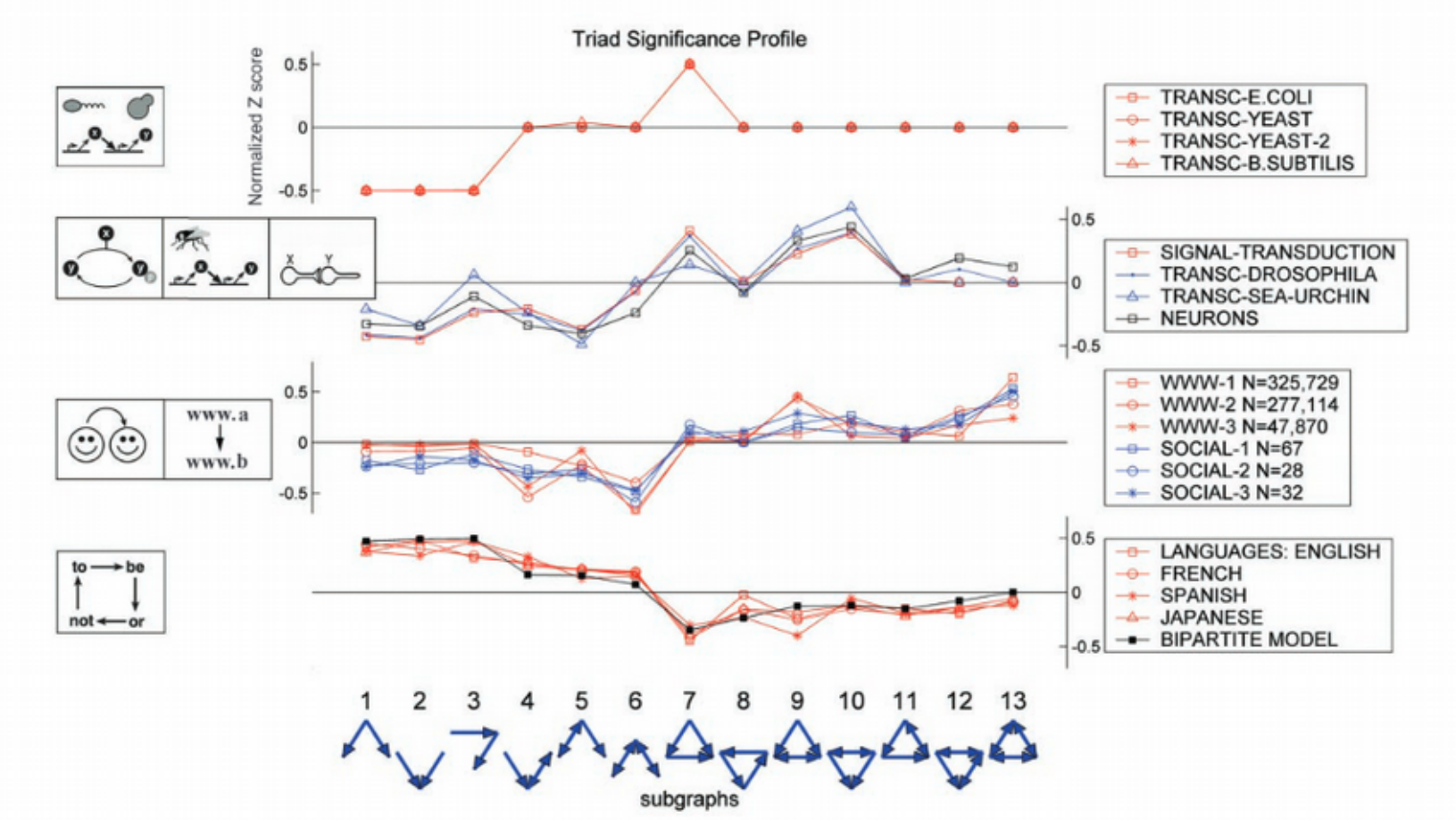
- Degree – directed (in- & out-), weighted (strength)
- Edges → Degree correlation (assortativity)
- Triangles → clustering coefficient
- Shortest paths → betweeness
  - Geodetic vs. random paths
- Whole network – degree, distribution

