Psych 253
Advanced Statistical Modeling

Graph theory

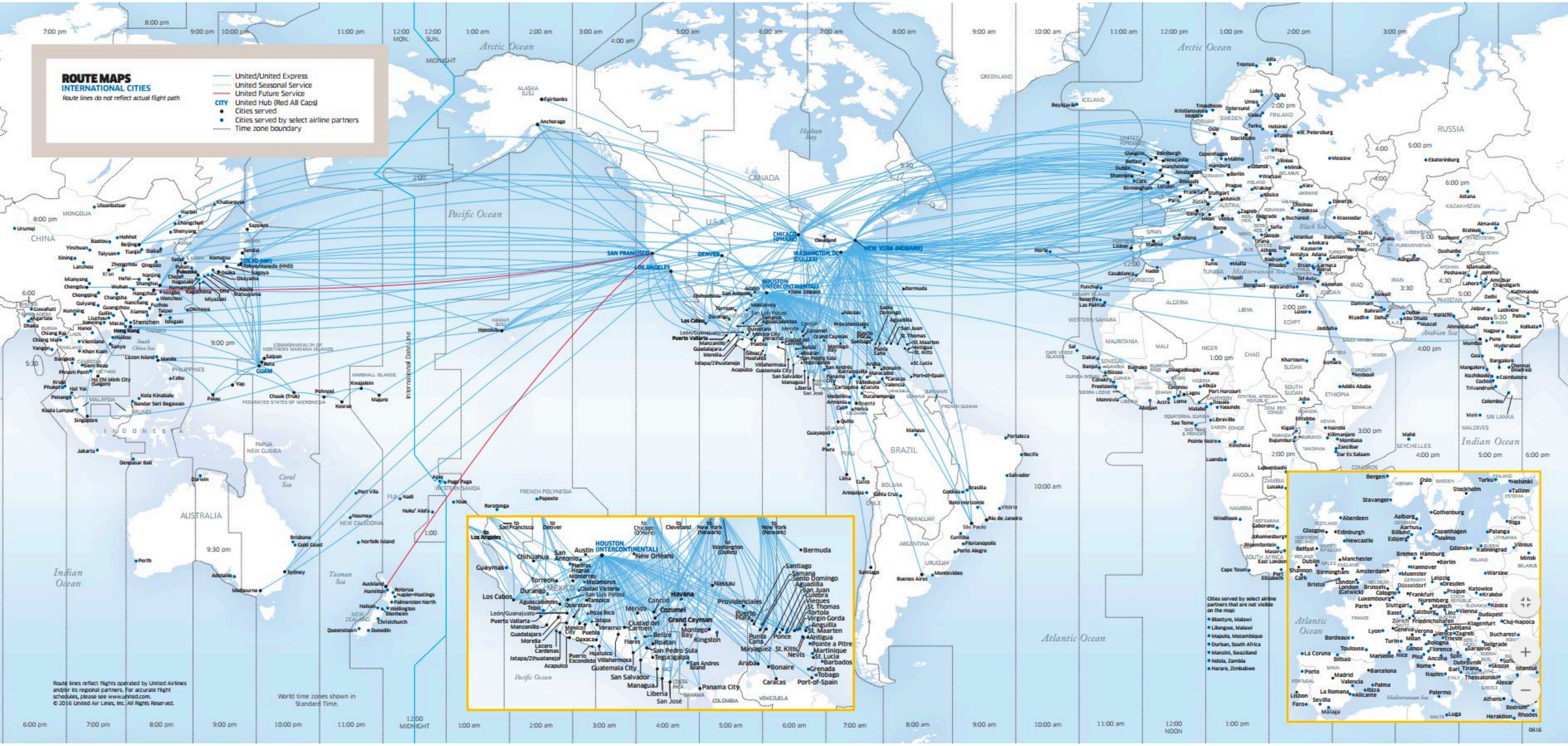
Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford Artificial Intelligence Laboratory
Stanford University

Russ Poldrack
Department of Psychology
Stanford University

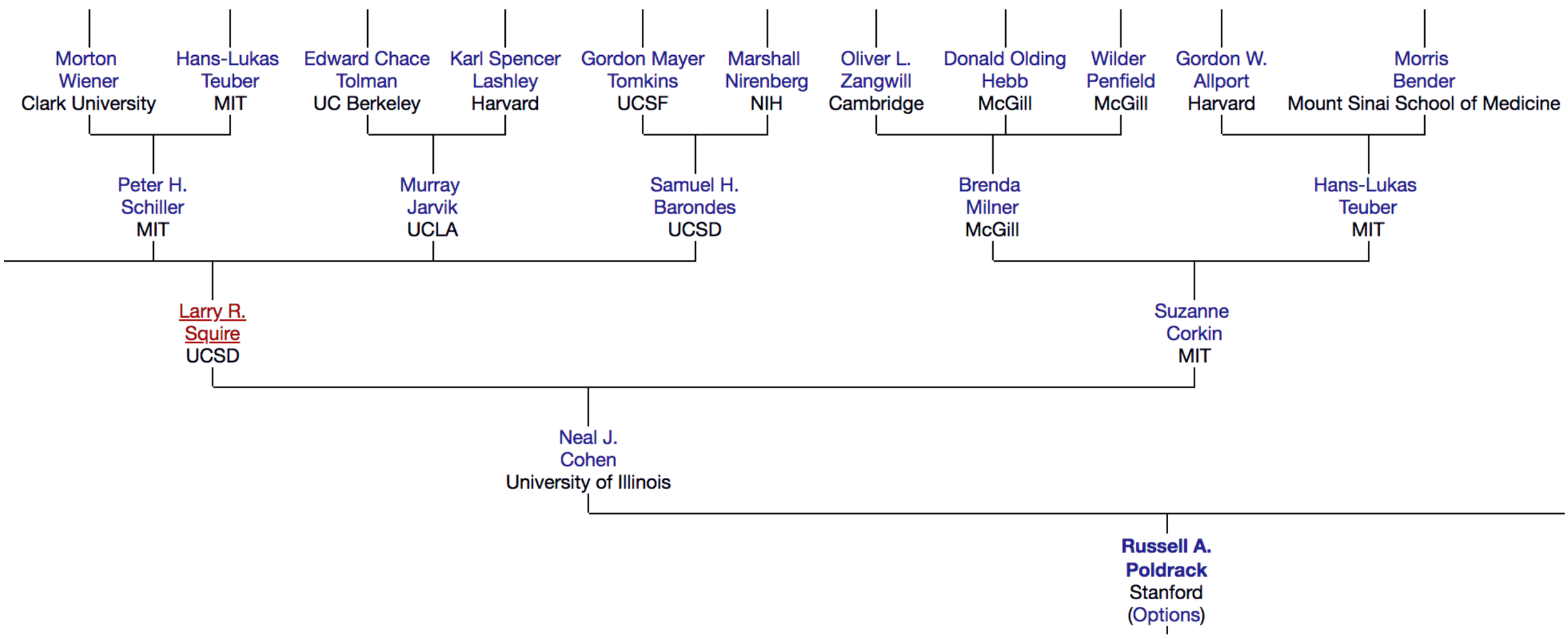
What is a graph?

- A formal description of a set of things (“nodes”) and their relationships (“edges”)

What are the nodes and edges?



What are the nodes and edges?



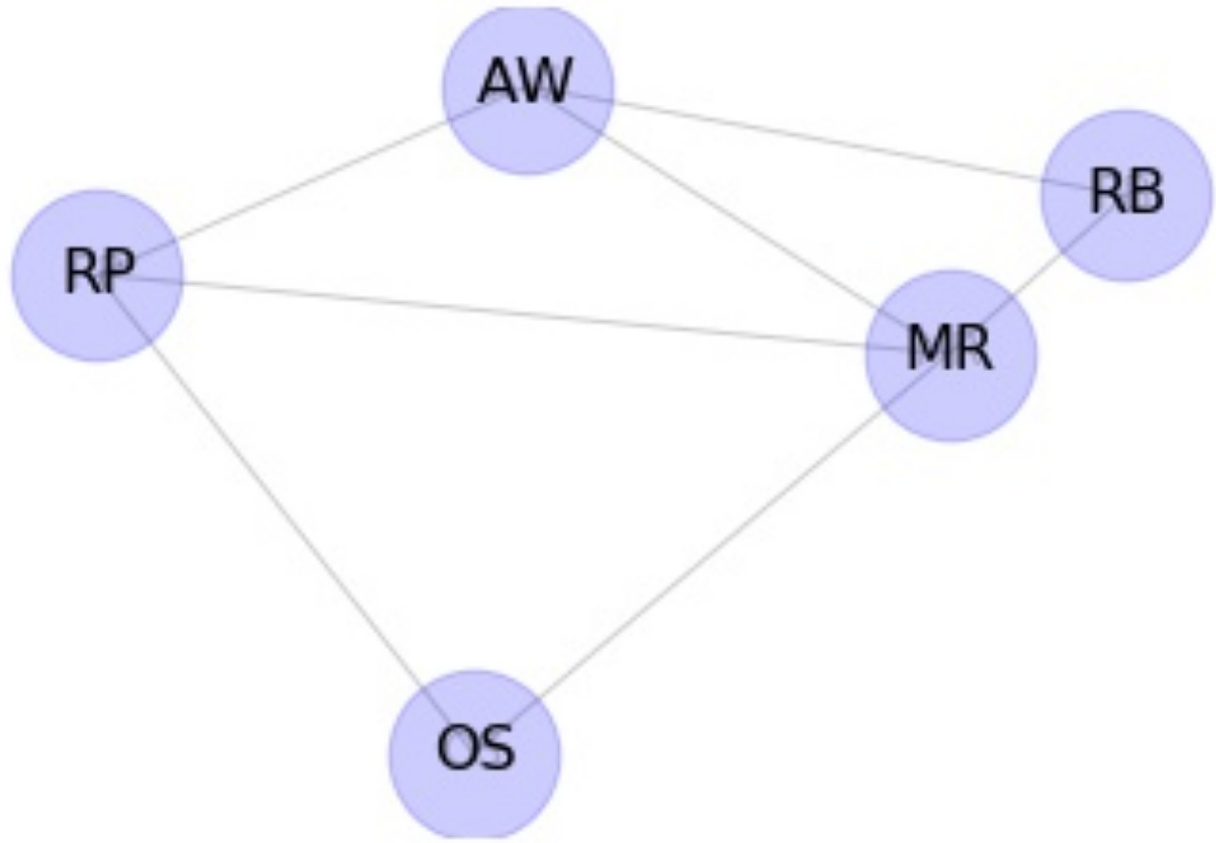
https://neurotree.org/neurotree/tree.php?pid=4016

Describing a graph

- G = (V,E)
  - V: a set of *vertices* (i.e. nodes)
  - E: a set of *edges* (i.e. connections)
    - e = (a,b) where a and b are vertices in V

Co-publication data

```
RP : Russ Poldrack
AW : Anthony Wagner
RB : Randy Buckner
MR : Marcus Raichle
OS : Olaf Sporns
```



V: (‘RP’, 'AW', 'RB', 'MR', 'OS')

