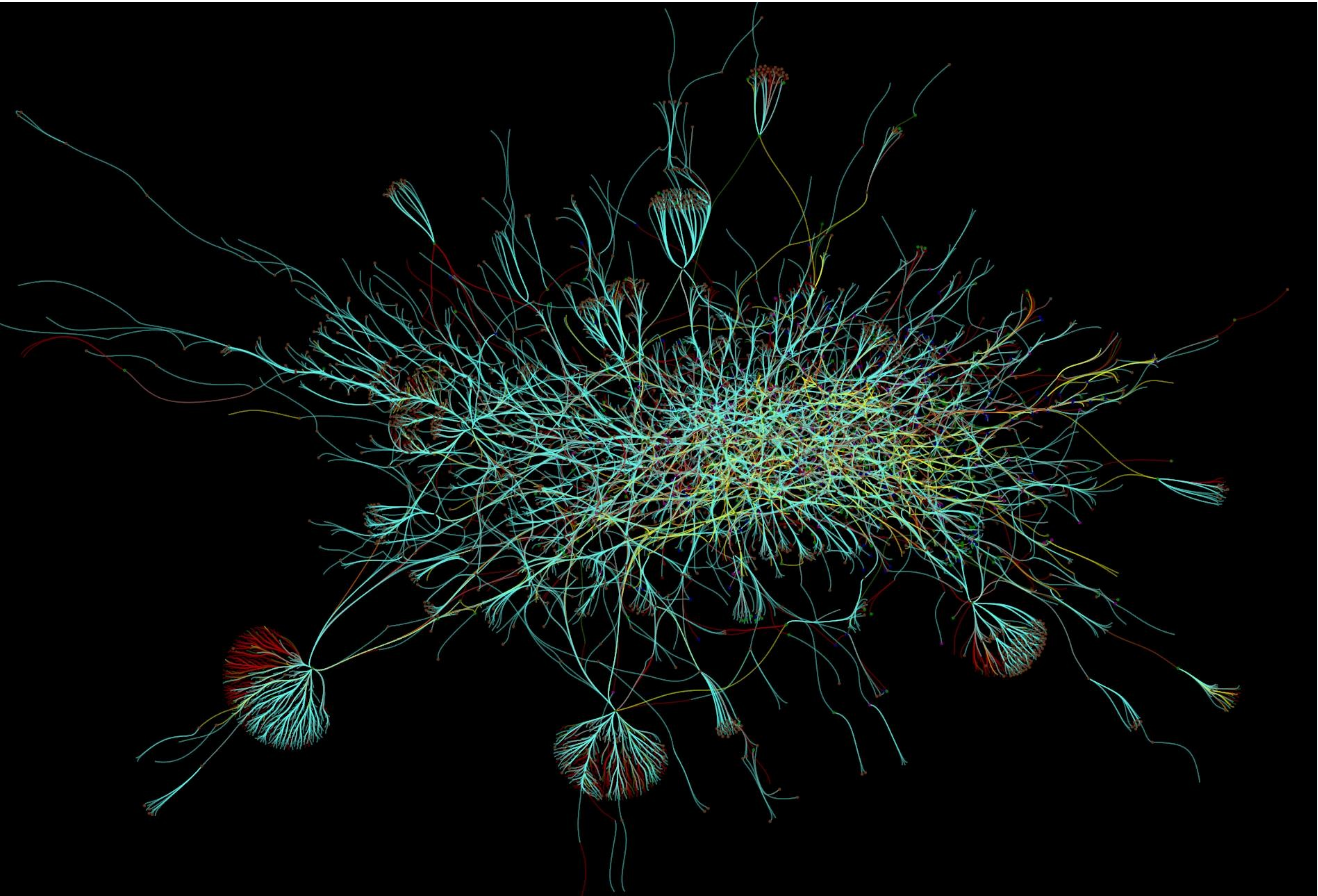
Networks



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

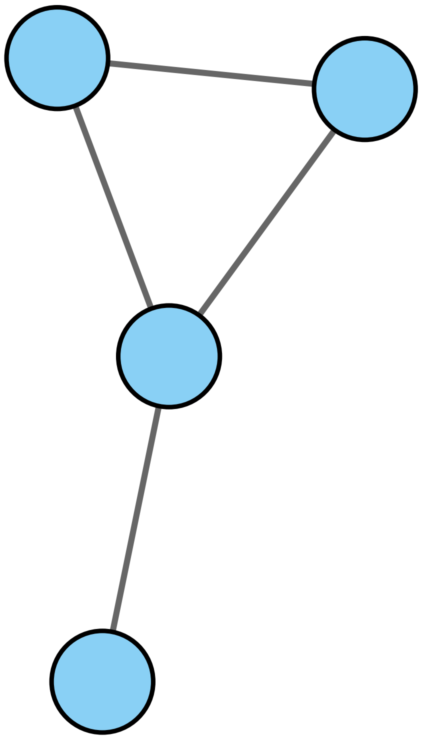
Graph theory vs. network science vs. systems science vs. ….

- 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”

Defining a network

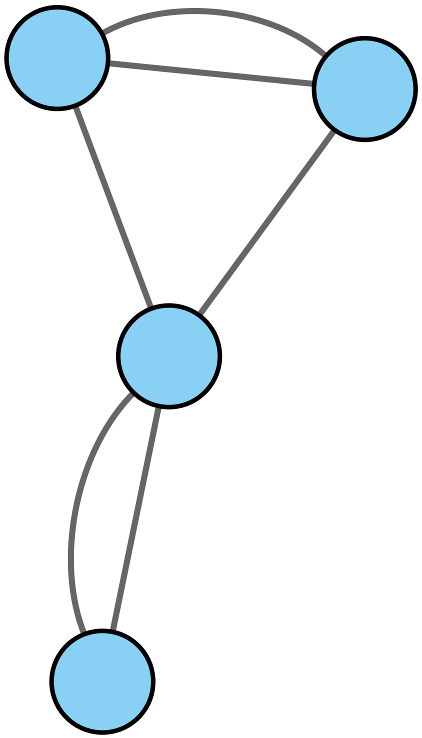
Defining a network: graphs



Simple graph:

- Nodes
- Edges

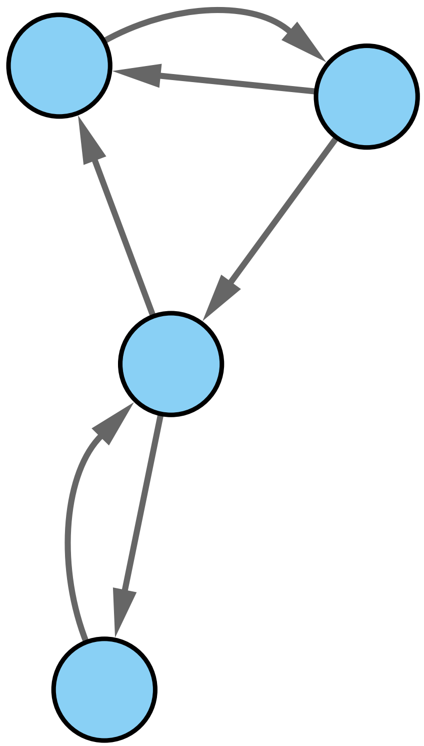
Defining a network: graphs



Multigraph:

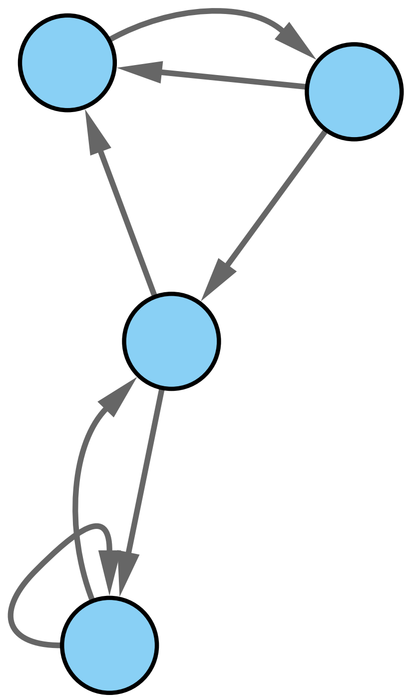
– Several, parallel edges

Defining a network: graphs



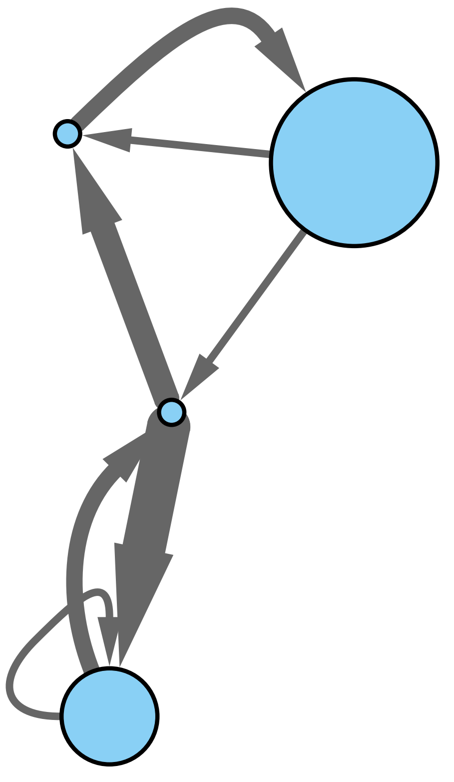
Directed graph

Defining a network: graphs



Self-edges

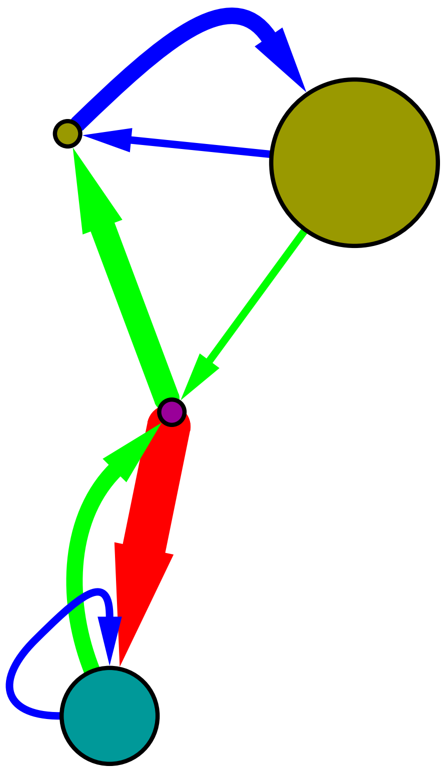
Defining a network: attributes



Additional data:

– Size / importance

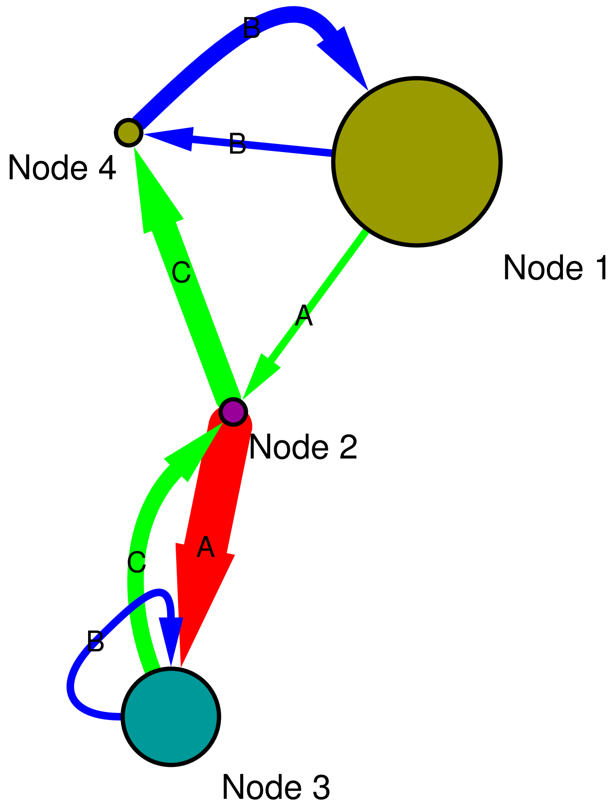
Defining a network: attributes



Extra data:

- Size / importance
- Type

Defining a network: attributes



Extra data:

- Size / importance
- Type
- Anything else

Bipartite network

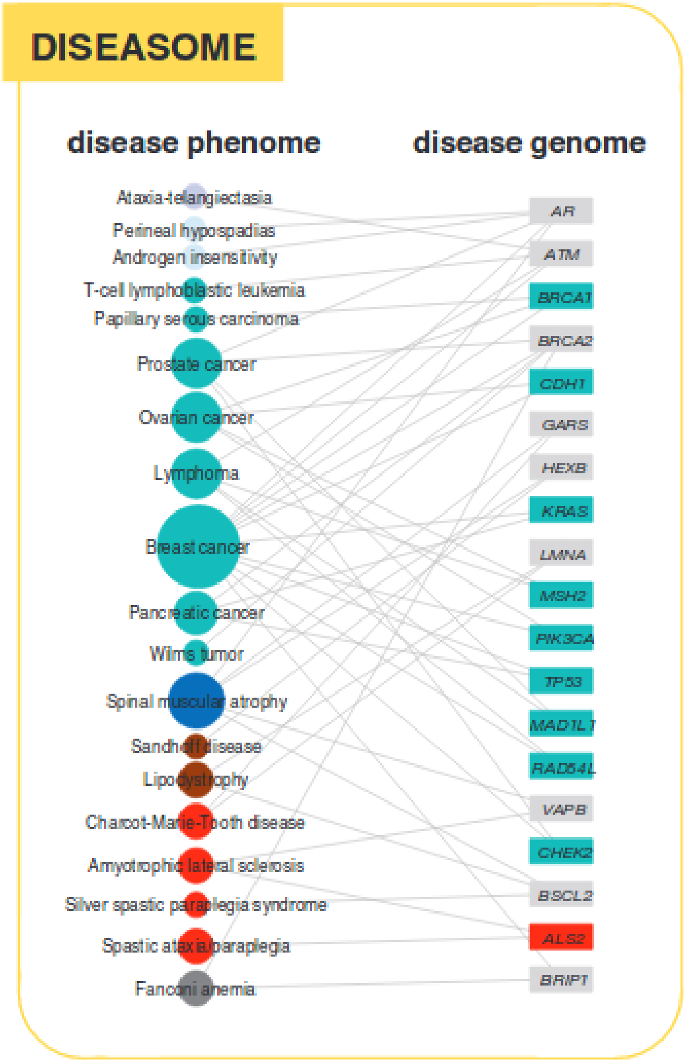


Image source: Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. PNAS (2007)