Various centrality measures

- A) Betweenness centrality
- B) Closeness centrality
- C) Eigenvector centrality
- D) Degree centrality
- E) Harmonic centrality
- F) Katz centrality



Motifs



Milo, R; Itzkovitz, S; Kashtan, N; Levitt, R; Shen-Orr, S; Ayzenshtat, I; Sheffer, M; Alon, U
Superfamilies of Evolved and Designed Networks
Science , (2004)

Random models

Models in general

- Qualitative model: what are the fundamental mechanisms?
- Quantitative model: act as reference

Erdős-Rényi model

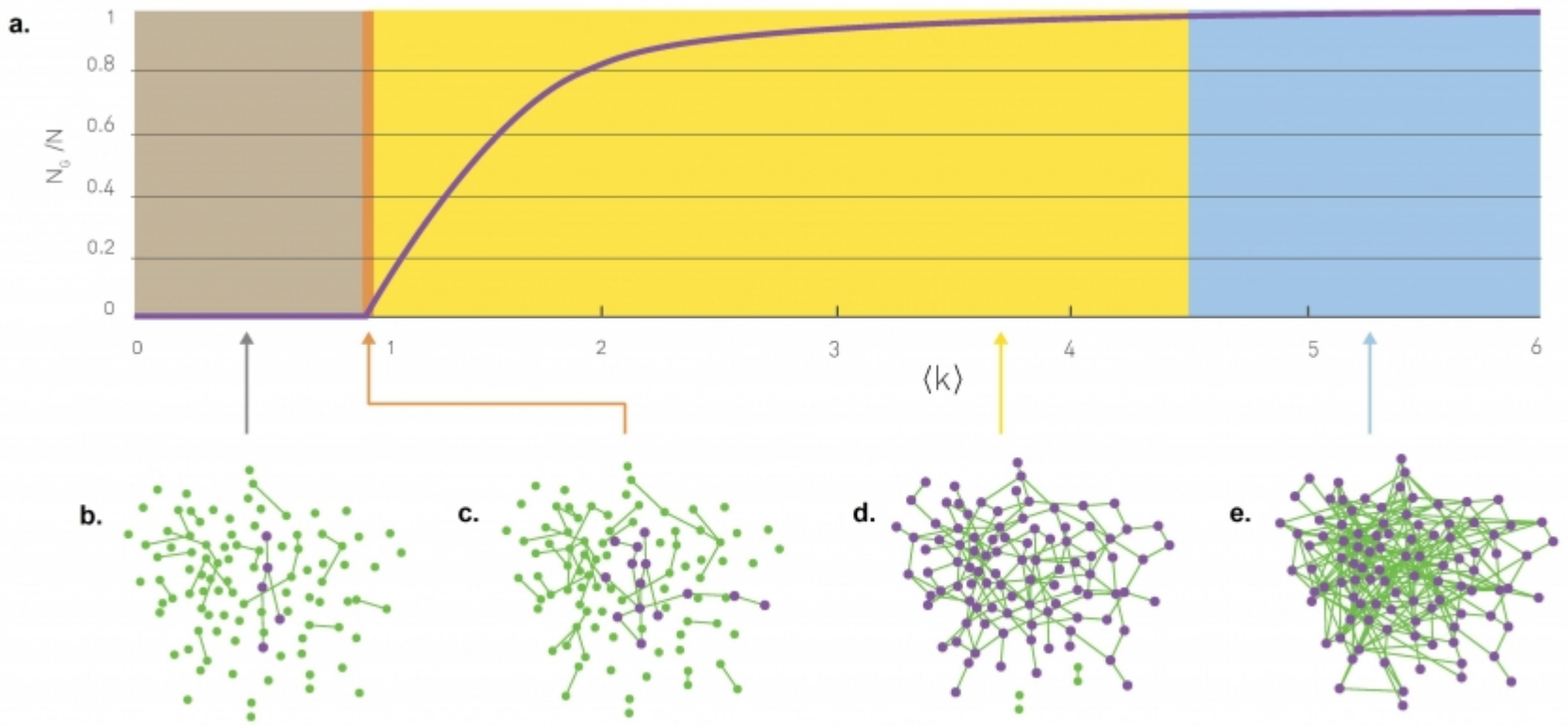
- N nodes and:
  - M edges

    or:
  - Every pair of nodes connected with p probability

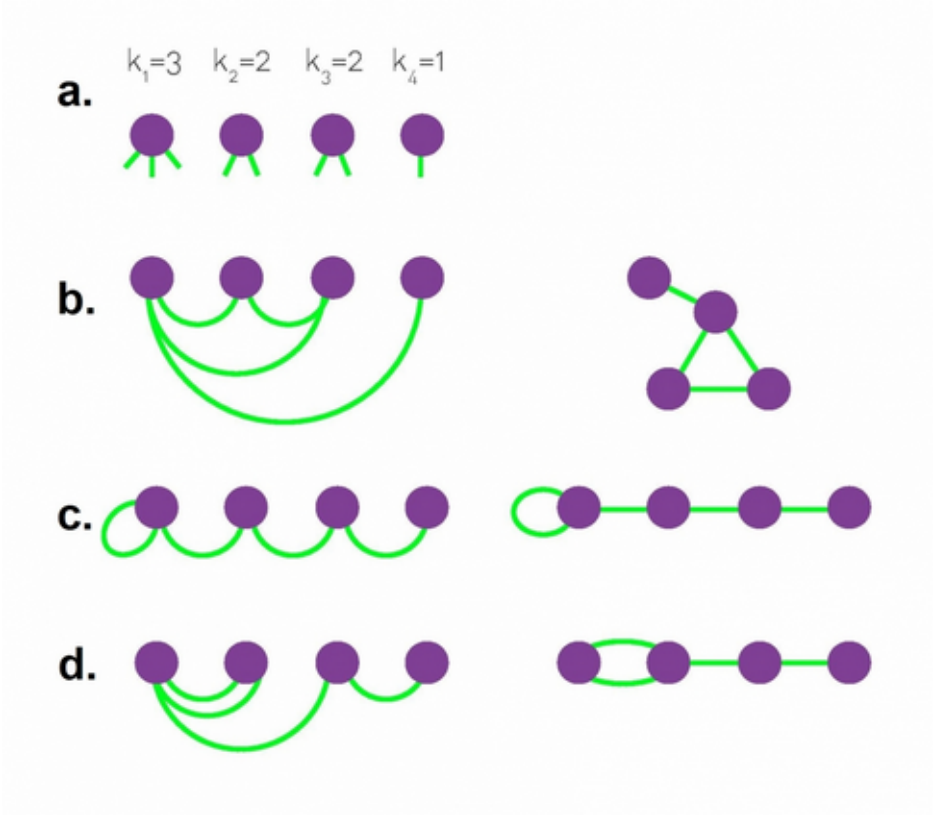
N=100, p=0.03 – Image 3.3 from Barabási's book

Erdős-Rényi model

- If adding edges one by one: percolation



Configuration model: given degrees

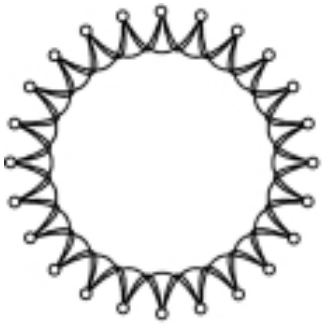


- Nodes & edge ends
- Connecting edge ends
- (not all sequences suitable)