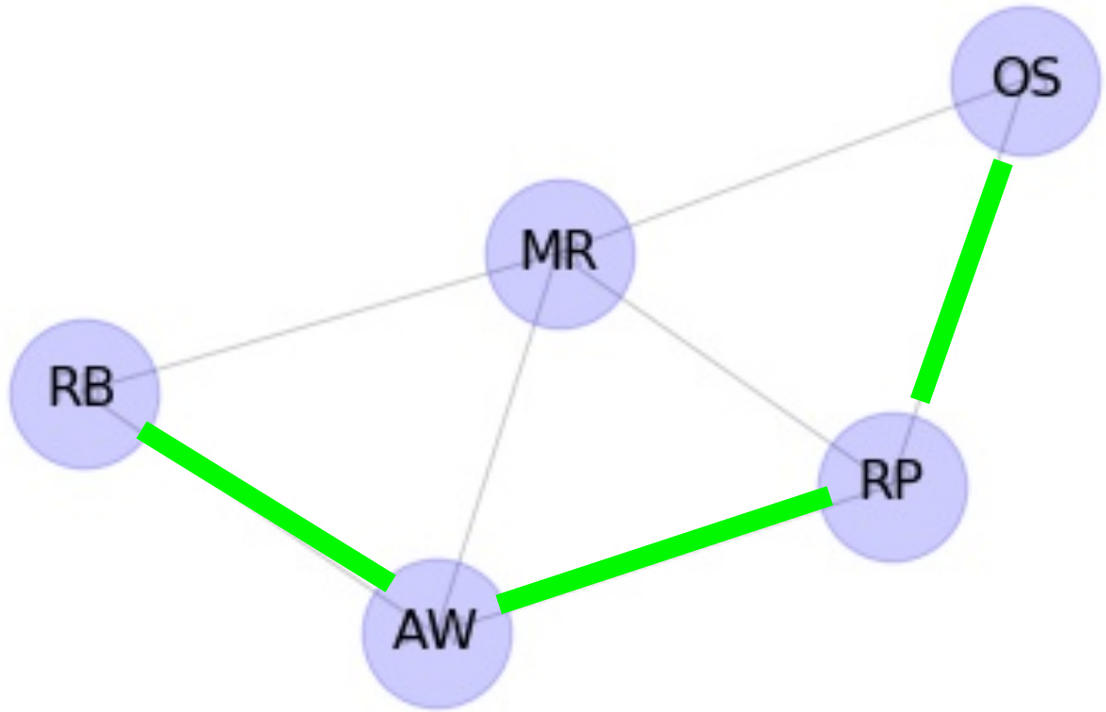
E = [(‘RP', 'AW'), ('RP', 'MR'), ('RP', 'OS'), ('AW', 'RB'), (‘AW’, 'MR'), ('RB', 'MR'), (‘MR', 'OS')]

Types of graphs

- Directed vs undirected
  - Does the edge have a direction? (i.e. arrow or arrowless?)
- Weighted vs. unweighted
  - Do the edges differ in their strength, or are they binary (present/absent)?
- Examples
  - Directed/unweighted: parenthood
  - Undirected/unweighted: social network friendship
  - Directed/weighted: airline connections (weighted by # of flights)
  - Undirected/weighted: co-publication (weighted by # of publications)

Traversing a graph

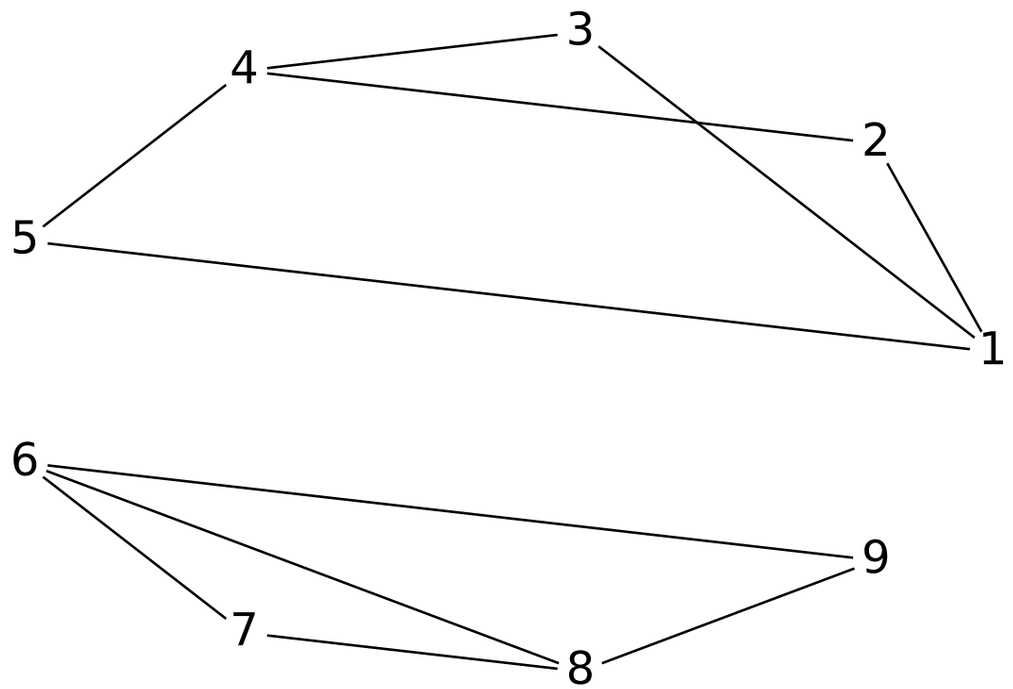
- A *walk* is a set of alternating vertices and edges that define a traversal through the graph
  - A walk is *closed* if the starting vertex is the same as the ending vertex, otherwise it is *open*
- A *path* is a particular type of open walk that that has no repeated edges or vertices



Path: (RB,AW,RP,OS)

Subgraphs and components

- A *subgraph* is a subset of vertices/edges in a graph
- A *component* is a subgraph in which each pair of vertices is connected by a path

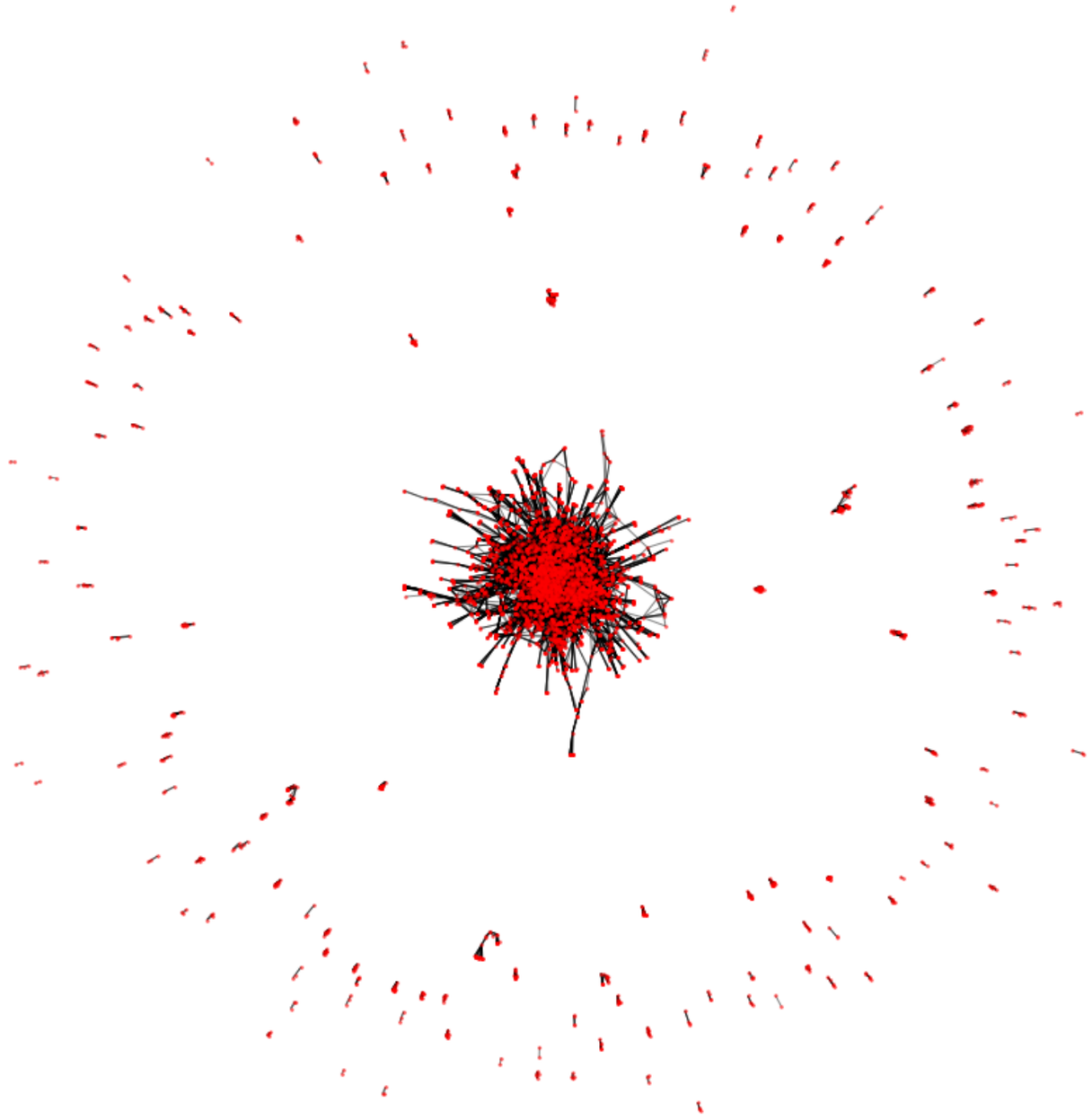



```
comp = nx.connected_component_subgraphs(G_2comp)
for i, c in enumerate(comp):
    print(i, c.nodes)
```

0	[1,	2,	3,	4,	5]
1	[8,	9,	6,	7]	

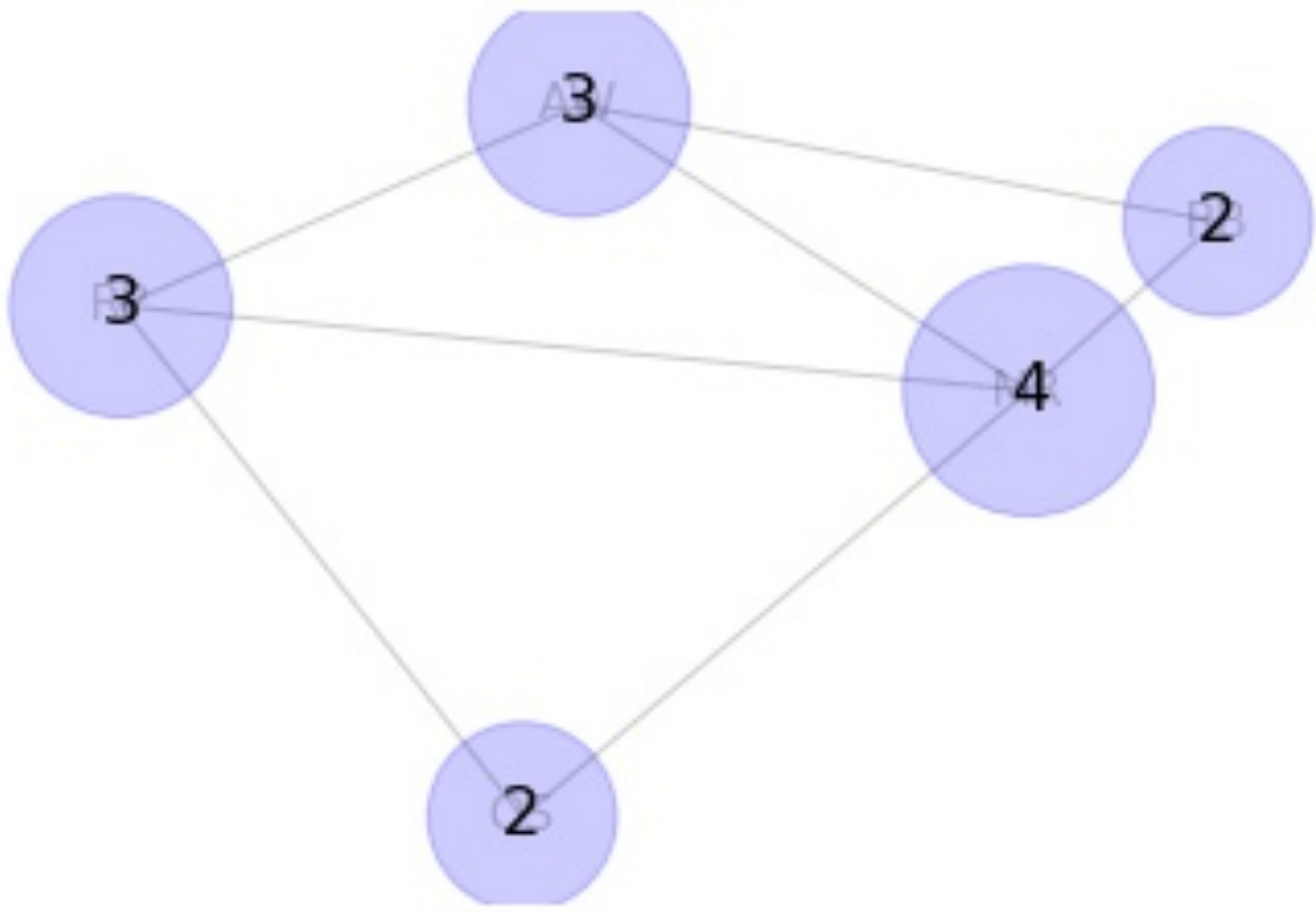
Giant component

- In many real-world networks there is one main component and then many smaller ones
- The main component is referred to as the “giant component”



Describing a graph: Node features

- Degree: how many edges does a node have?