Bipartite network and its projections

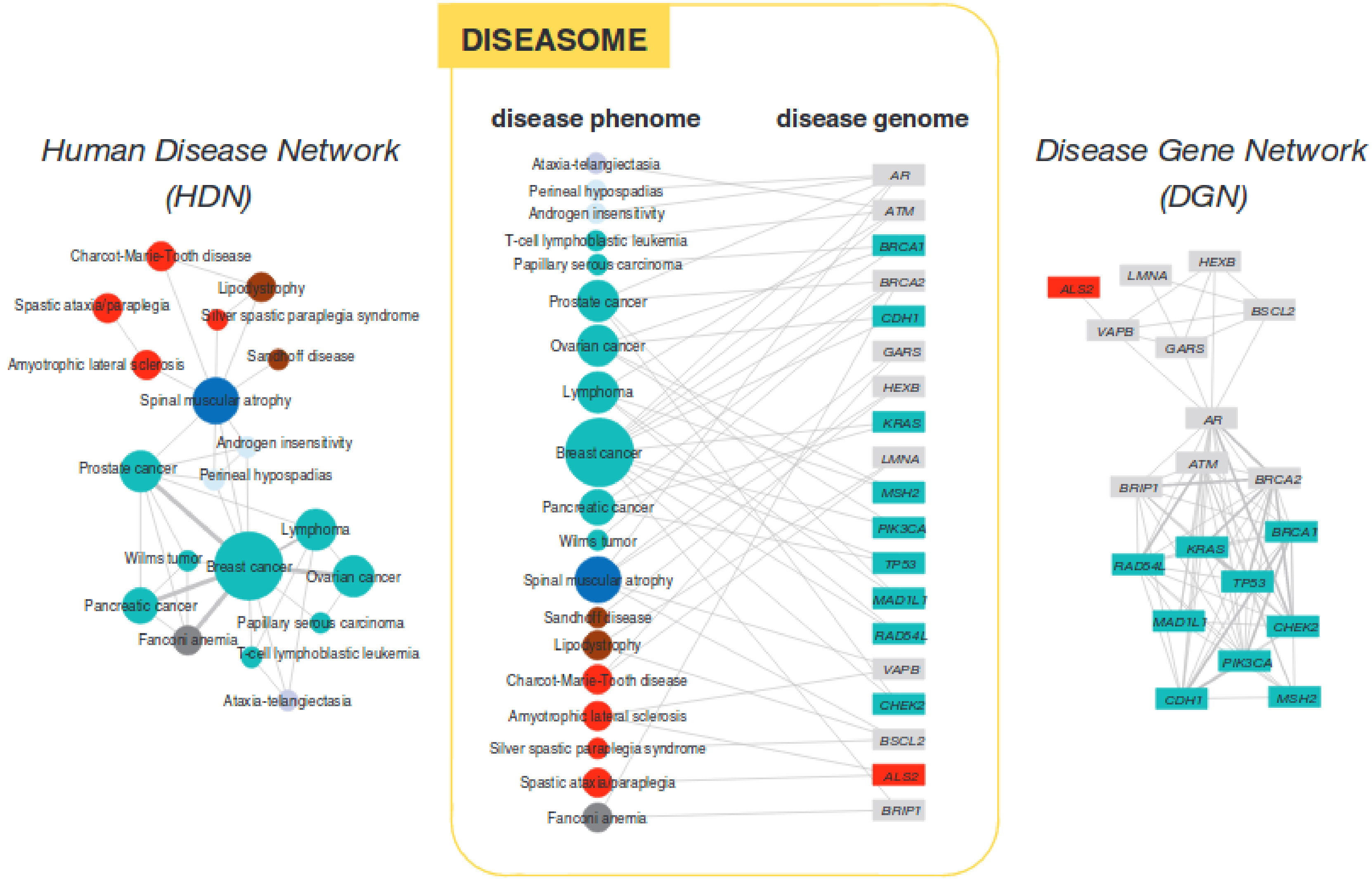
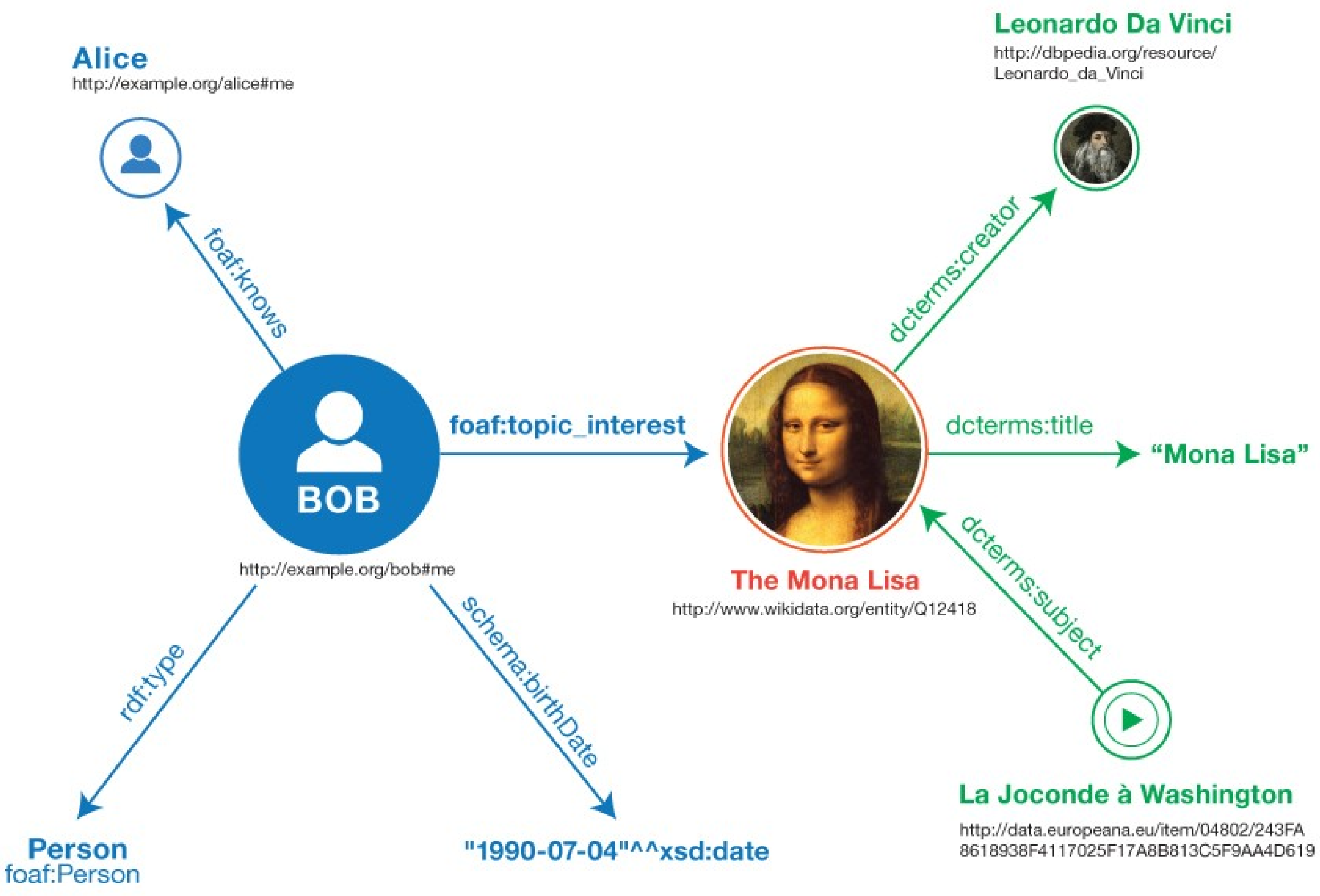
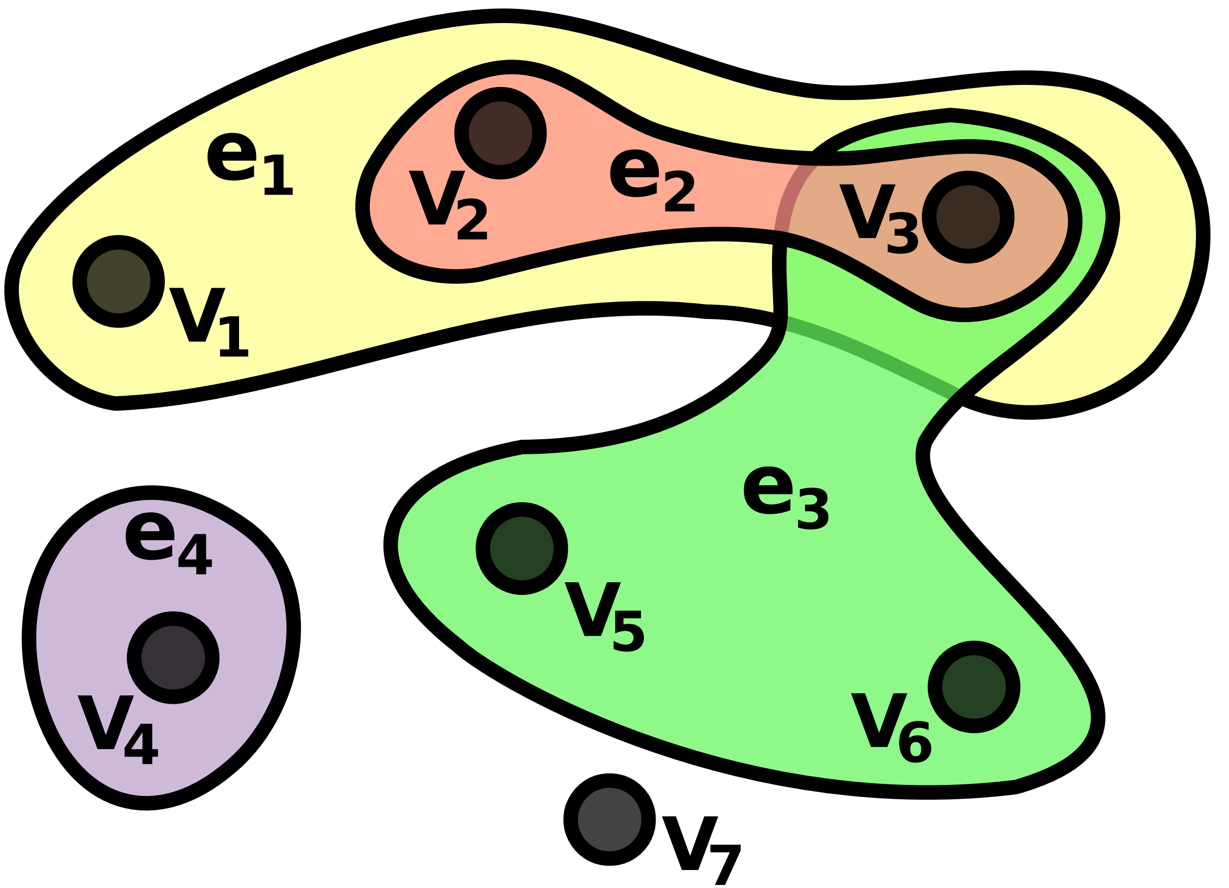


Image source: Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. PNAS (2007)

Tripartite network: RDF

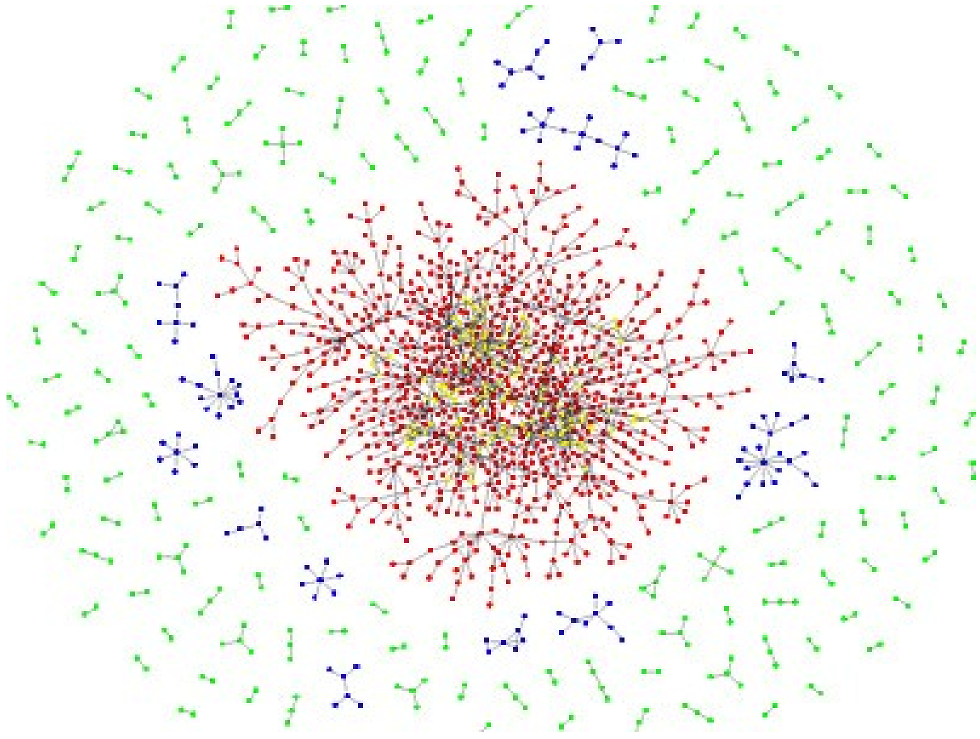


Hypergraph



Common basic properties

Components



- Isolated parts
- One large connected component

→ complications for:

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

Sparseness

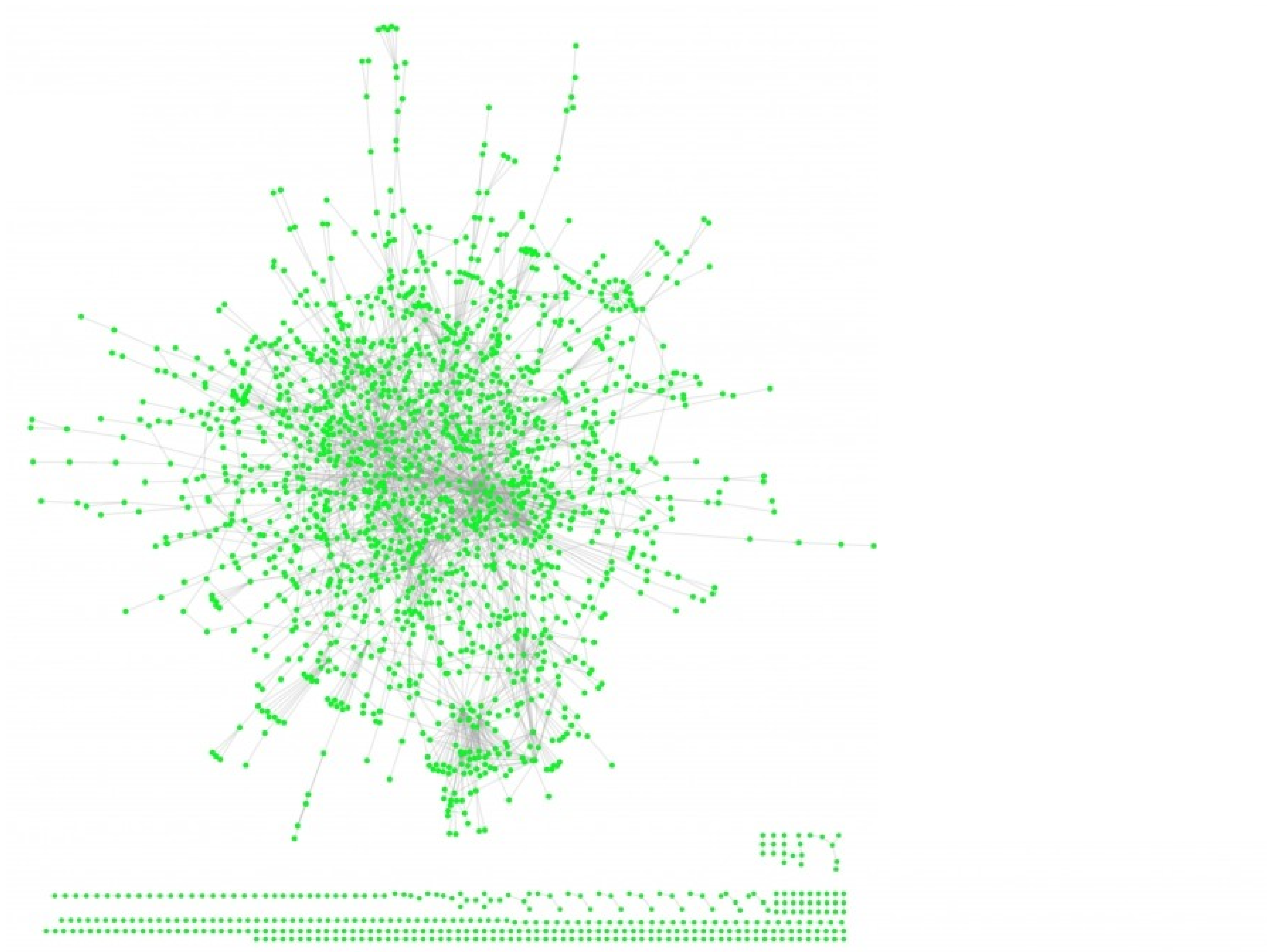


Image 2.4 and 2.7 from Barabasi's book

Sparseness

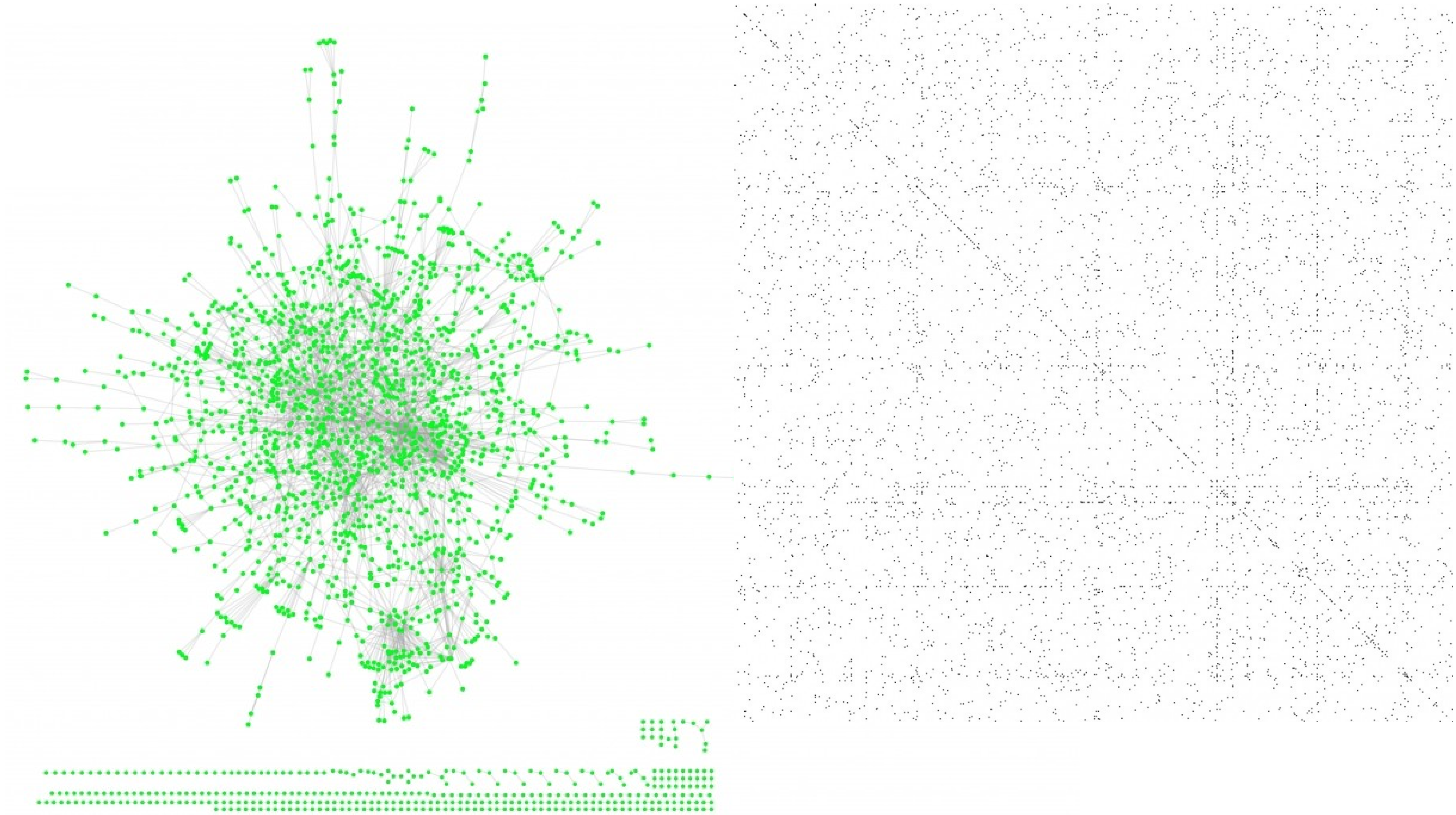
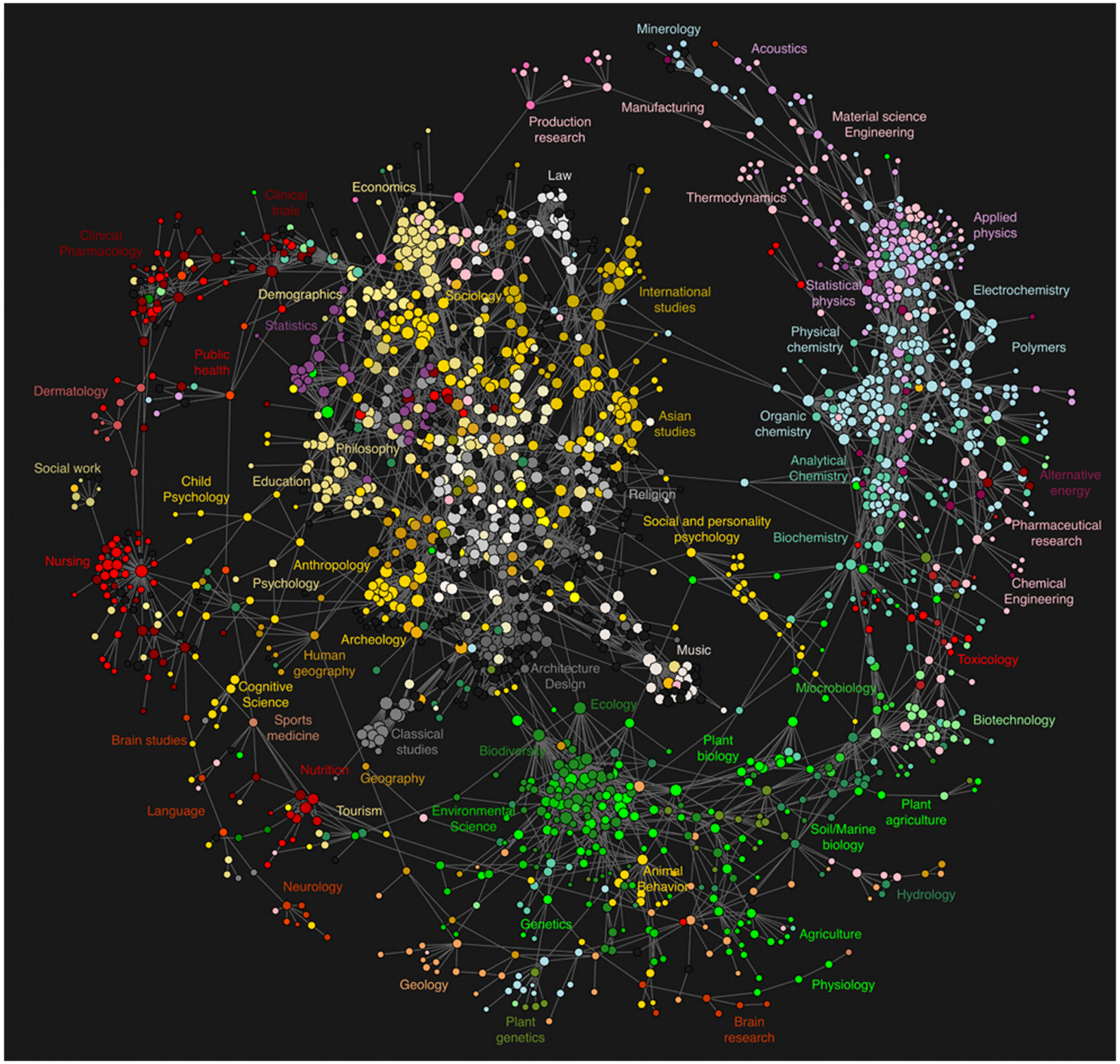