Image 4.15 from Barabási's book

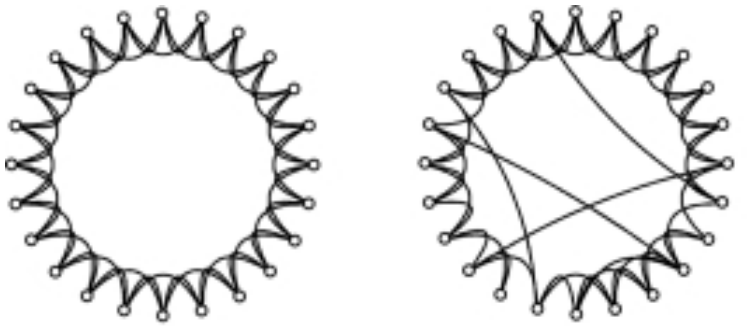
Watts-Strogatz – small world

- Ring, close neighbors connected



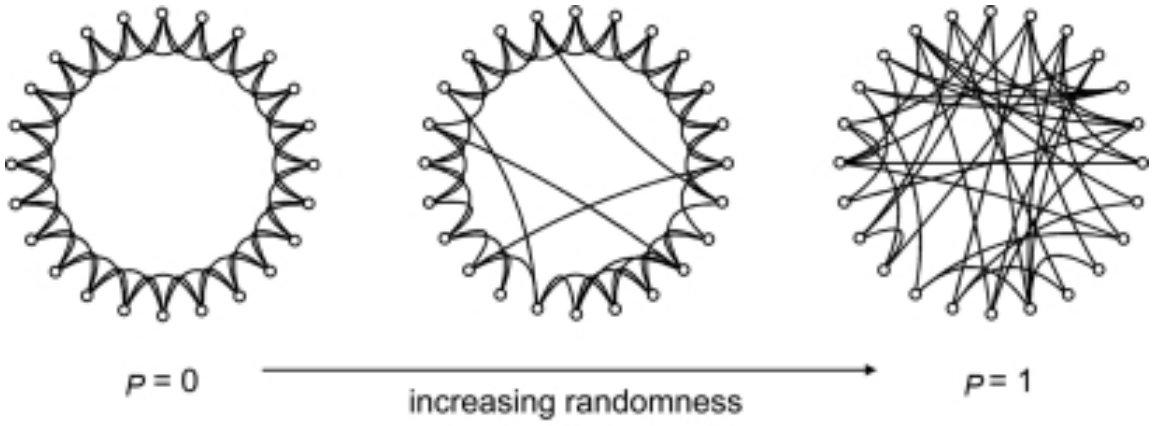
Watts-Strogatz – small world

- Ring, close neighbors connected
- Randomly move edges



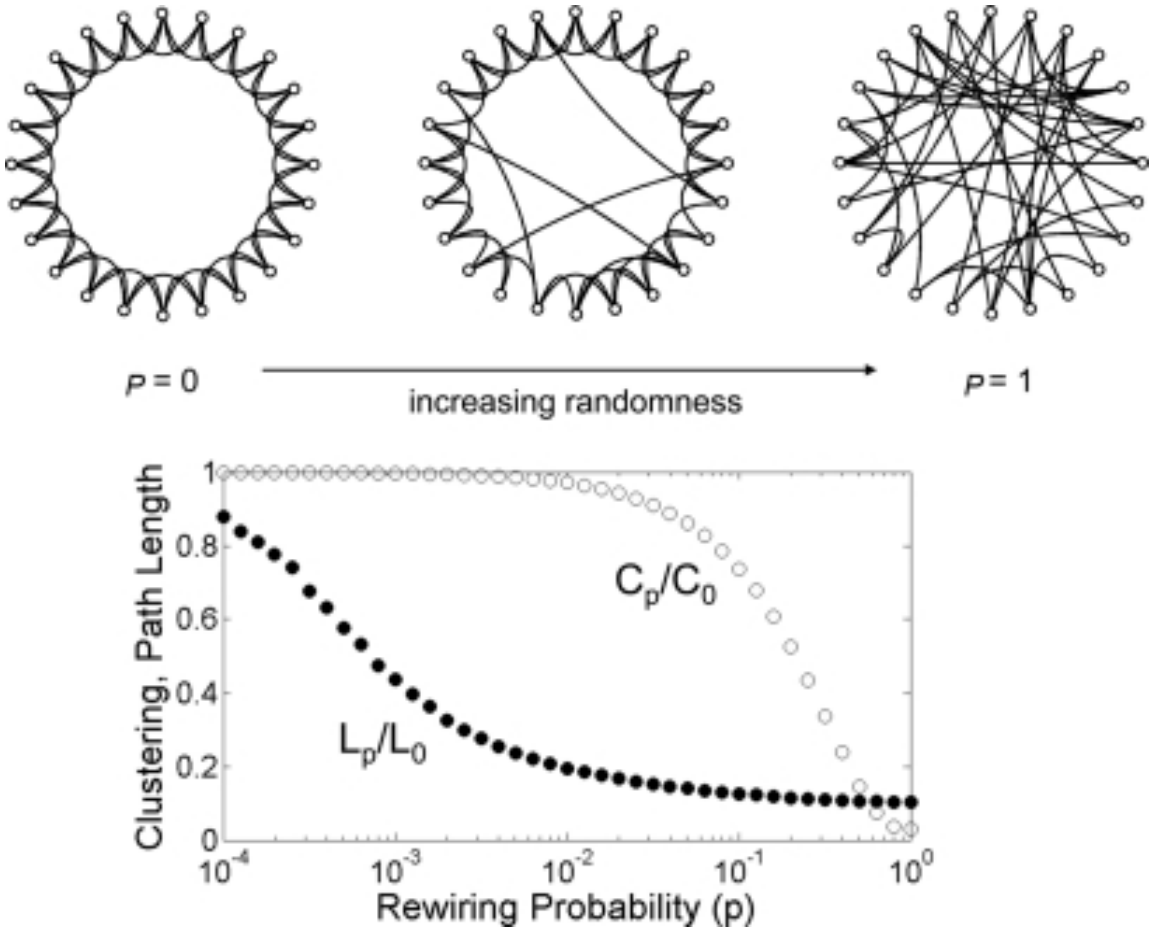
Watts-Strogatz – small world

- Ring, close neighbors connected
- Randomly move edges
- At end: essentially Erdős-Rényi network



Watts-Strogatz – small world

- Ring, close neighbors connected
- Randomly move edges
- At end: essentially Erdős-Rényi network
- Interim state: high clustering, low diameter