Describing a graph as a matrix

- Adjacency matrix A (for an undirected graph)
  - A symmetric N x N matrix (where N is the number of vertices)

$$A_{i,j} = \begin{cases} 1 \; if \; (i,j) \; or \; (j,i) \; in \; E \\ 0 \; if \; (i,j) \; and \; (j,i) \; not \; in \; E \end{cases}$$

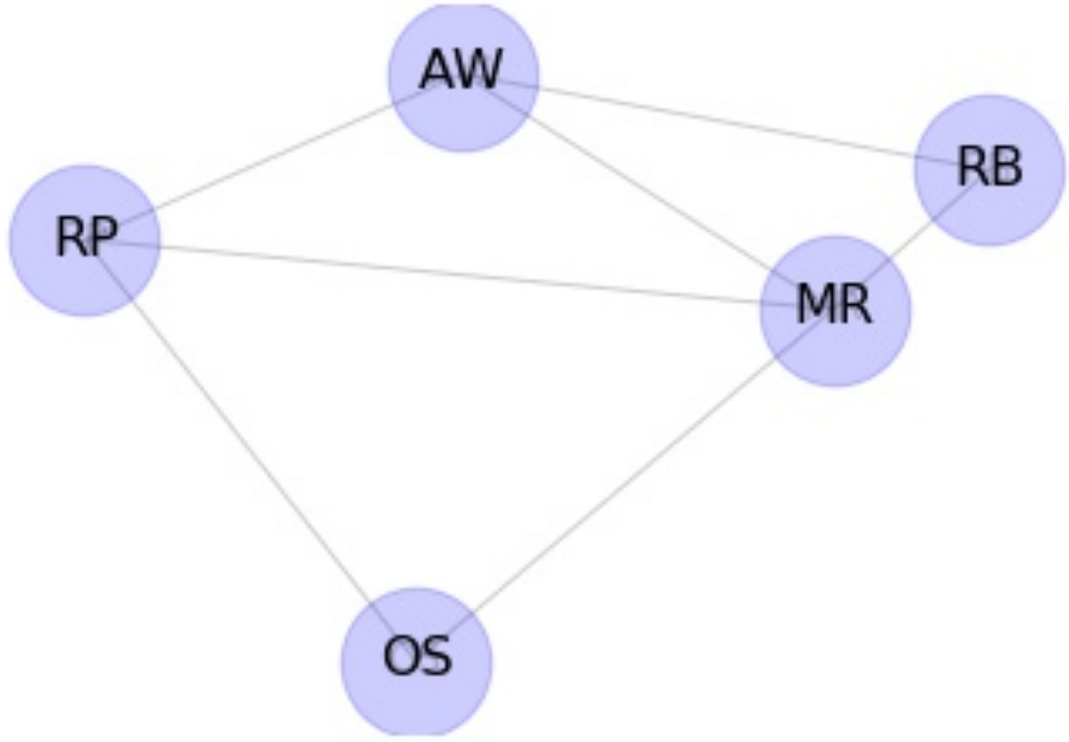
  - Can also be weighted:

$$A_{i,j} = \begin{cases} w_{i,j} \; if \; (i,j) \; or \; (j,i) \; in \; E \\ 0 \; if \; (i,j) \; and \; (j,i) \; not \; in \; E \end{cases}$$

Co-publication data

V: (‘RP’, 'AW', 'RB', 'MR', 'OS')

E = [(‘RP', 'AW'), ('RP', 'MR'), ('RP', 'OS'), ('AW', 'RB'), (‘AW’, 'MR'), ('RB', 'MR'), (‘MR', 'OS')]



Weighted

	RP	AW	RB	MR	OS
RP	0	7	0	2	1
AW	7	0	6	2	0
RB	0	6	0	22	0
MR	2	2	22	0	2
OS	1	0	0	2	0

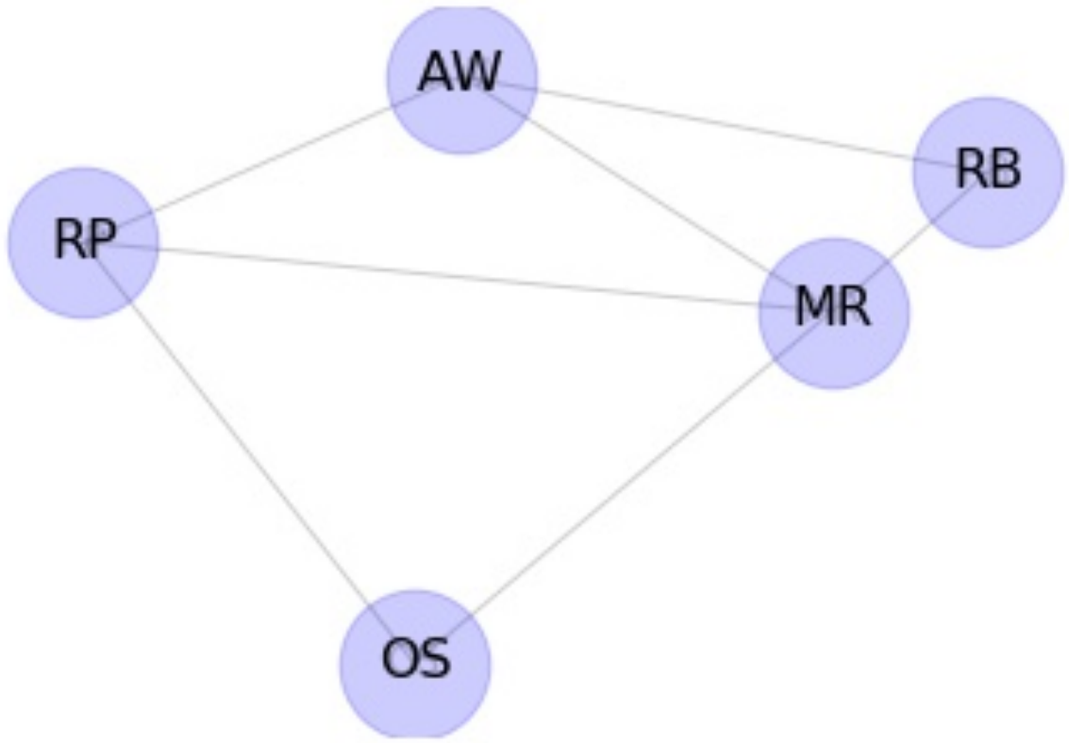
Binary

	RP	AW	RB	MR	OS
RP	0	1	0	1	1
AW	1	0	1	1	0
RB	0	1	0	1	0
MR	1	1	1	0	1
OS	1	0	0	1	0

Co-publication data

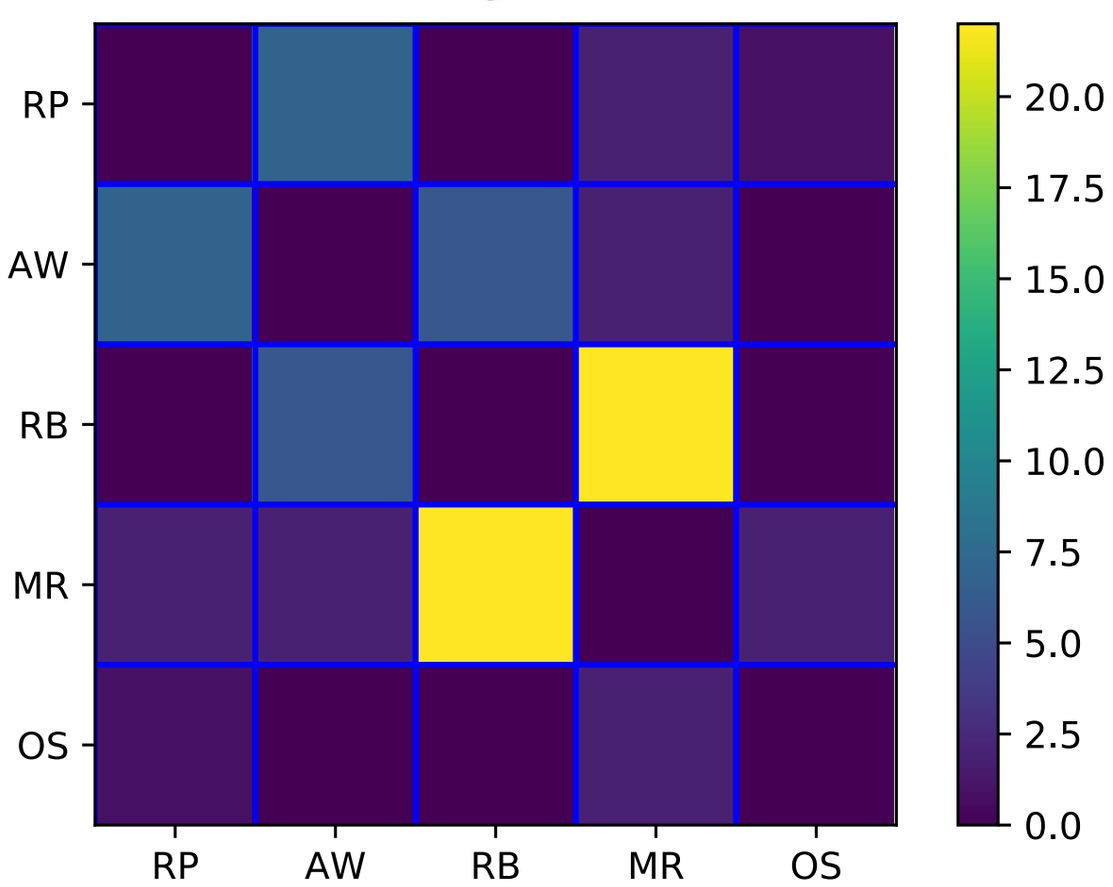
V: (‘RP’, 'AW', 'RB', 'MR', 'OS')

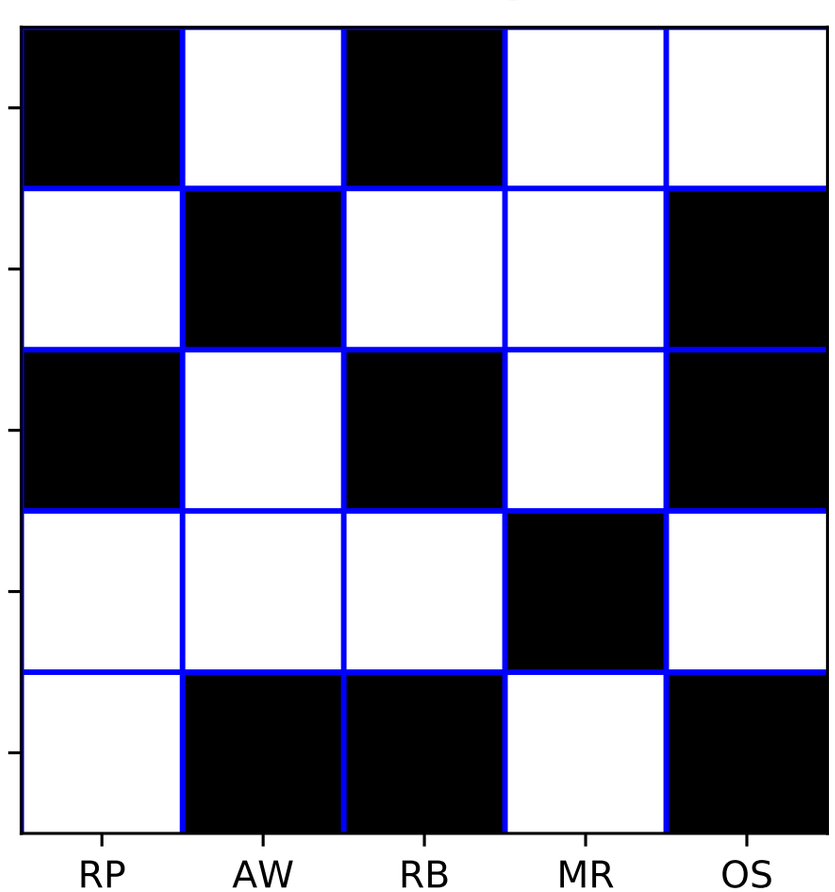
E = [(‘RP', 'AW'), ('RP', 'MR'), ('RP', 'OS'), ('AW', 'RB'), (‘AW’, 'MR'), ('RB', 'MR'), (‘MR', 'OS')]