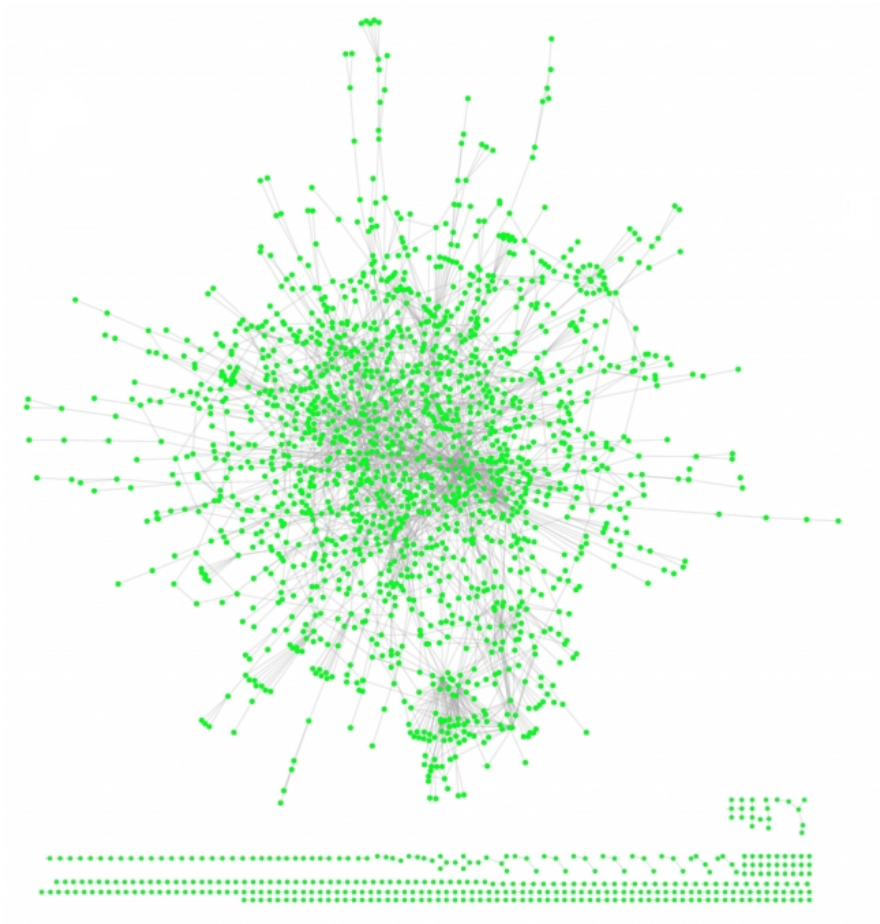
Image 2.4 and 2.7 from Barabasi's book

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

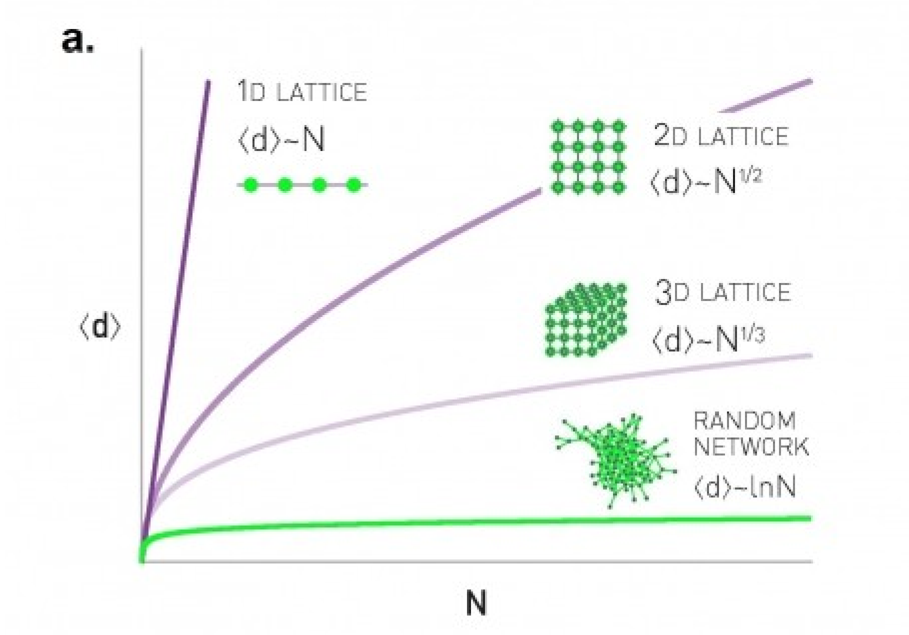


Image 3.11 from Barabasi's book

Degree distribution

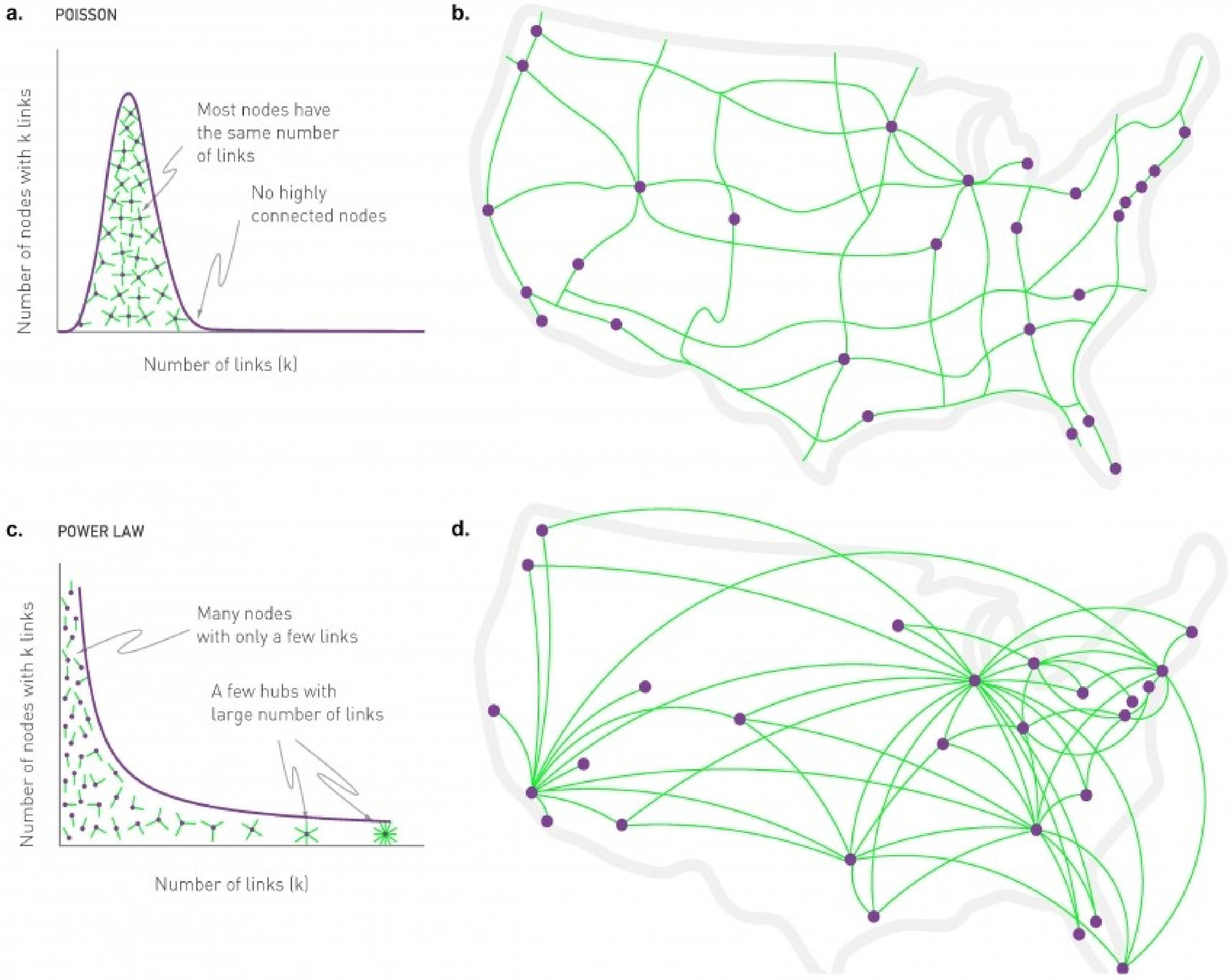
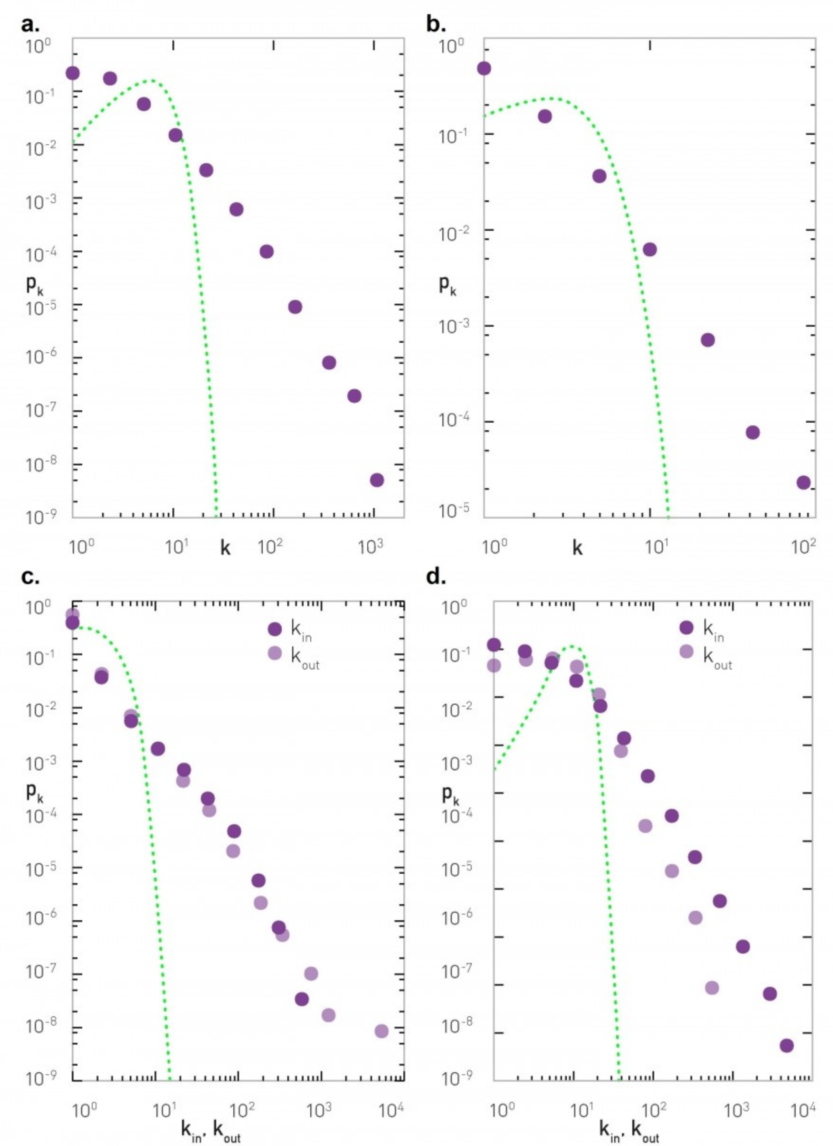


Image 4.6 from Barabasi’s book

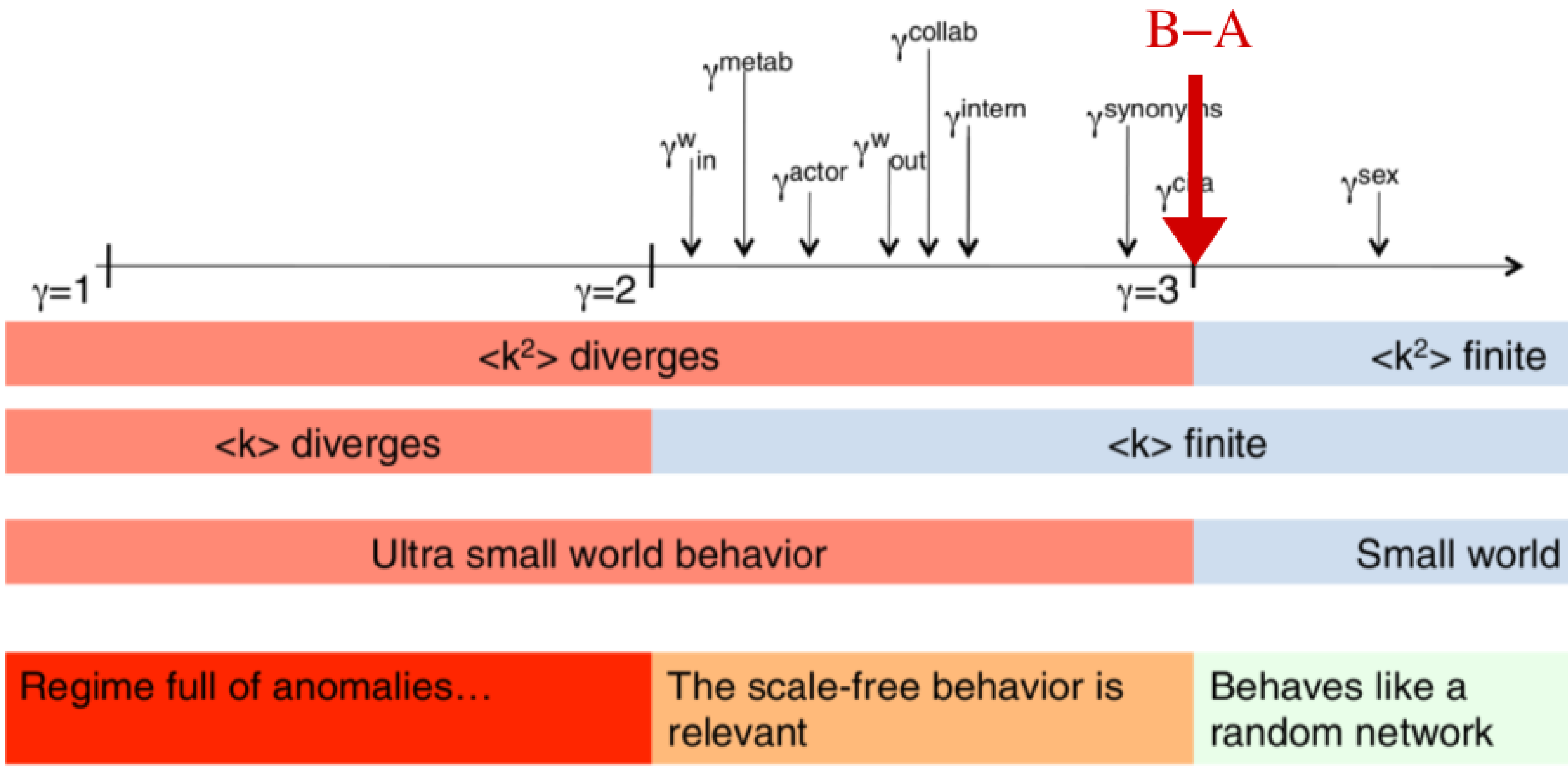
Degree distribution



- 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	$x^{-\alpha}$	$(\alpha - 1)x_{\min}^{\alpha-1}$
continuous	power law with cutoff	$x^{-\alpha}e^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha, \lambda x_{\min})}$
continuous	exponential	$e^{-\lambda x}$	$\lambda e^{\lambda x_{\min}}$
continuous	stretched exponential	$x^{\beta-1}e^{-\lambda x^{\beta}}$	$\beta\lambda e^{\lambda x_{\min}^{\beta}}$
continuous	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	$x^{-\alpha}$	$1/\zeta(\alpha, x_{\min})$
discrete	Yule distribution	$\frac{\Gamma(x)}{\Gamma(x+\alpha)}$	$(\alpha - 1)\frac{\Gamma(x_{\min}+\alpha-1)}{\Gamma(x_{\min})}$
discrete	exponential	$e^{-\lambda x}$	$(1 - e^{-\lambda})\,e^{\lambda x_{\min}}$
discrete	Poisson	$\mu^x/x!$	$\left[e^{\mu} - \sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$