Barabási-Albert model

Growth process– preferential attachment

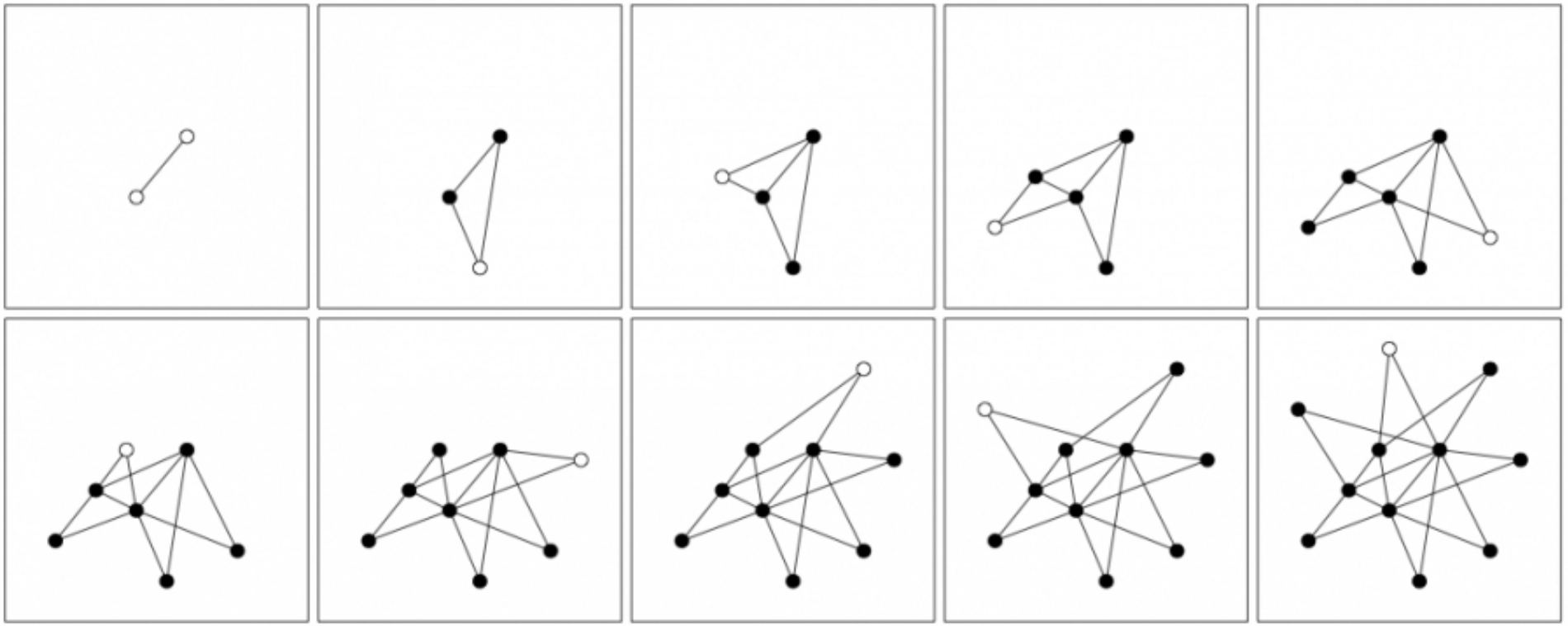


Image 5.3 from Barabási's book

Random models – how to use them?

- Consider parameters
- Use as large network as possible
- Average over multiple runs

  (but one might be enough)

“create from scratch”

vs.

“randomizing existing data”

Clustering / Community finding

- From “globally sparse, locally dense” property → find dense sub-sets of nodes
- Modules / Clusters / Communities
  - Different names for the same general idea
  - (note: “clustering” in this case has nothing to do with “clustering coefficient”!)
- Very under-defined problem, hard to measure accuracy
  - Everyone will have their favorite definition / method
- Very helpful for visualization
  - Can use for high-level overview
- Many methods from other disciplines:
  - Data mining, unsupervised learning, statistics, etc.

Varieties of definitions of groups

- Disjunct: each node placed in only one group
- Fuzzy: each node is in every group, but with different weights
- Overlapping: each node might be in several groups
- Hierarchical: groups defined at various levels

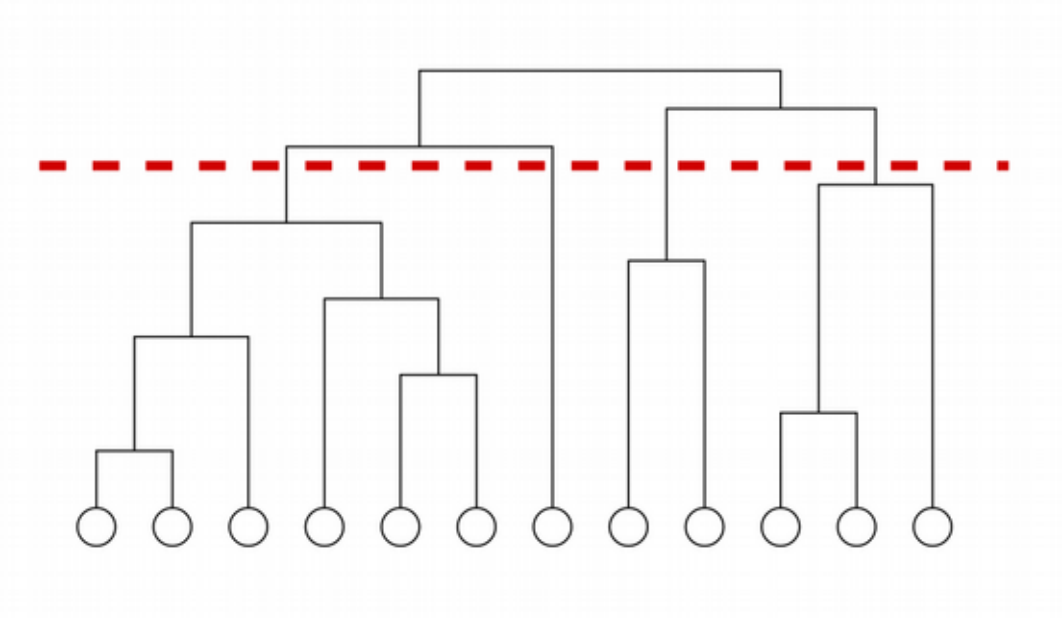
And also: is every node placed in a group, or might some be left out?