Binary

Weighted

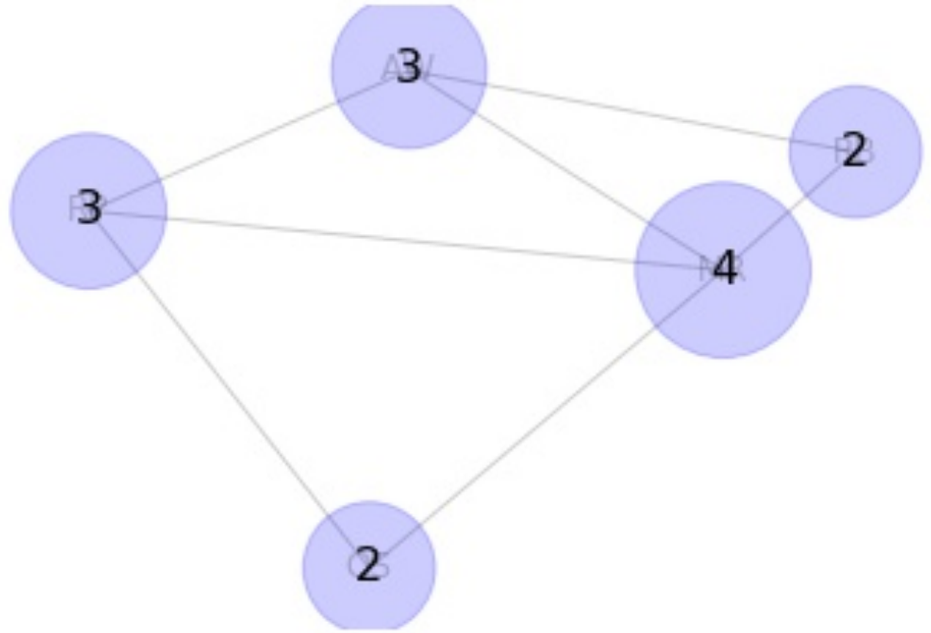




Computing graph degree

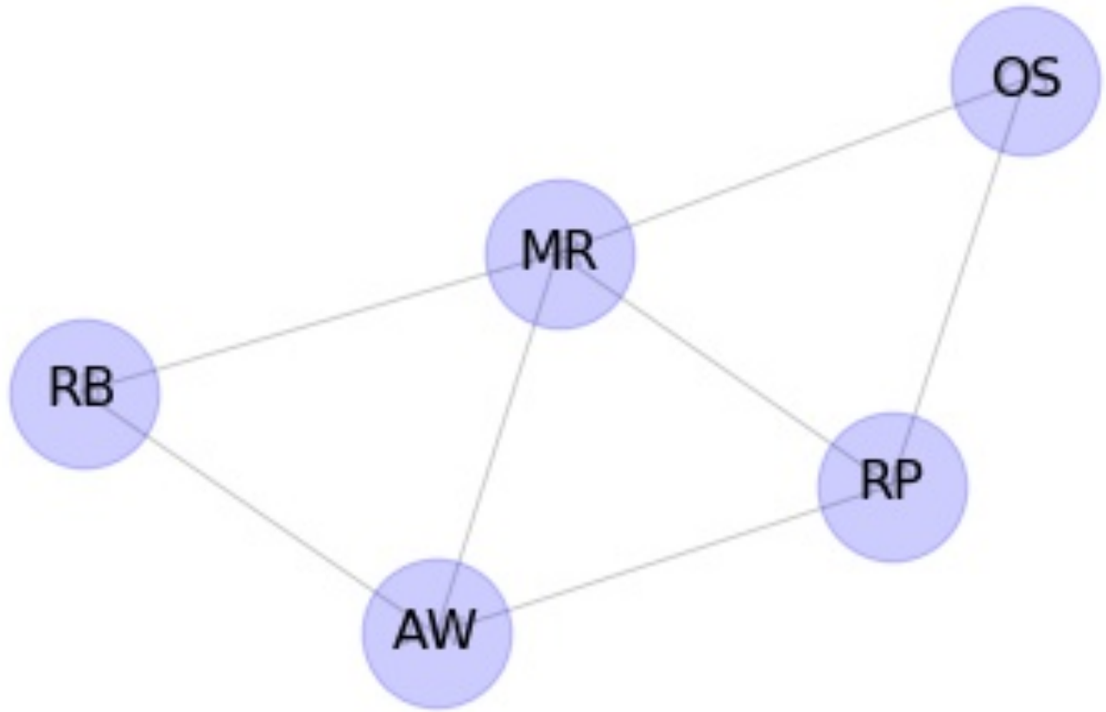
- Sum binarized adjacency matrix across rows (or columns)

	RP	AW	RB	MR	OS	degree
RP	0	1	0	1	1	3
AW	1	0	1	1	0	3
RB	0	1	0	1	0	2
MR	1	1	1	0	1	4
OS	1	0	0	1	0	2



Finding paths in a graph

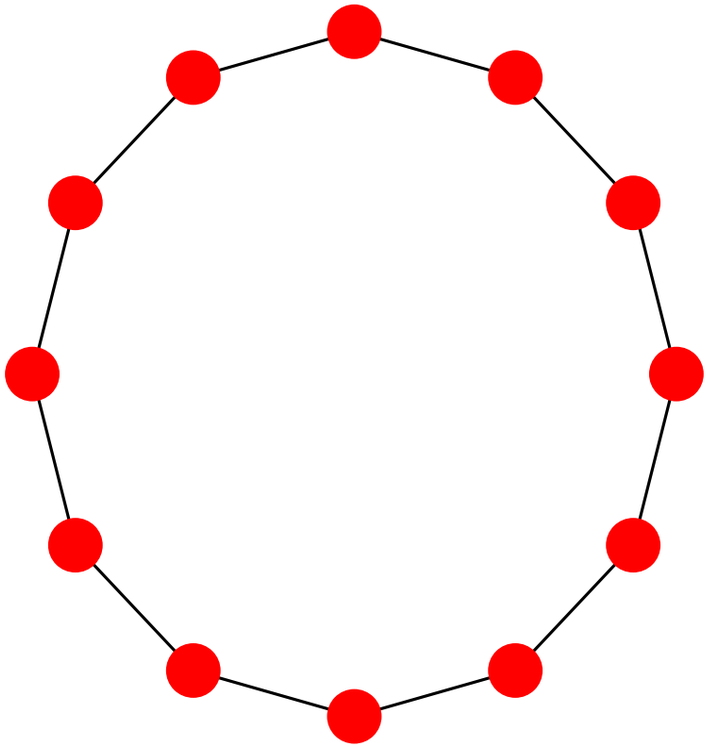
- What is the shortest path between RB and OS?
  - Easy by eye with a small graph
- Polynomial time in general
  - Dijkstra's algorithm (best-first search) is $O(V^2)$



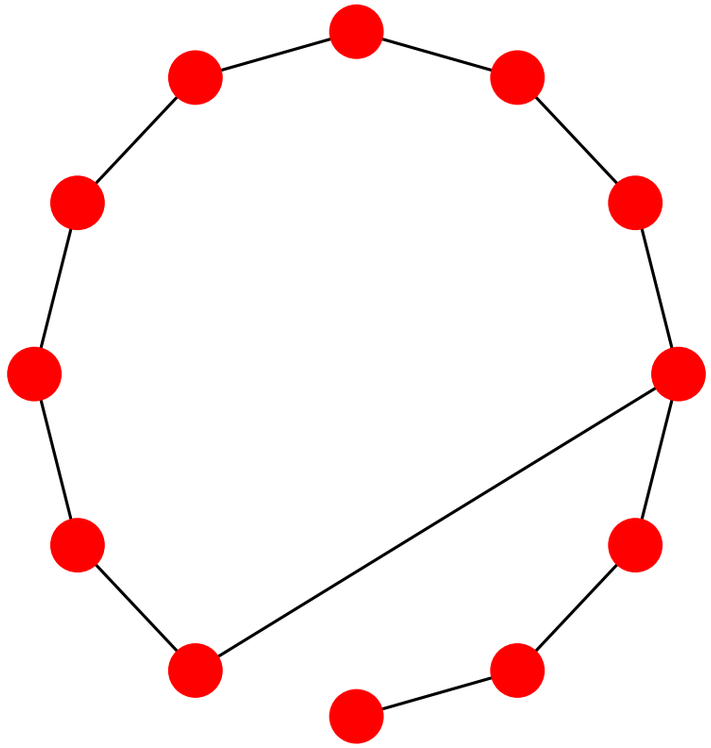
Average shortest path length

- How easy is it for nodes to communicate?

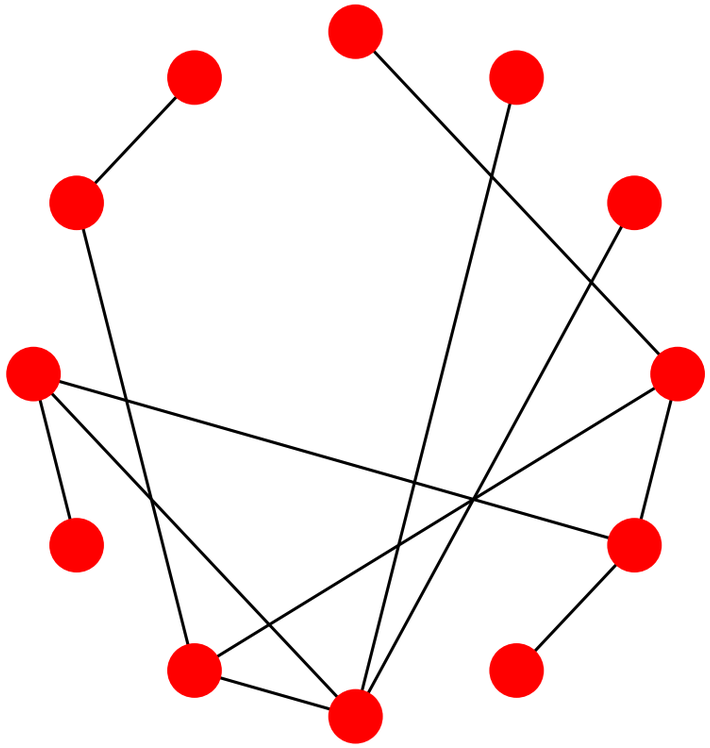
avg path length:3.273



avg path length:3.152



avg path length:2.591



Centrality

- What are the most important nodes in the graph?
  - Quantified by measures of *centrality*
- Two dimensions along with centrality measures vary (Borgatti & Everett, 2006)
  - How is the node involved?
    - Radial: focused on paths that start or end at a given node
    - Medial: focused on paths that traverse a given node
  - What is the quantity being assessed?
    - volume: how many paths?
    - length: how long are the paths?

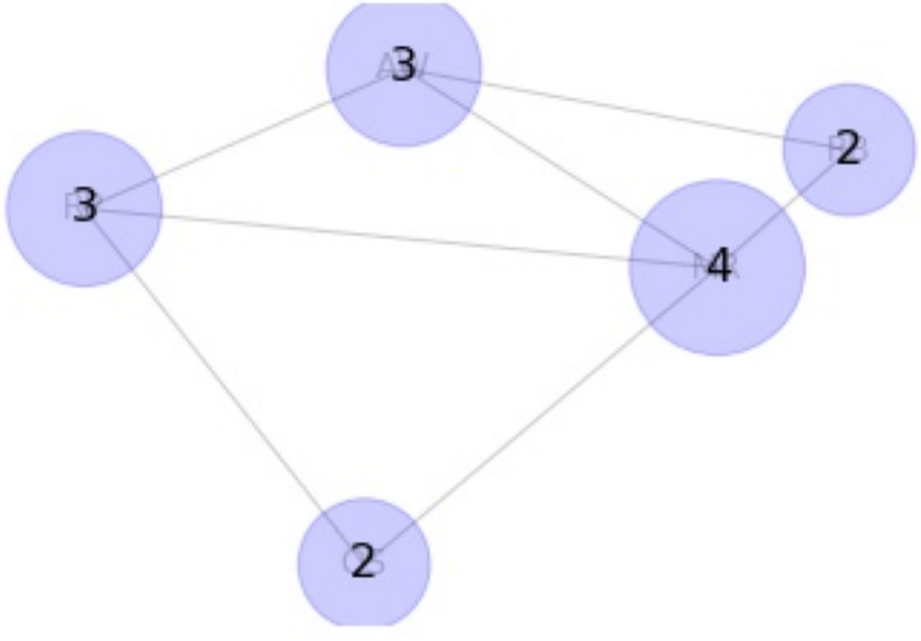
Degree centrality

- How many walks of length 1 (i.e. edges) are there from a given node?