TABLE 2.1

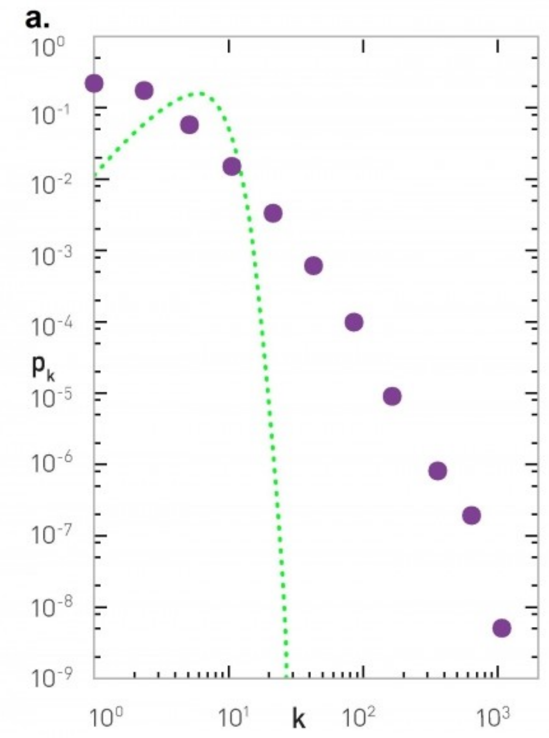


Table from:
http://tuvalu.santafe.edu/~aaronc/powerlaws/

What wasn’t visible on these slides

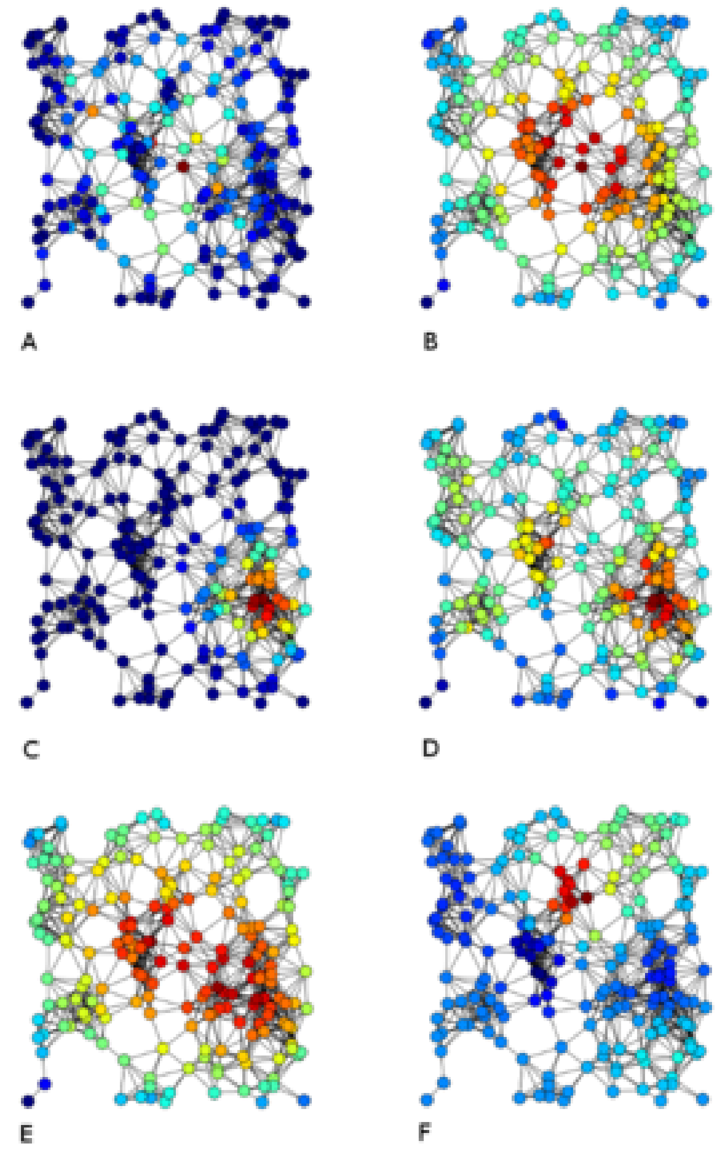
- Dirty data, artifacts
- “you get what you pay for”
- Network analysis is more sensitive to data cleaning than other methods
- But: for data analysis, less of a difference between “artifact” and “interesting phenomena”

Quantities

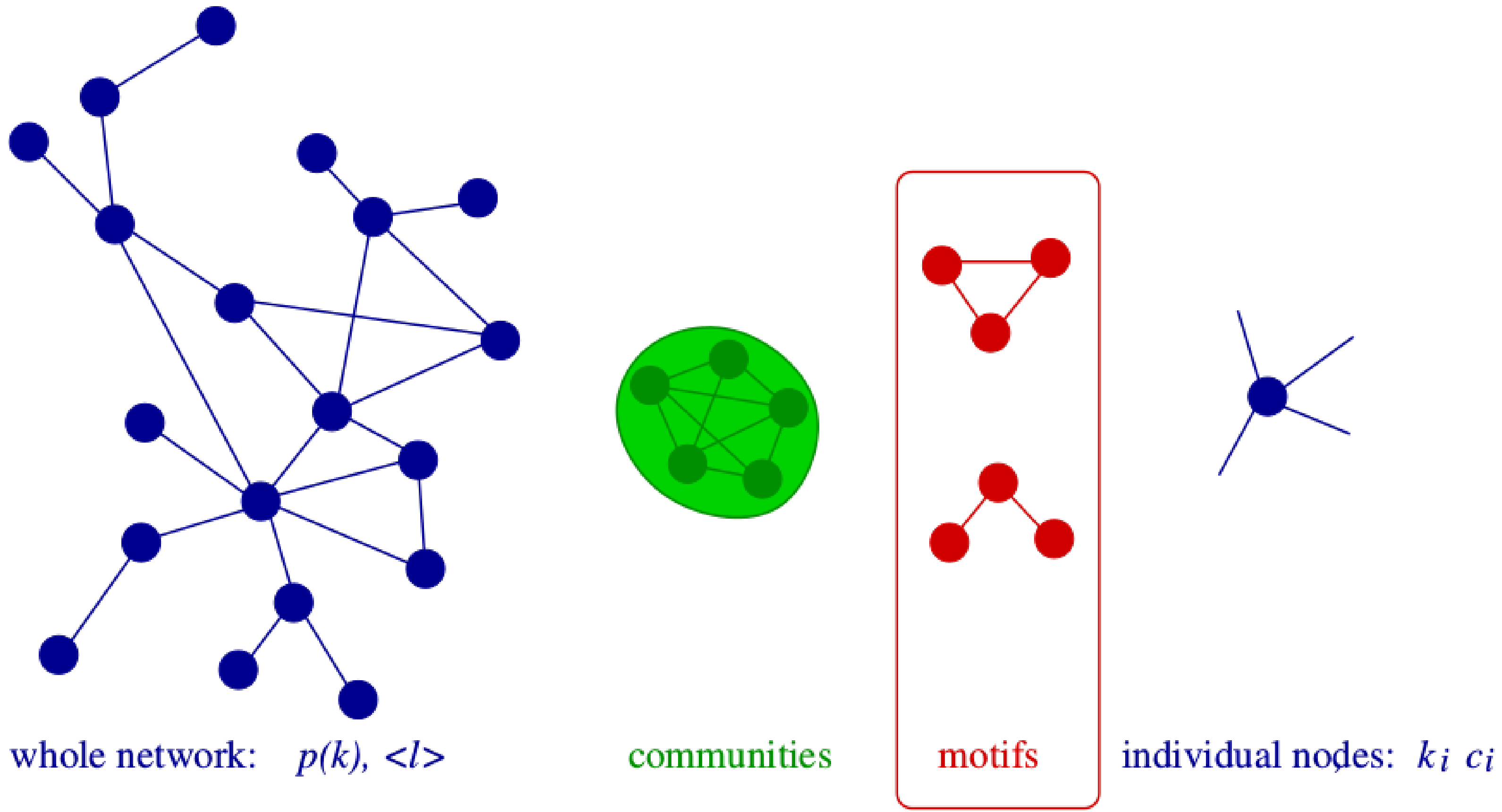
- Degree – directed (in- & out-), weighted (strength)
- Edges →Degree correlation (assortativity)
- Triangles → clustering coefficient
- Shortest paths→ betweeness
  - Geodetic vs. random paths
- Whole network – degree, distribution

Various centrality measures

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

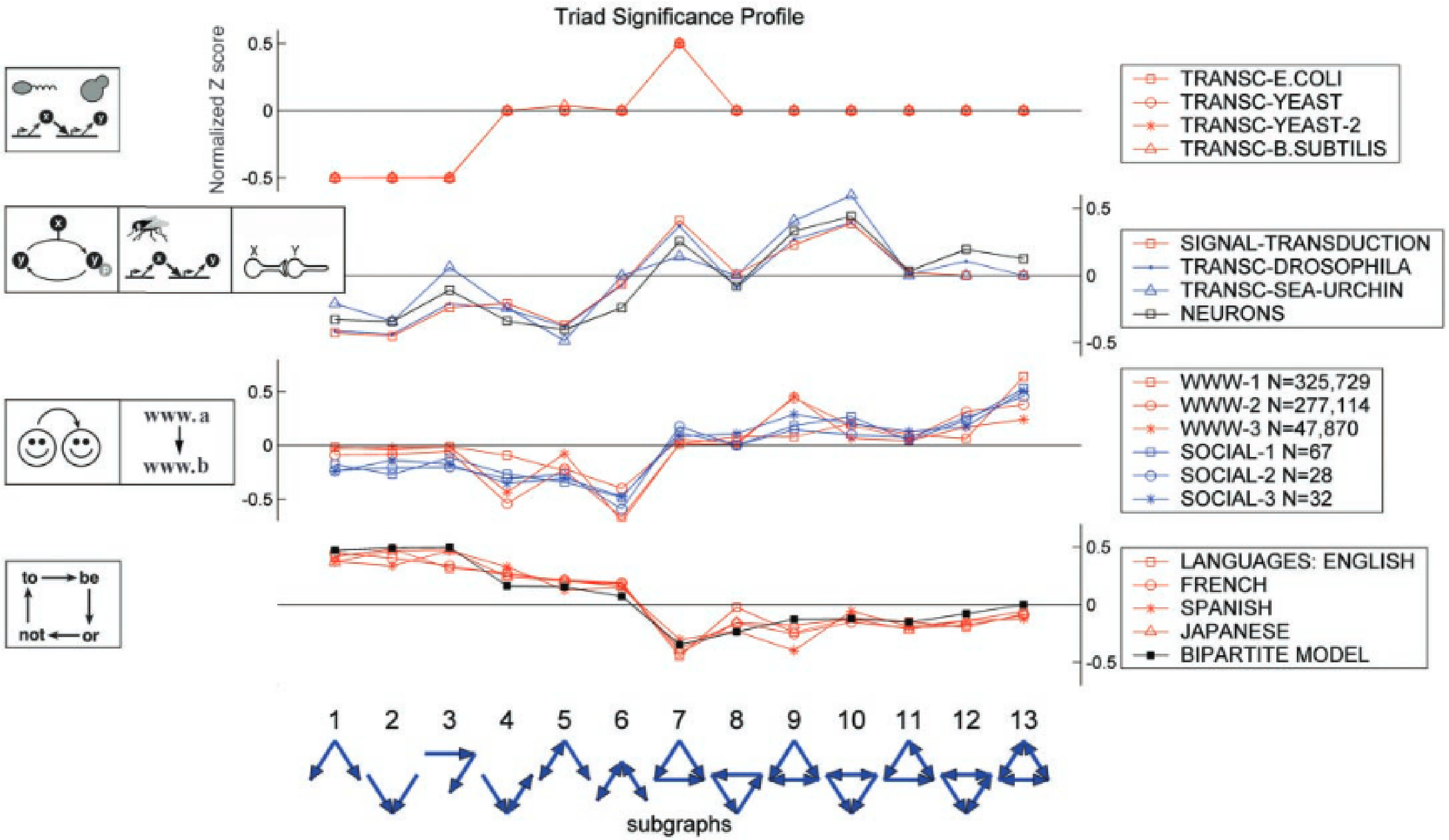


Scales



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

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)