Agglomerative / divisive methods

- Agglomerative: start with isolated nodes, consider each node a separate community, at each step join two communities
- Divisive: start with one single community, at each step divide one community into two
- Easy to define greedy algorithms
- An example for divisive method: (Girvan-Newman algorithm) cut edges by always removing the edge with the highest betweenness
- Note: need separate criteria for when to stop

Dendrogram

- Steps of joining / separating communities
- Individual nodes at the bottom
- Final structure: horizontal cut at determined height



Optimizing a function

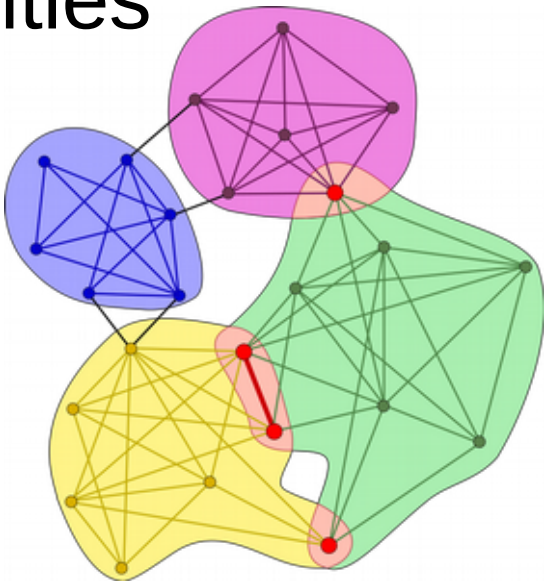
- If a single score can be defined for the goodness of the communities, we can optimize that
- Most popular variant: Modularity:

  Compares number of edges within / between communities to those expected based on configuration model

$$Q = \frac{1}{(2m)} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w) = \sum_{i=1}^{c} (e_{ij} - a_i^2)$$

Clique percolation method

- Use k-clique (fully connected network of k nodes)
- Community is a part of network where clique can be moved to by moving one node at a time (“rolling the clique”)
- Allows distinct, overlapping communities



Software & tools

Cytoscape

- GUI program, point & click interface
  - Introductory tutorials: https://github.com/cytoscape/cytoscape-tutorials/wiki
- Very good for assembling visualizations
- Large number of plugins: http://apps.cytoscape.org/
- Drawbacks:
  - Gui program, so scripting is problematic
  - Performance limitations for large networks

networkx

- https://networkx.github.io/
- Very user-friendly library for python
- Many algorithms implemented
- Drawbacks:
  - Configuring visualization more fiddly than with cytoscape (gui style editor vs. writing python code)
  - Not performance oriented, other libs will work better for huge networks

Other tools

- igraph – http://igraph.org/
  - C library, R, C++ and python wrapper
  - Harder to use than networkx, but more optimized for performance
- Gephi – http://gephi.org
  - Very similar to cytoscape (Java desktop app, plugin-system)
- Neo4j – http://neo4j.org
  - Graph database, not “graph compute engine”
  - Not needed for small data
- Distributed parallel computation systems
  - Graphx (https://spark.apache.org/graphx/)
  - Giraph (http://giraph.apache.org/)
  - Etc.

Online materials

- Barabási's book (also in print, also in hungarian):
  - http://barabasi.com/networksciencebook/
- Online courses:
  - https://github.com/ladamalina/coursera-sna
  - https://www.coursera.org/learn/python-social-network-analysis
- Wikipedia