Radial

Volume

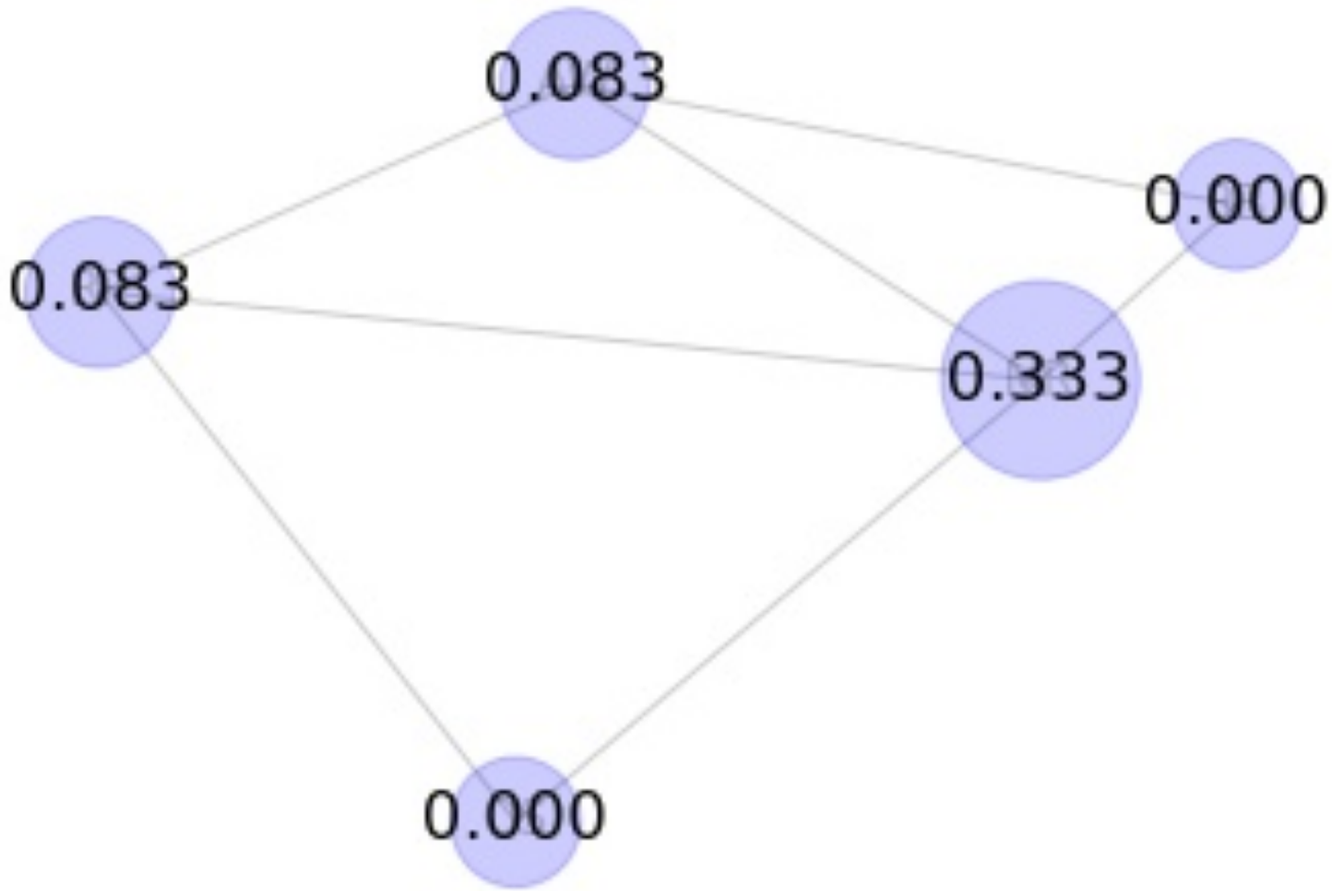
	RP	AW	RB	MR	OS	degree
RP	0	1	0	1	1	3
AW	1	0	1	1	0	3
RB	0	1	0	1	0	2
MR	1	1	1	0	1	4
OS	1	0	0	1	0	2



Betweenness centrality

- The number of shortest paths that traverse a given node

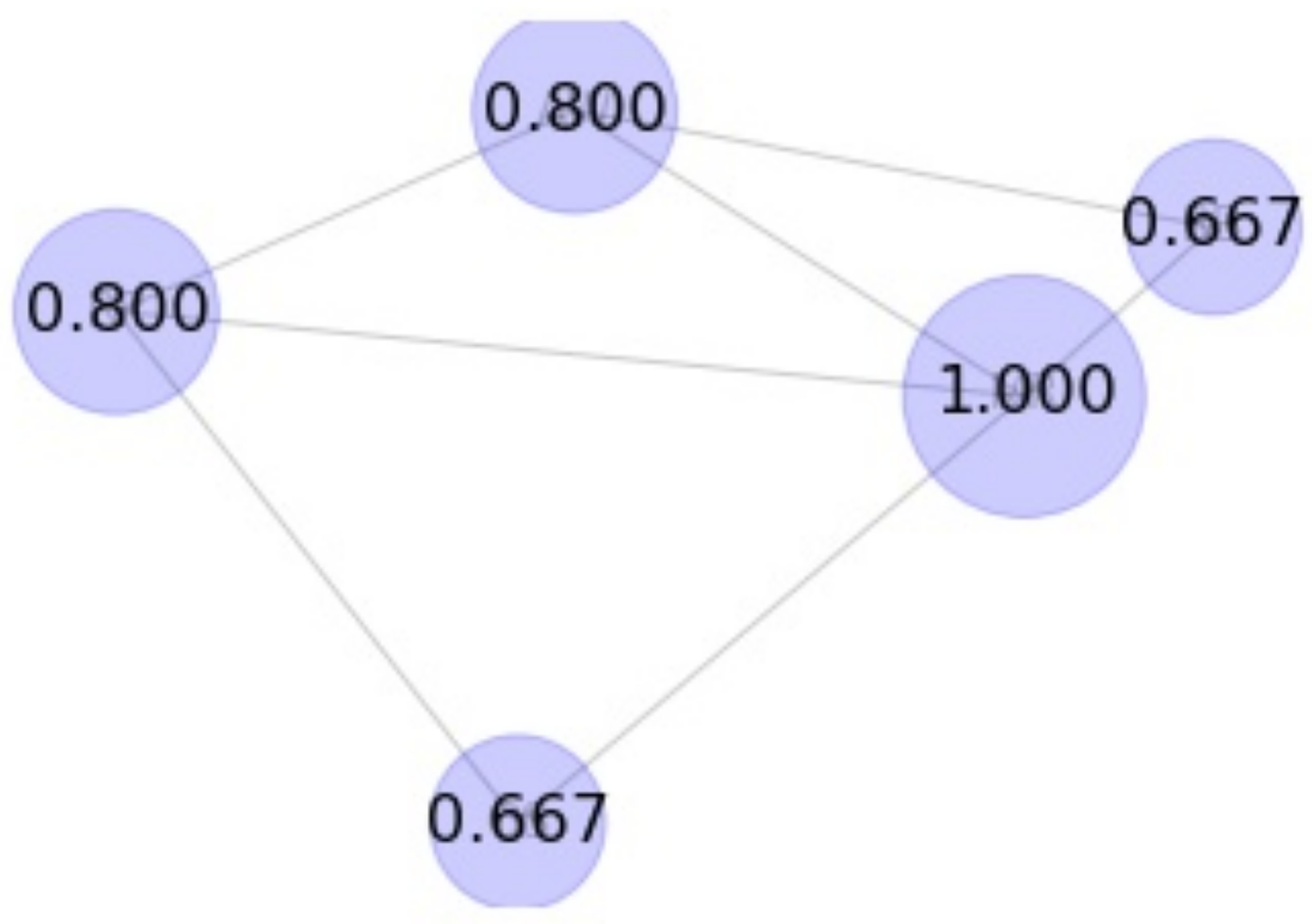
Medial
Volume



Closeness centrality

- Total distance from any vertex to all other vertices via shortest paths

Radial
Length

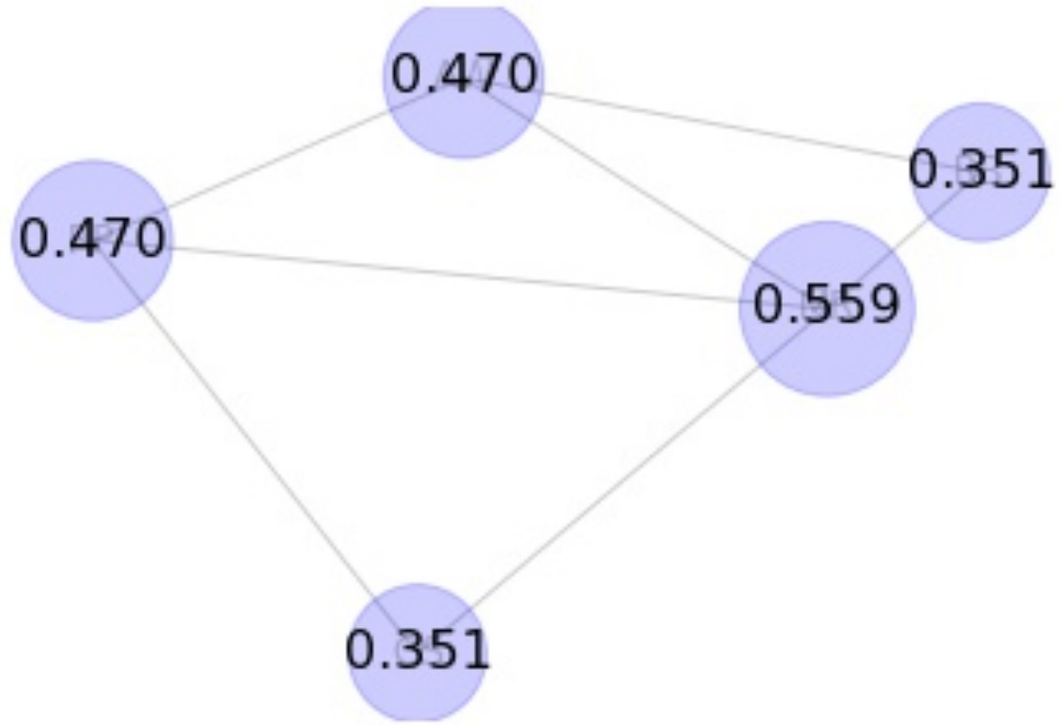


Eigenvector centrality

- How many walks (of all lengths) are there from a given node?
  - Like degree centrality, but weights by how connected the adjacent nodes are
- Idea underlying Google's PageRank algorithm
  - A web page can have lots of links, but if they are all from web pages that are not highly linked, then its Pagerank value will be low