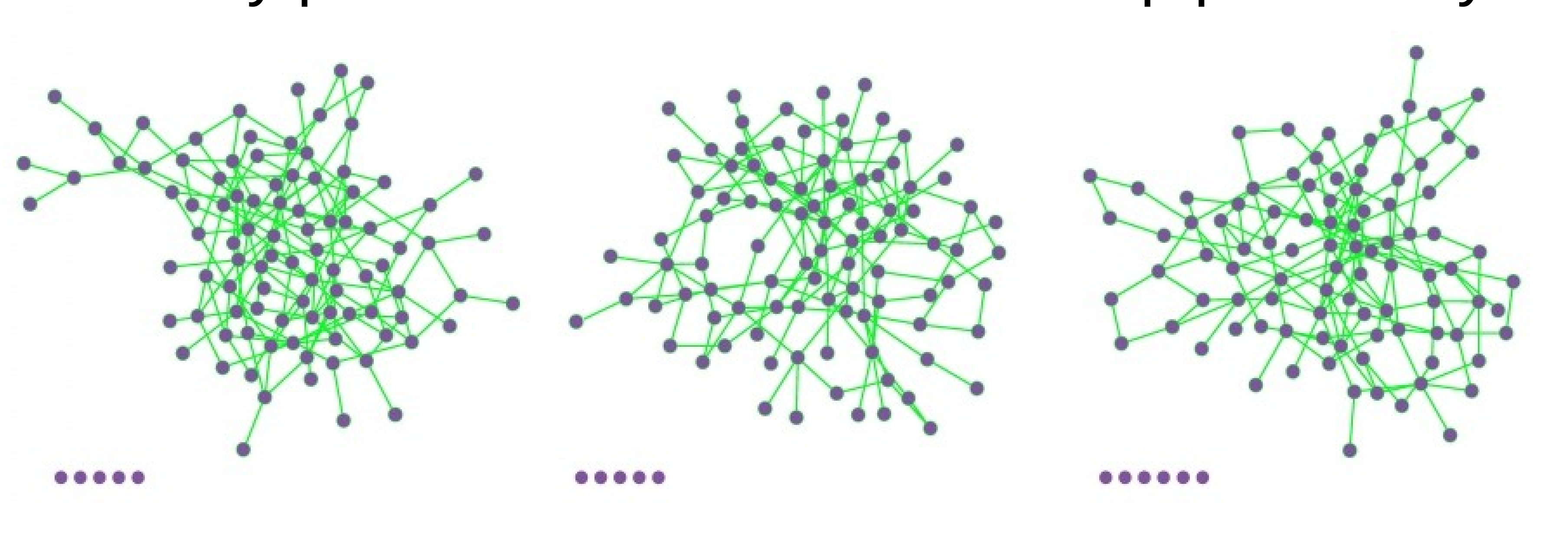
Random models

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

Erdős-Rényi model

- N nodes and:
  - M edges

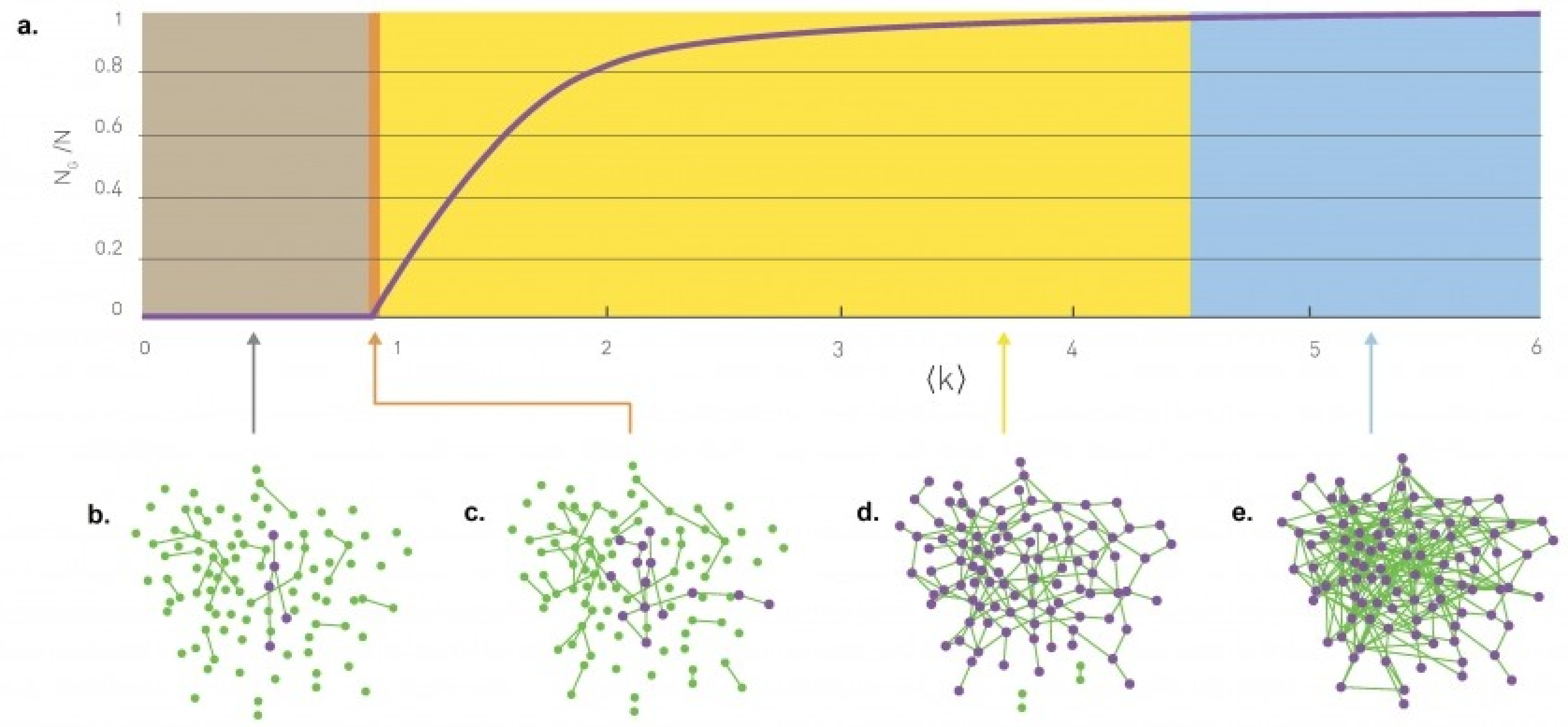
    or:
  - Every pair of nodes connected with p probability



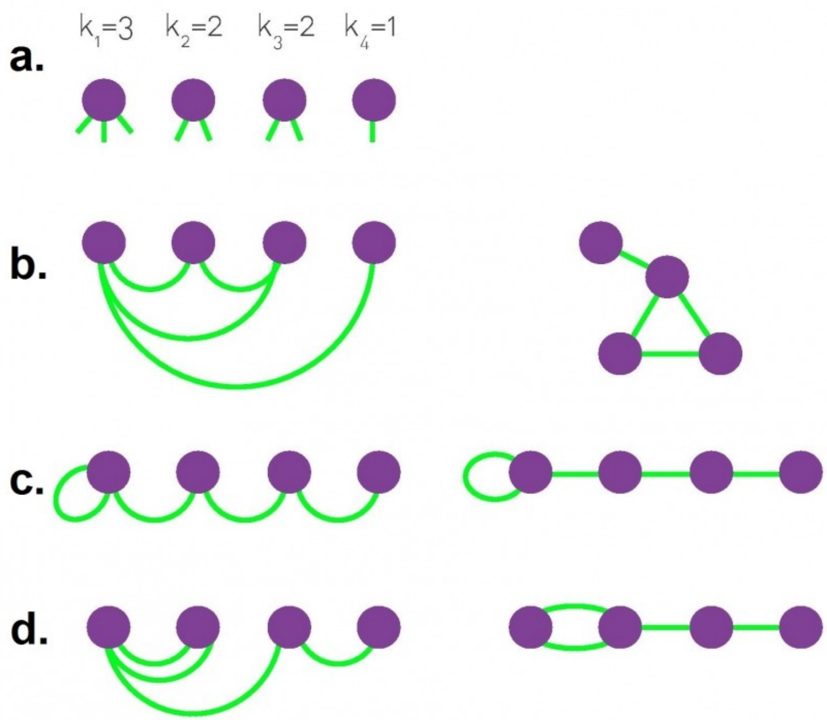
N=100, p=0.03 – Image 3.3 from Barabási’s book

Erdős-Rényi model

- If adding edges one by one: percolation



Configuration model: given degrees

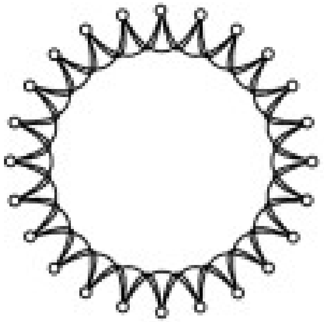


- Nodes & edge ends
- Connecting edge ends
- (not all sequences suitable)

Image 4.15 from Barabási's book

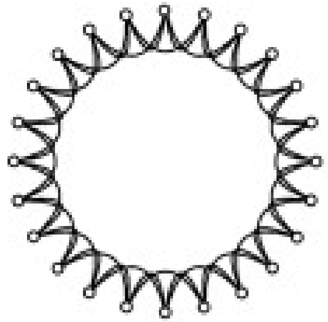
Watts-Strogatz – small world

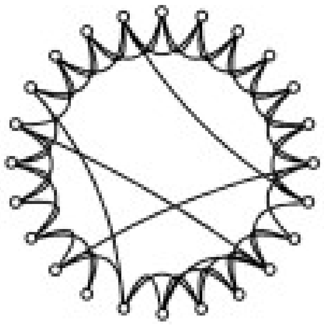
- Ring, close neighbors connected



Watts-Strogatz – small world

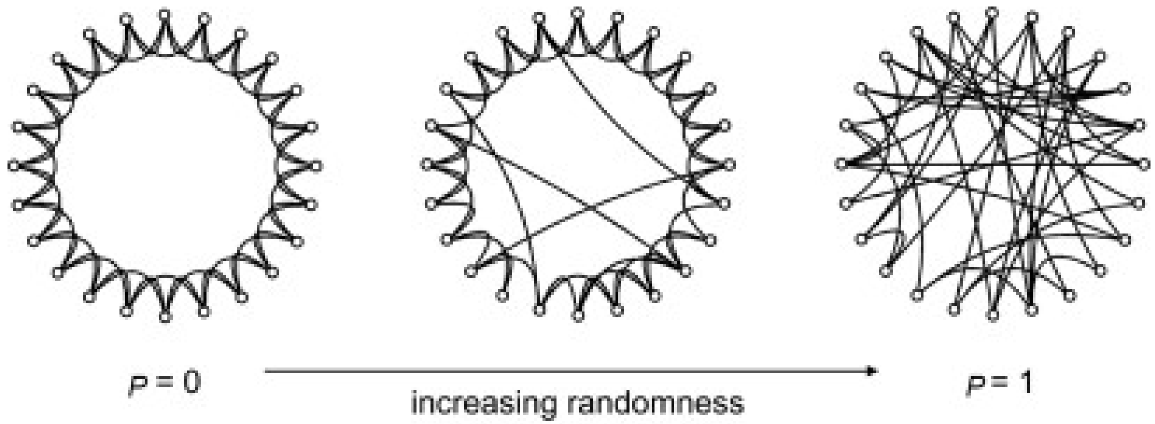
- Ring, close neighbors connected
- Randomly move edges