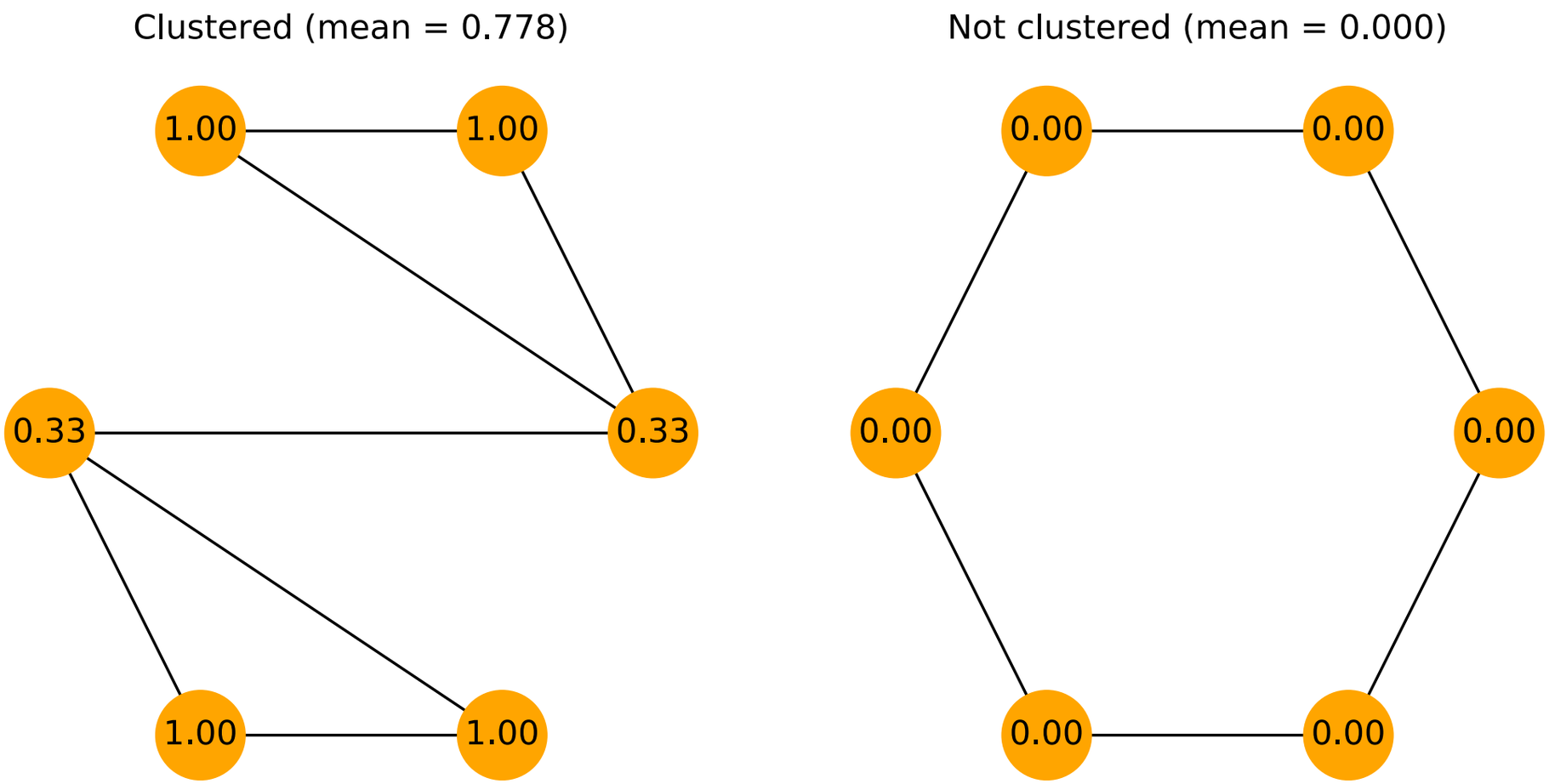
Radial

Volume



Beyond individual nodes: Clustering

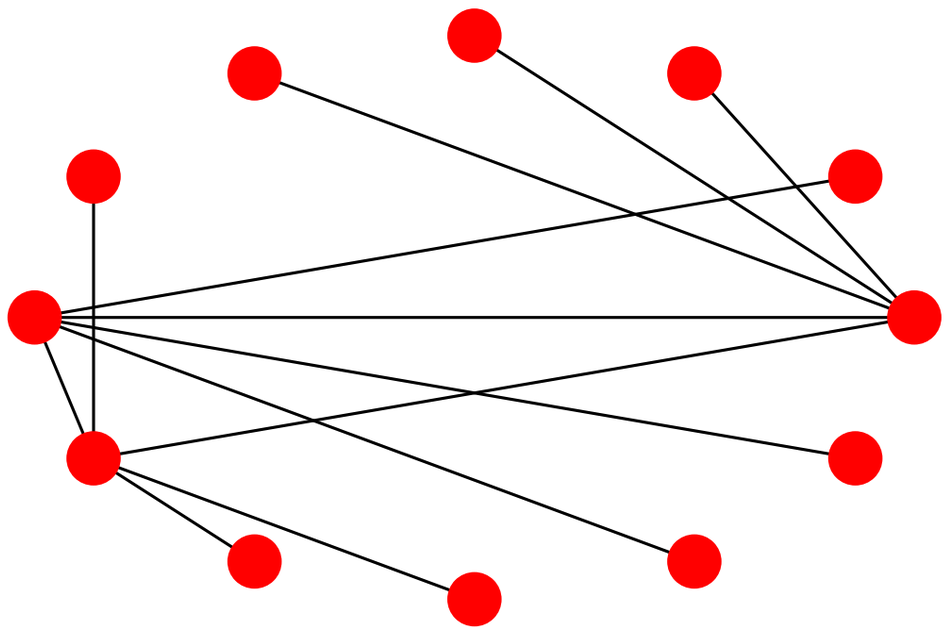
- A cluster is a set of nodes that are more heavily connected to each other than they are to other nodes around the network
  - Clustering coefficient: what proportion of a node's neighbors are connected to one another?



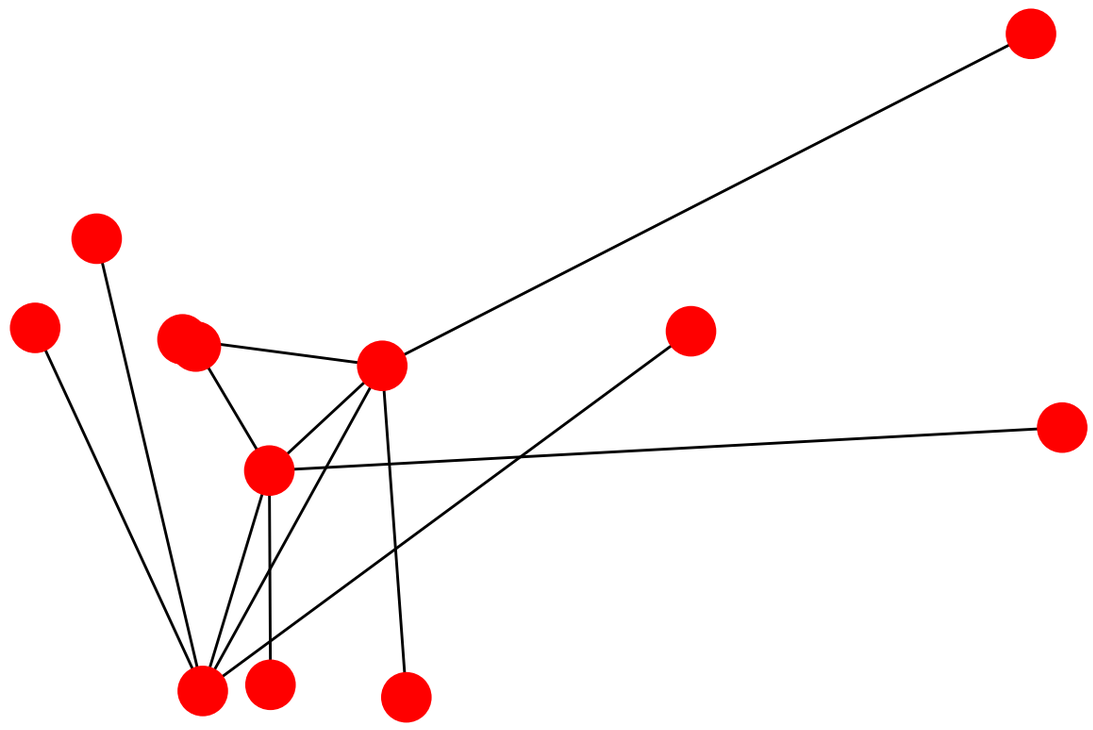
Visualizing graphs

- Project network onto a 2-d space
  - How do we position the nodes in 2-d?

Circular

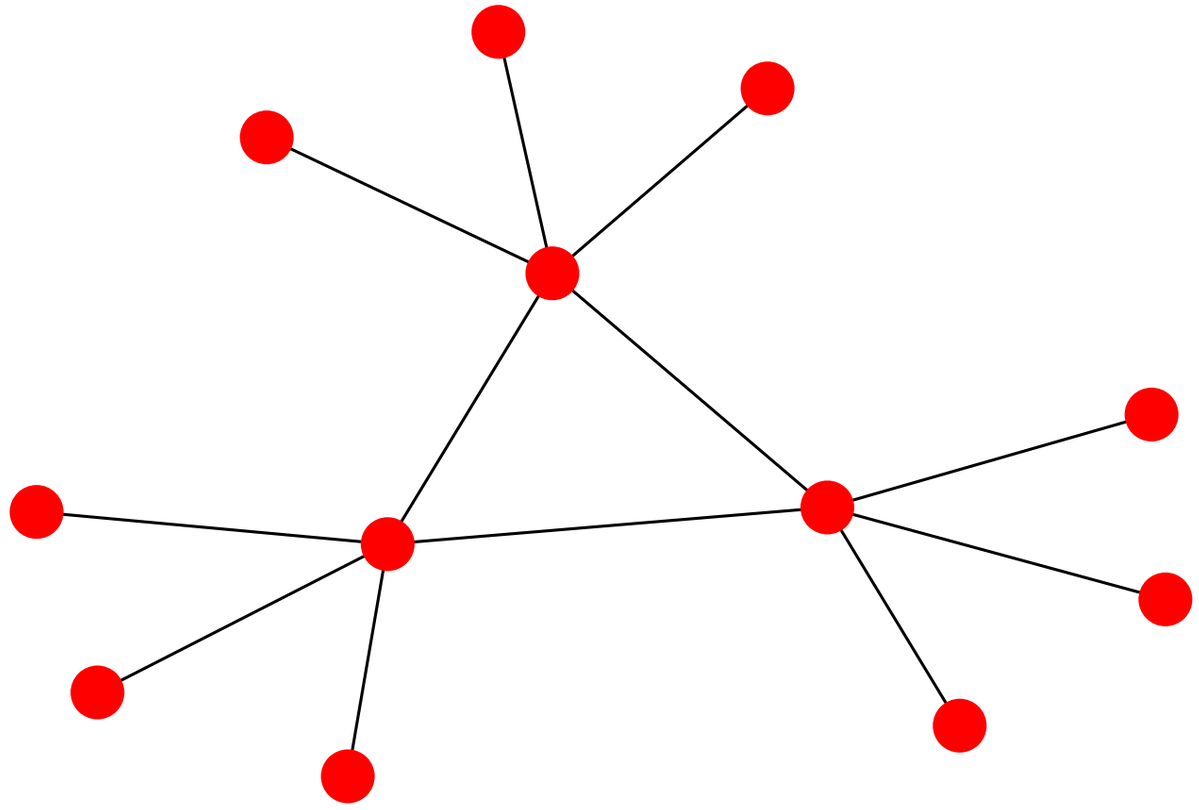


Random



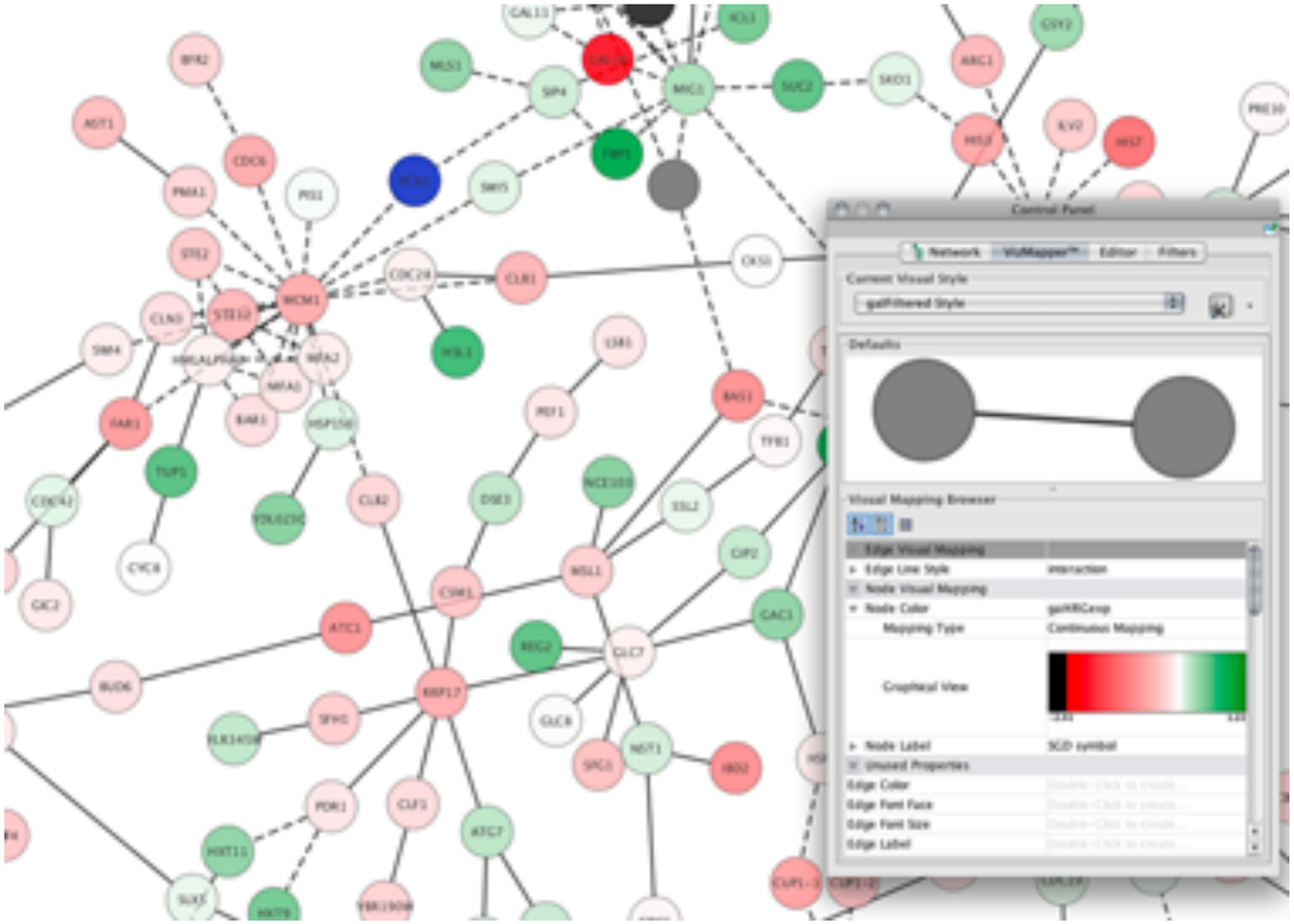
Force-directed graph embedding

- Nodes repel one another
- Edges attract by a spring-like force
- Algorithm minimizes total energy in the system



Tools for graph visualization

- NetworkX
  - Limited set of built-in visualization method
- Production-level tools
- Cytoscape
  - Very powerful open-source tool
  - Can import .graphml files saved by NetworkX



http://cytoscape.org/

Psych 253
Advanced Statistical Modeling

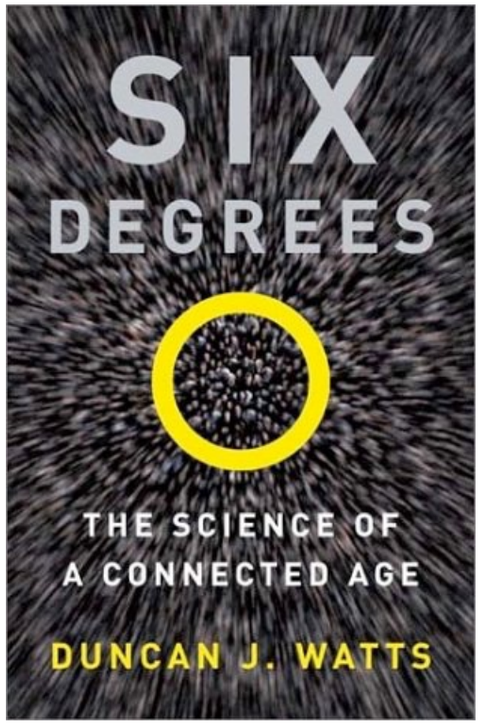
Network modeling

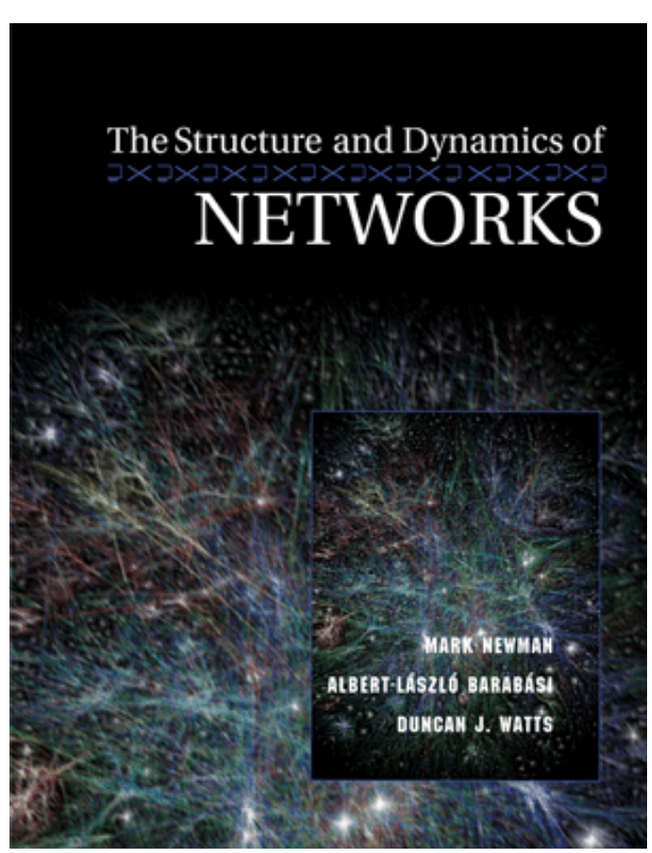
Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford Artificial Intelligence Laboratory
Stanford University

Russ Poldrack
Department of Psychology
Stanford University

Network science

- Emerged from early studies of social networks
- Grew quickly starting in the 1990's as physicists and mathematicians became involved
- Focused on the features of complex real-world networks





Community structure

- In most complex networks, there are sets of nodes that are more densely connected with each other than with other nodes

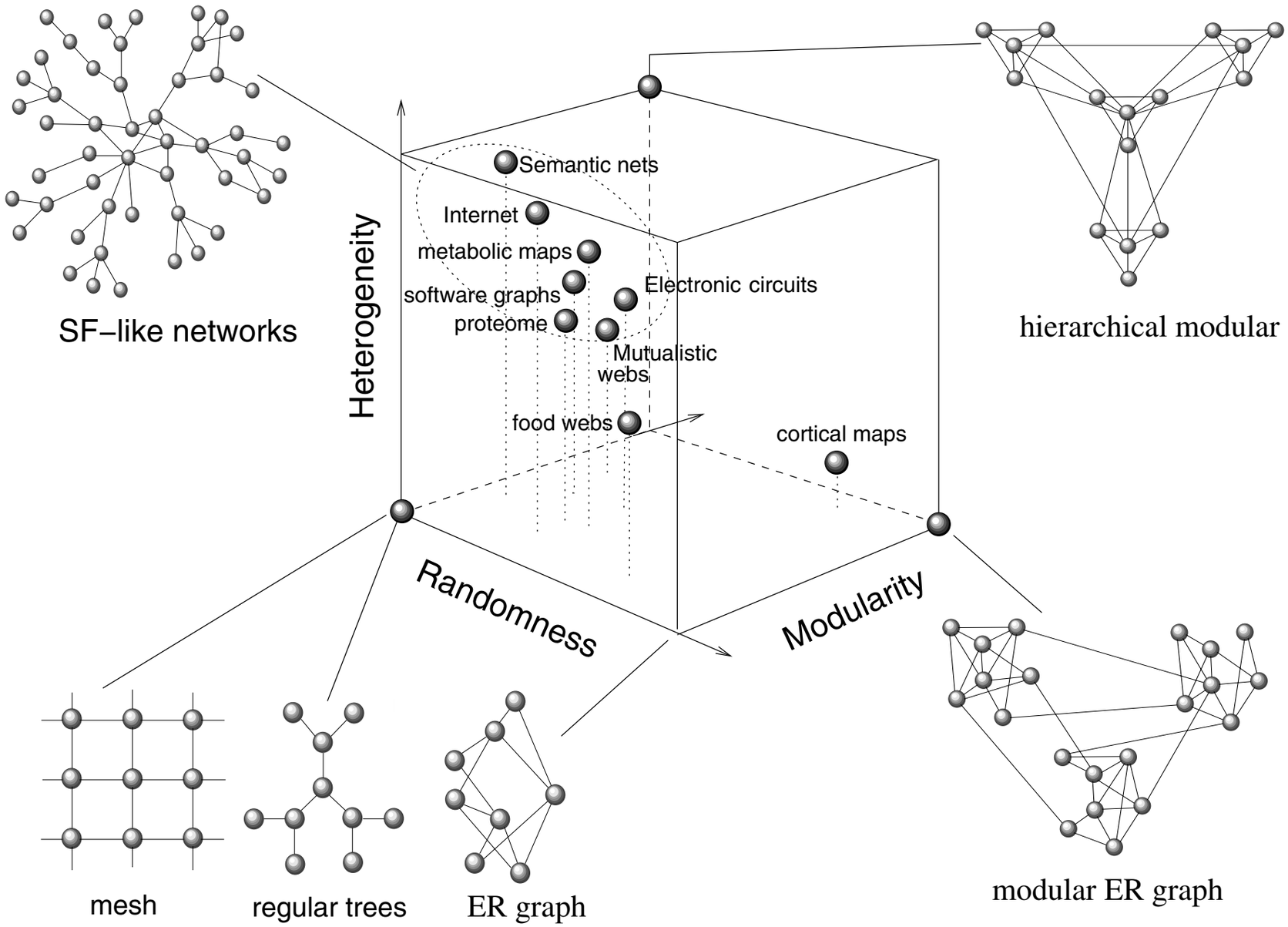
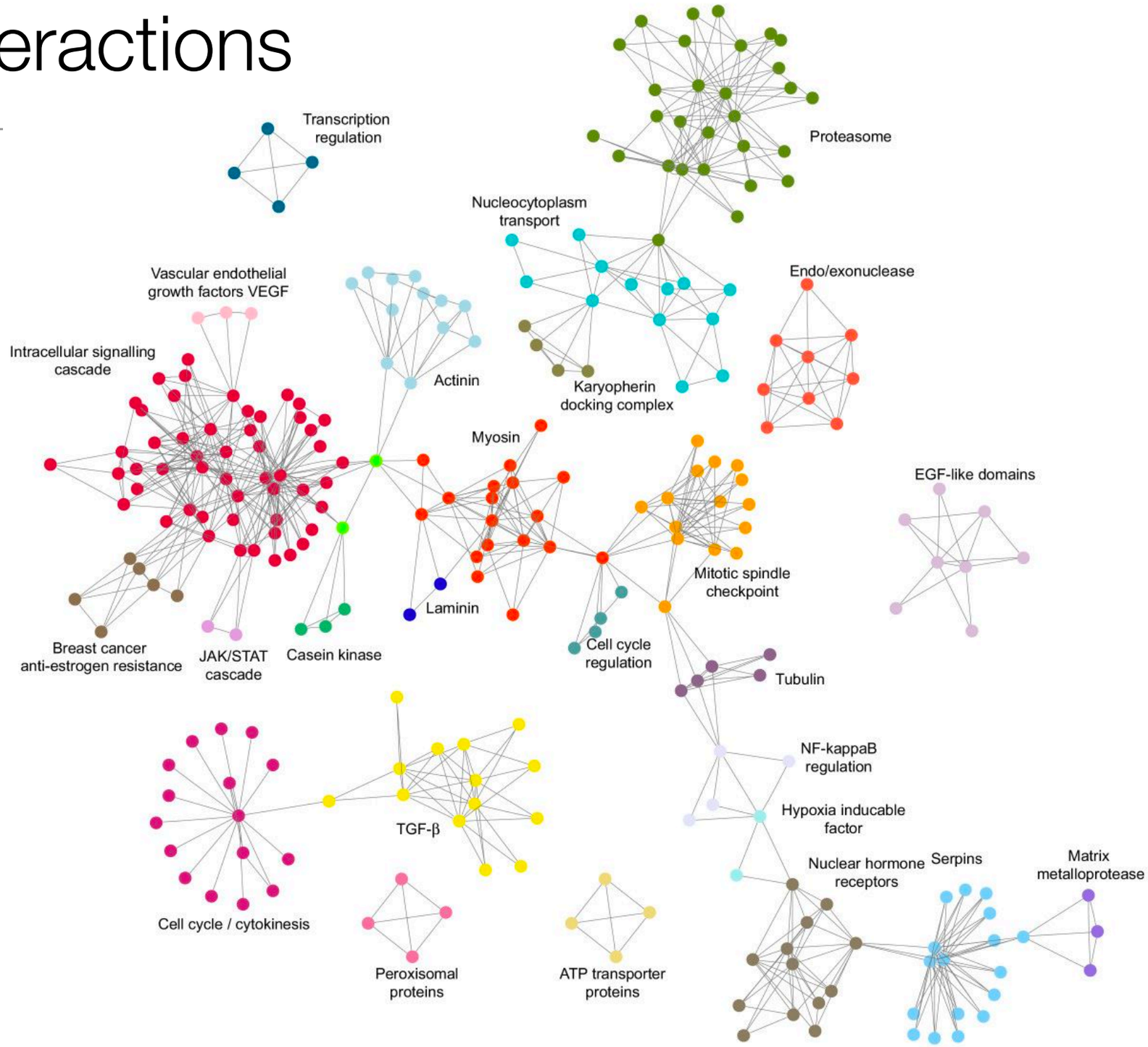