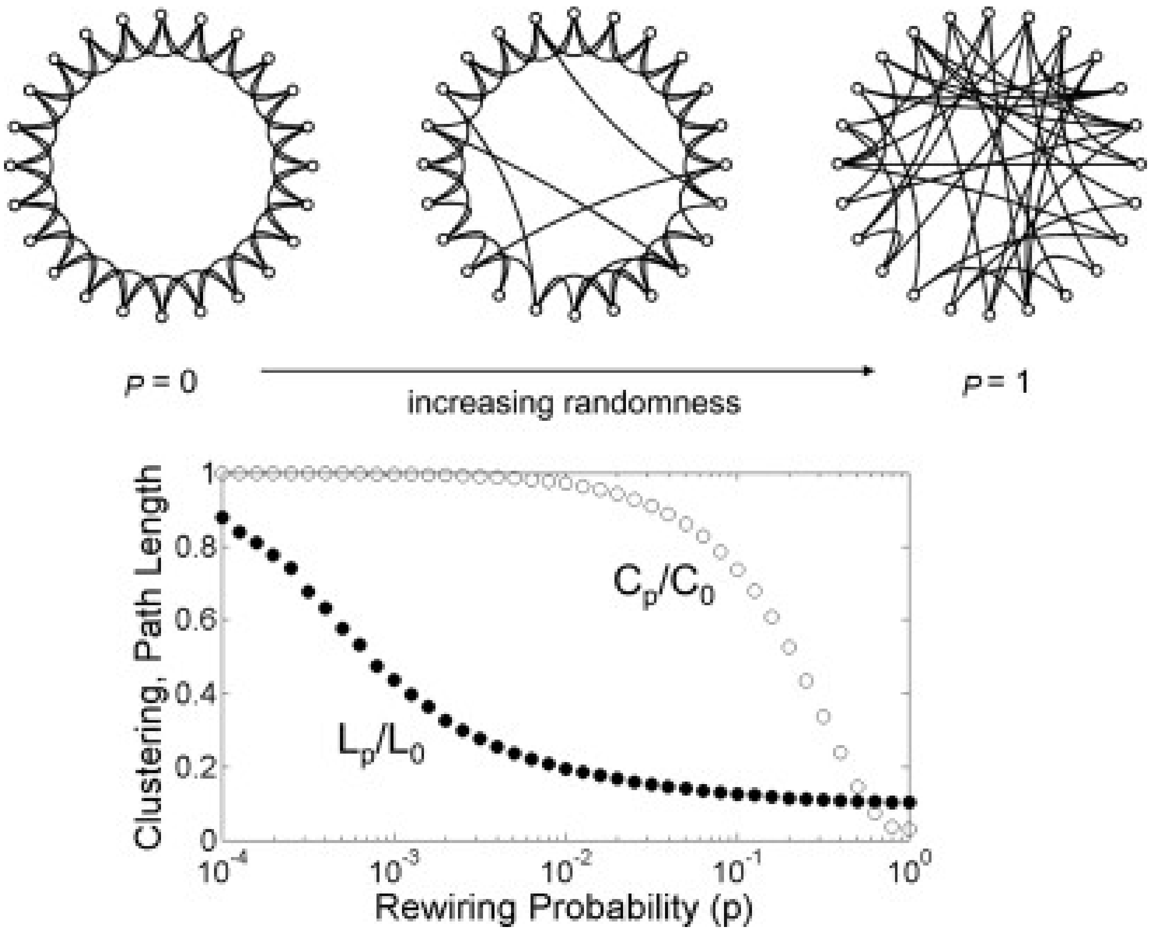
Watts-Strogatz – small world

- 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

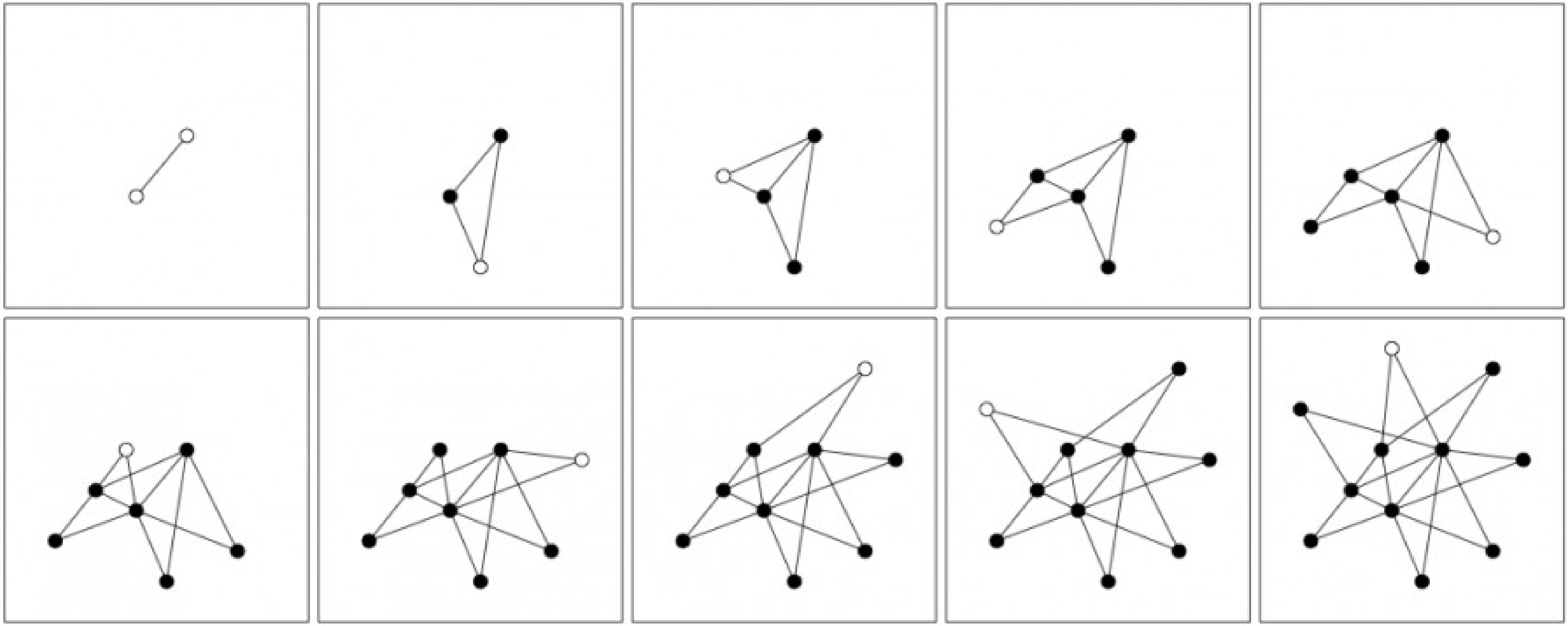


Image 5.3 from Barabási's book

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”

Other tools

- igraph – http://igraph.org/
  - C library, R, C++ and python wrapper
  - Harder to use than networkx, but more optimized
- 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.

DEMO

Egy matematikai eredmény levezetése

Section 5.13

Advanced Topic 4.A
Deriving the Degree Distribution

A number of analytical techniques are available to calculate the exact form of the degree exponent (5.11). Next we derive it using the rate equation approach [12, 13]. The method is sufficiently general to help explore the properties of a wide range of growing networks. Consequently, the calculations described here are of direct relevance for many systems, from models pertaining to the WWW [16, 17, 18] to describing the evolution of the protein interaction network via gene duplication [19, 20].

Let us denote with $N(k,t)$ the number of nodes with degree k at time t. The degree distribution $p_k(t)$ relates to this quantity via $p_k(t) = N(k,t)/N(t)$. Since at each time-step we add a new node to the network, we have $N = t$. That is, at any moment the total number of nodes equals the number of timesteps (BOX 5.2).

We write preferential attachment as

$$\Pi(k) = \frac{k}{\sum_j k_j} = \frac{k}{2mt} \qquad (5.31)$$

where the $2m$ term captures the fact that in an undirected network each link contributes to the degree of two nodes. Our goal is to calculate the changes in the number of nodes with degree k after a new node is added to the network. For this we inspect the two events that alter $N(k,t)$ and $p_k(t)$ following the arrival of a new node:

- A new node can link to a degree-k node, turning it into a degree (k+1) node, hence *decreasing* $N(k,t)$.
- A new node can link to a degree (k-1) node, turning it into a degree k node, hence *increasing* $N(k,t)$.

The number of links that are expected to connect to degree k nodes after the arrival of a new node is

$$\frac{k}{2mt} \times Np_k(t) \times m = \frac{k}{2} p_k(t) \qquad (5.32)$$

In (5.32) the first term on the l.h.s. captures the probability that the new node will link to a degree-k node (preferential attachment); the second term provides the total number of nodes with degree k, as the more nodes are in this category, the higher the chance that a new node will attach to one of them; the third term is the degree of the incoming node, as the higher is m, the higher is the chance that the new node will link to a degree-k node. We next apply (5.32) to cases (i) and (ii) above:

- The number of degree k nodes that acquire a new link and turn into (k+1) degree nodes is
$$\frac{k}{2} p_k(t) \qquad (5.33)$$
- The number of degree (k-1) nodes that acquire a new link, increasing their degree to k is
$$\frac{k-1}{2} p_{k-1}(t) \qquad (5.34)$$

Combining (5.33) and (5.34) we obtain the expected number of degree-k nodes after the addition of a new node

$$(N+1)p_k(t+1) = Np_k(t) + \frac{k-1}{2} p_{k-1}(t) - \frac{k}{2} p_k(t) \qquad (5.35)$$

This equation applies to all nodes with degree $k > m$. As we lack nodes with degree k=0,1, ... , m-1 in the network (each new node arrives with degree m) we need a separate equation for degree-m modes. Following the same arguments we used to derive (5.35), we obtain

$$(N+1)p_m(t+1) = Np_m(t) + 1 - \frac{m}{2} p_m(t) \qquad (5.36)$$

Equations (5.35) and (5.36) are the starting point of the recursive process that provides p_k. Let us use the fact that we are looking for a stationary degree distribution, an expectation supported by numerical simulations (Image 5.6). This means that in the $N = t \to \infty$ limit, $p_k(\infty) = p_k$. Using this we can write the l.h.s. of (5.35) and (5.36) as

$$(N+1)p_k(t+1) - Np_k(t) \to Np_k(\infty) + p_k(\infty) - Np_k(\infty) = p_k(\infty) = p_k,$$
$$(N+1)p_m(t+1) - Np_m(t) \to p_m$$

Therefore the rate equations (5.35) and (5.36) take the form:

$$p_k = \frac{k-1}{k+2} p_{k-1} \qquad k > m \qquad (5.37)$$
$$p_m = \frac{2}{m+2} \qquad (5.38)$$

Note that (5.37) can be rewritten as

$$p_{k+1} = \frac{k}{k+3} p_k \qquad (5.39)$$

via a $k \to k+1$ variable change.

We use a recursive approach to obtain the degree distribution. That is, we write the degree distribution for the smallest degree, $k=m$, using (5.38) and then use (5.39) to calculate p_k for the higher degrees:

$$p_{m+1} = \frac{m}{m+3} p_m = \frac{2m}{(m+2)(m+3)}$$
$$p_{m+2} = \frac{m+1}{m+4} p_{m+1} = \frac{2m(m+1)}{(m+2)(m+3)(m+4)} \qquad (5.40)$$
$$p_{m+3} = \frac{m+2}{m+5} p_{m+2} = \frac{2m(m+1)}{(m+3)(m+4)(m+5)}$$

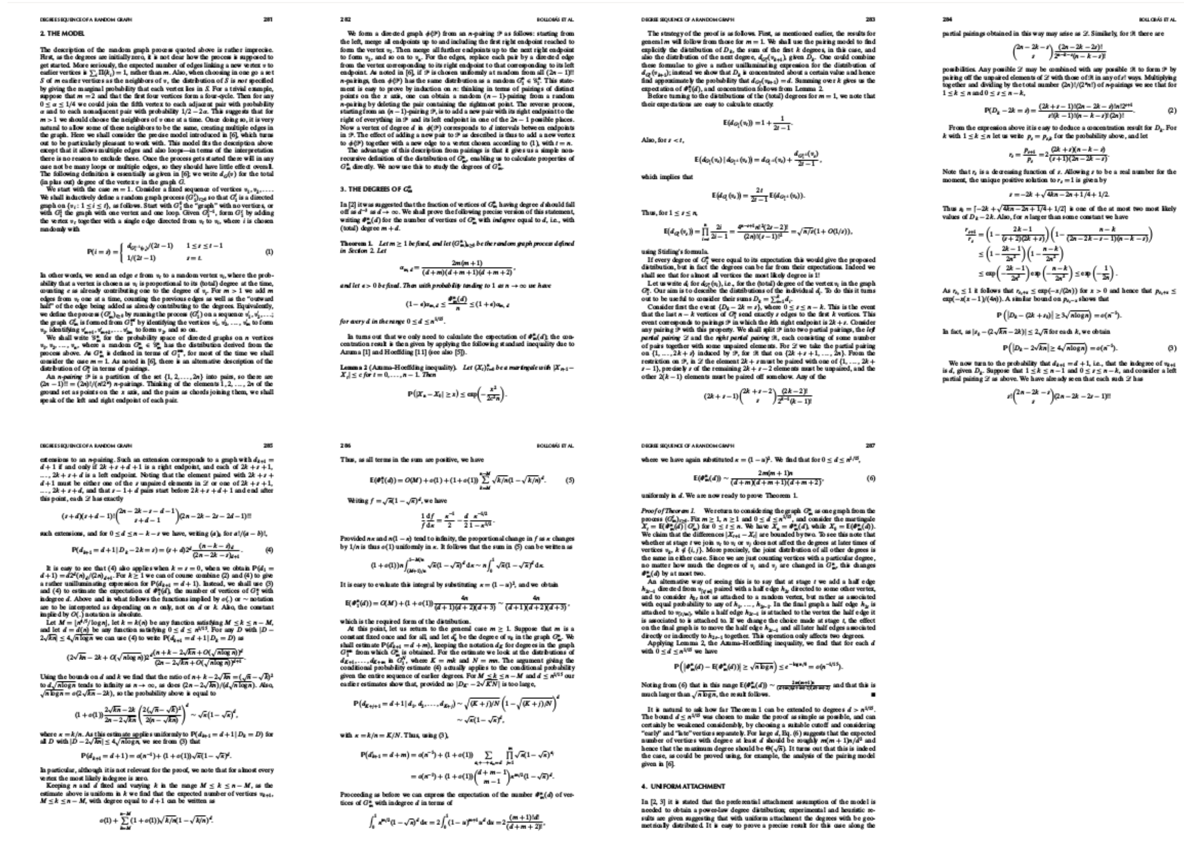
At this point we notice a simple recursive pattern: By replacing in the denomerator m+3 with k we obtain the probability to observe a node with degree k

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \qquad (5.41)$$

which represents the exact form of the degree distribution of the Barabási-Albert model.

Note that:

- For large k (5.41) becomes $p_k \sim k^{-3}$, in agreement with the numerical result.
- The prefactor of (5.11) or (5.41) is different from the prefactor of (5.9).
- This form was derived independently in [12] and [13], and the exact mathematical proof of its validity is provided in [10].



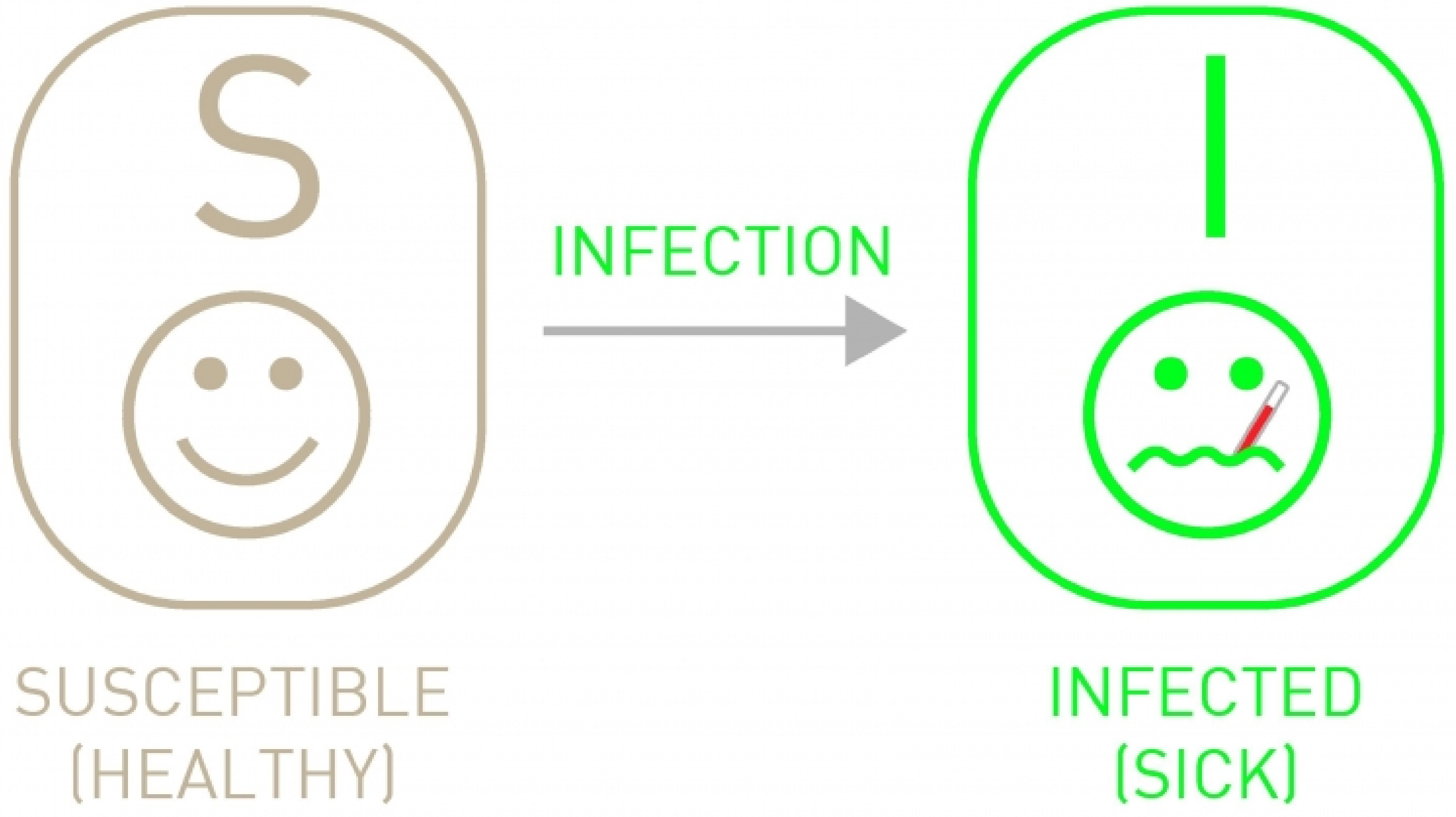
“Matematikusosan”: 7 oldal
“Fizikusosan”: 2 oldal

Modellezésről

Modellezésről

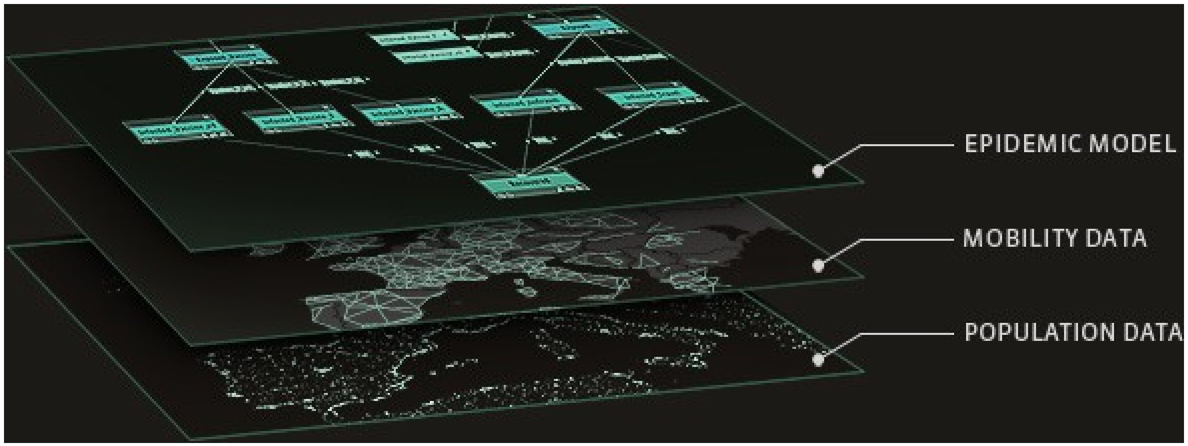
- Mi a cél?
  - Szimuláció – pontos, kvantitatív előrejelzés
  - kvalitatív modell – alaptulajdonságok megértése
- Neumann János: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."
- Occam borotvája

Kvalitatív modell

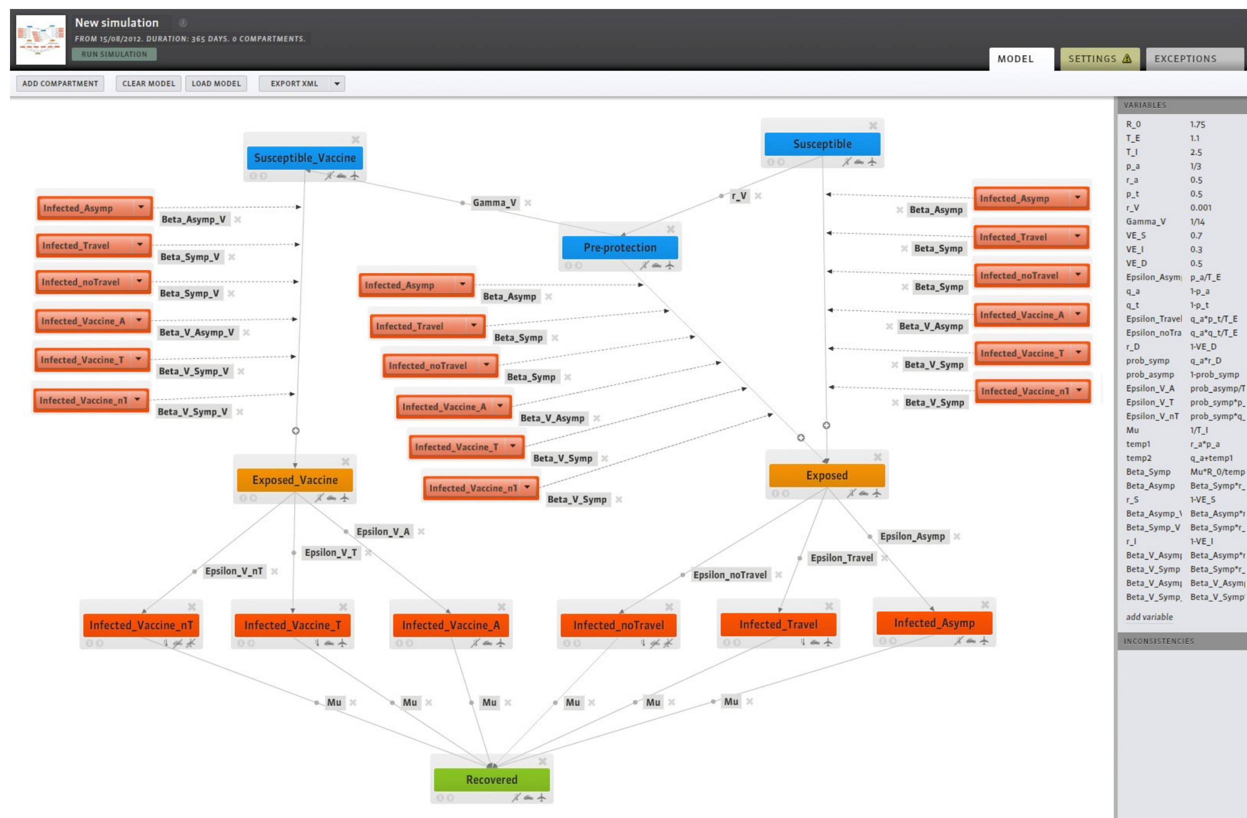


Csúcsok S állapotból p valószinűséggel I állapotba

Kvantitatív szimuláció



http://www.gleamviz.org/



Adatelemzés folyamata: absztrakciós szintek

1) Komplex rendszer
2) Adat
3) Hálózat
4) Matematikai mennyiség
5) Értelmezett mennyiség
6) Konklúzió

Adatelemzés folyamata: absztrakciós szintek

1) Komplex rendszer – Bp-i tömegközlekedés
2) Adat – jármű-mozgások
3) Hálózat – megállók & utak
4) Matematikai mennyiség – fokszám
5) Értelmezett mennyiség – megálló fontossága
6) Konklúzió – hova érdemes elősször online kijelzőt rakni

Fázis-átmenetek, kritikus pontok, univerzalizás

- Ami a fizikusoknak érdekes

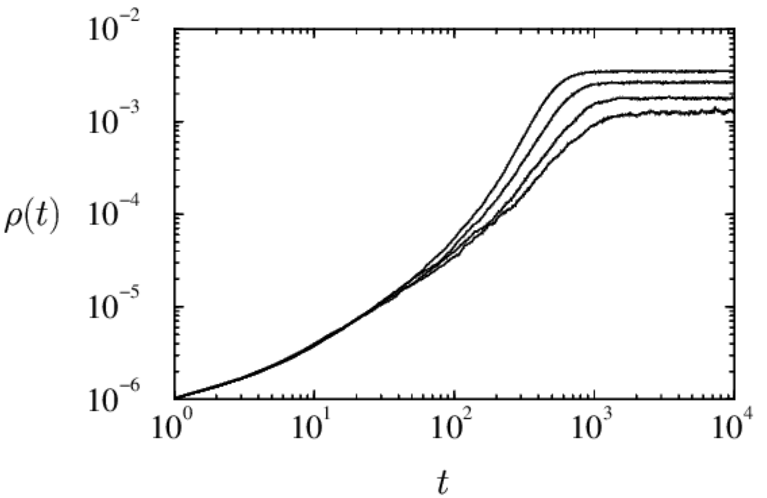


FIG. 3. Density of infected nodes $\rho(t)$ as a function of time in supercritical spreading experiments in the WS network. Network size $N = 1.5 \times 10^6$. Spreading rates range from $\lambda - \lambda_c = 0.002$ to 0.0007 (top to bottom).

Fázis-átmenetek, kritikus pontok, univerzalizás

- Ami a fizikusoknak érdekes

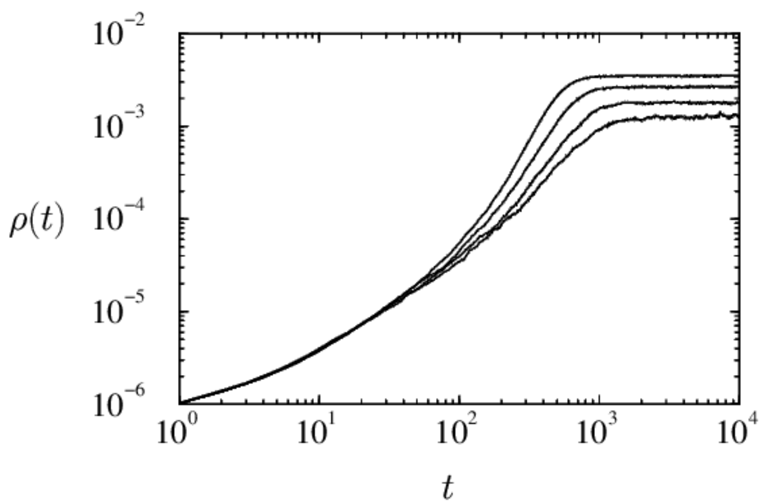


FIG. 3. Density of infected nodes $\rho(t)$ as a function of time in supercritical spreading experiments in the WS network. Network size $N = 1.5 \times 10^6$. Spreading rates range from $\lambda - \lambda_c = 0.002$ to 0.0007 (top to bottom).

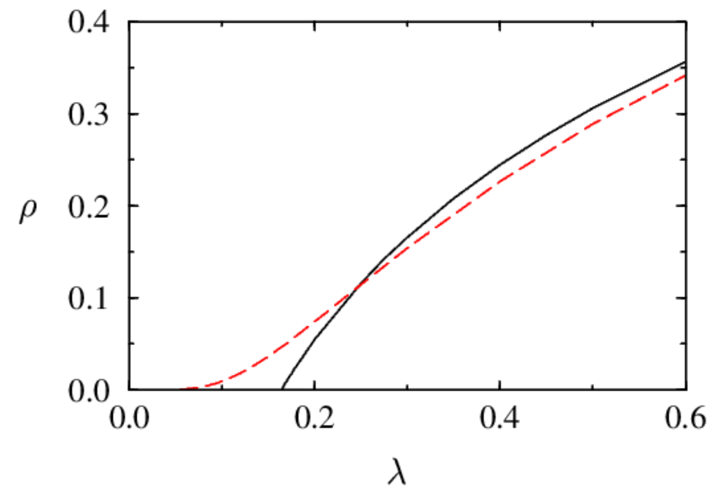
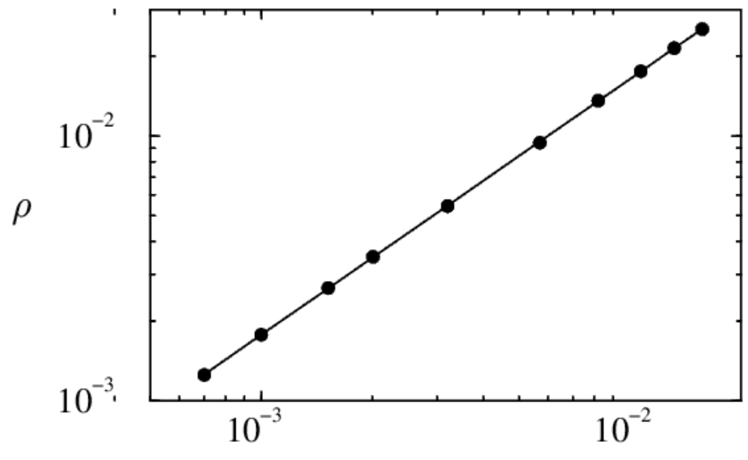


FIG. 1. Density of infected nodes ρ as a function of λ in the WS network (full line) and the BA network (dashed line).



Grafikus programok

- http://cytoscape.org/
- https://gephi.org/
- Import-hoz adatszerkezet: él lista (xlsx, csv)
- Továbbiak jövő héten