Fig. 3. A zoo of complex networks. In this qualitative space, three relevant characteristics are included: randomness, heterogeneity and modularity. The first introduces the amount of randomness involved in the process of network's building. The second measures how diverse is the link distribution and the third would measure how modular is the architecture. The position of different examples are only a visual guide. The domain of highly heterogeneous, random hierarchical networks appears much more occupied than others. Scale-free like networks belong to this domain.

Sola & Valverde, 2004

Protein-protein interactions

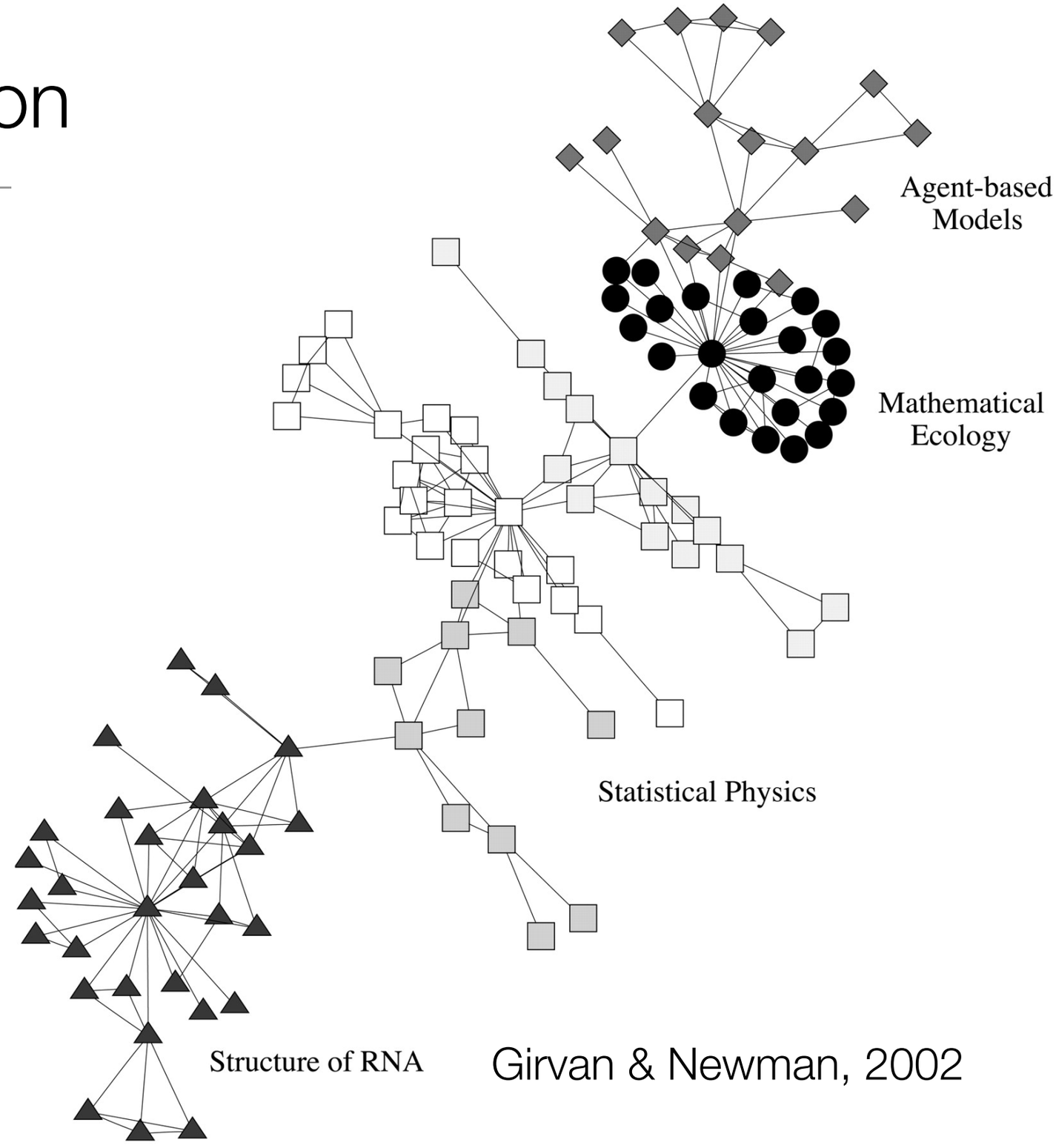
- Inferred interactions by homology to known interacting proteins, and scored based on supporting data
- Community detection identified known functional pathways



Jonsson et al., 2006

Scientific collaboration

- Vertices: 118 scientists at Santa Fe Institute
- Edges: coauthorship
- Community detection identifies researchers working on different topics
- Most communities revolve around a central member



Girvan & Newman, 2002

What is community structure?

- Connectivity within communities is stronger than connectivity between communities
- Similar to clustering methods
  - E.g. Ward's method minimizes within-cluster variance

Why is community structure important?

- Herb Simon (1962)
  - Makes it easier to build a complex system
    - The modules can be built independently and then assembled into a larger whole
  - Makes it easier to describe a system (if we find the correct representation)

THE ARCHITECTURE OF COMPLEXITY

HERBERT A. SIMON*

Professor of Administration, Carnegie Institute of Technology

(*Read April 26, 1962*)

A NUMBER of proposals have been advanced in recent years for the development of "general systems theory" which, abstracting from properties peculiar to physical, biological, or social systems, would be applicable to all of them.[1] We might well feel that, while the goal is laudable, systems of such diverse kinds could hardly be expected to have any nontrivial properties in common. Metaphor and analogy can be helpful, or they can be misleading. All depends on whether the similarities the metaphor captures are significant or superficial.

It may not be entirely vain, however, to search for common properties among diverse kinds of complex systems. The ideas that go by the name of cybernetics constitute, if not a theory, at least a point of view that has been proving fruitful over a wide range of applications.[2] It has been useful to look at the behavior of adaptive systems in terms of the concepts of feedback and homeostasis, and to analyze adaptiveness in terms of the theory of selective information.[3] The ideas of feedback and information provide a frame of reference for viewing a wide range of situations, just as do the ideas of evolution, of relativism, of axiomatic method, and of operationalism.

In this paper I should like to report on some things we have been learning about particular kinds of complex systems encountered in the behavioral sciences. The developments I shall discuss arose in the context of specific phenomena, but the theoretical formulations themselves make little reference to details of structure. Instead they refer primarily to the complexity of the systems under view without specifying the exact content of that complexity. Because of their abstractness, the theories may have relevance—application would be too strong a term—to other kinds of complex systems that are observed in the social, biological, and physical sciences.

In recounting these developments, I shall avoid technical detail, which can generally be found elsewhere. I shall describe each theory in the particular context in which it arose. Then, I shall cite some examples of complex systems, from areas of science other than the initial application, to which the theoretical framework appears relevant. In doing so, I shall make reference to areas of knowledge where I am not expert—perhaps not even literate. I feel quite comfortable in doing so before the members of this society, representing as it does the whole span of the scientific and scholarly endeavor. Collectively you will have little difficulty, I am sure, in distinguishing instances based on idle fancy or sheer ignorance from instances that cast some light on the ways in which complexity exhibits itself wherever it is found in nature. I shall leave to you the final judgment of relevance in your respective fields.

I shall not undertake a formal definition of

* The ideas in this paper have been the topic of many conversations with my colleague, Allen Newell. George W. Corner suggested important improvements in biological content as well as editorial form. I am also indebted, for valuable comments on the manuscript, to Richard H. Meier, John R. Platt, and Warren Weaver. Some of the conjectures about the nearly decomposable structure of the nucleus-atom-molecule hierarchy were checked against the available quantitative data by Andrew Schoene and William Wise. My work in this area has been supported by a Ford Foundation grant for research in organizations and a Carnegie Corporation grant for research on cognitive processes. To all of the above, my warm thanks, and the usual absolution.

[1] See especially the yearbooks of the Society for General Systems Research. Prominent among the exponents of general systems theory are L. von Bertalanffy, K. Boulding, R. W. Gerard, and J. G. Miller. For a more skeptical view—perhaps too skeptical in the light of the present discussion—see H. A. Simon and A. Newell, Models: their uses and limitations, *in* L. D. White, ed., *The state of the social sciences,* 66–83, Chicago, Univ. of Chicago Press, 1956.

[2] N. Wiener, *Cybernetics,* New York, John Wiley & Sons, 1948. For an imaginative forerunner, see A. J. Lotka, *Elements of mathematical biology,* New York, Dover Publications, 1951, first published in 1924 as *Elements of physical biology.*

[3] C. Shannon and W. Weaver, *The mathematical theory of communication,* Urbana, Univ. of Illinois Press, 1949; W. R. Ashby, *Design for a brain,* New York, John Wiley & Sons, 1952.

Defining communities based on vertex similarity

- Use similarity/distance of vertices based on some property (e.g. topological distance)
  - Many techniques
    - Cluster methods discussed earlier in the course

Defining communities based on null models

- Focused on the entire graph
  - How does the structure of the graph relate to a *null model* on some feature of interest?
  - Example null model:
    - random graph with same degree distribution or degree sequence

Defining a “good” community partition

- Modularity (Girvan & Newman, 2004)

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

$$P_{ij} = \frac{k_i k_j}{2m}$$

m: # of edges
A: binary adjacency matrix
P: expected probability of edges under null model
C: partition assignment

Null model with no community structure but same degree sequence (on average)

$$\delta(C_i, C_j) = \begin{cases} 1 \ if \ C_i = C_j \\ 0 \ if \ C_i \neq C_j \end{cases}$$

Identifying community structure

- General idea
  - Minimize some particular loss function to optimize "good" community structure

Heuristic solutions for community detection

- Community detection is in general a hard problem
  - Many are **NP**-hard - unlikely to have a solution in polynomial time
- Most common algorithms are approximate
  - Can sometimes prove that they approximate the best solution
  - But they work much faster

Common assumptions

- Sparse graph
  - In general, # edges should be ≤ # of potential edges
- Unique partition
  - Each node is a member of only a single community
- Type of graph
  - We will focus on undirected binary (unweighted) graphs
  - Many techniques can be extended to directed or weighted graphs

Modularity maximization

- It's not possible to exhaustively test all possible partitions, even for a small network
- Most common algorithms use a greedy search
  - Start with each vertex in its own community
  - Find the next pairing that most increases modularity
  - Continue until no further increases are possible

Modularity maximization: Louvain algorithm

1. Each node assigned to its own community
2. For each node, assess change in modularity for changing the its community to each possible community of its neighbors
3. Combine each community into a single node, and do again at second level
4. Repeat until modularity does not improve

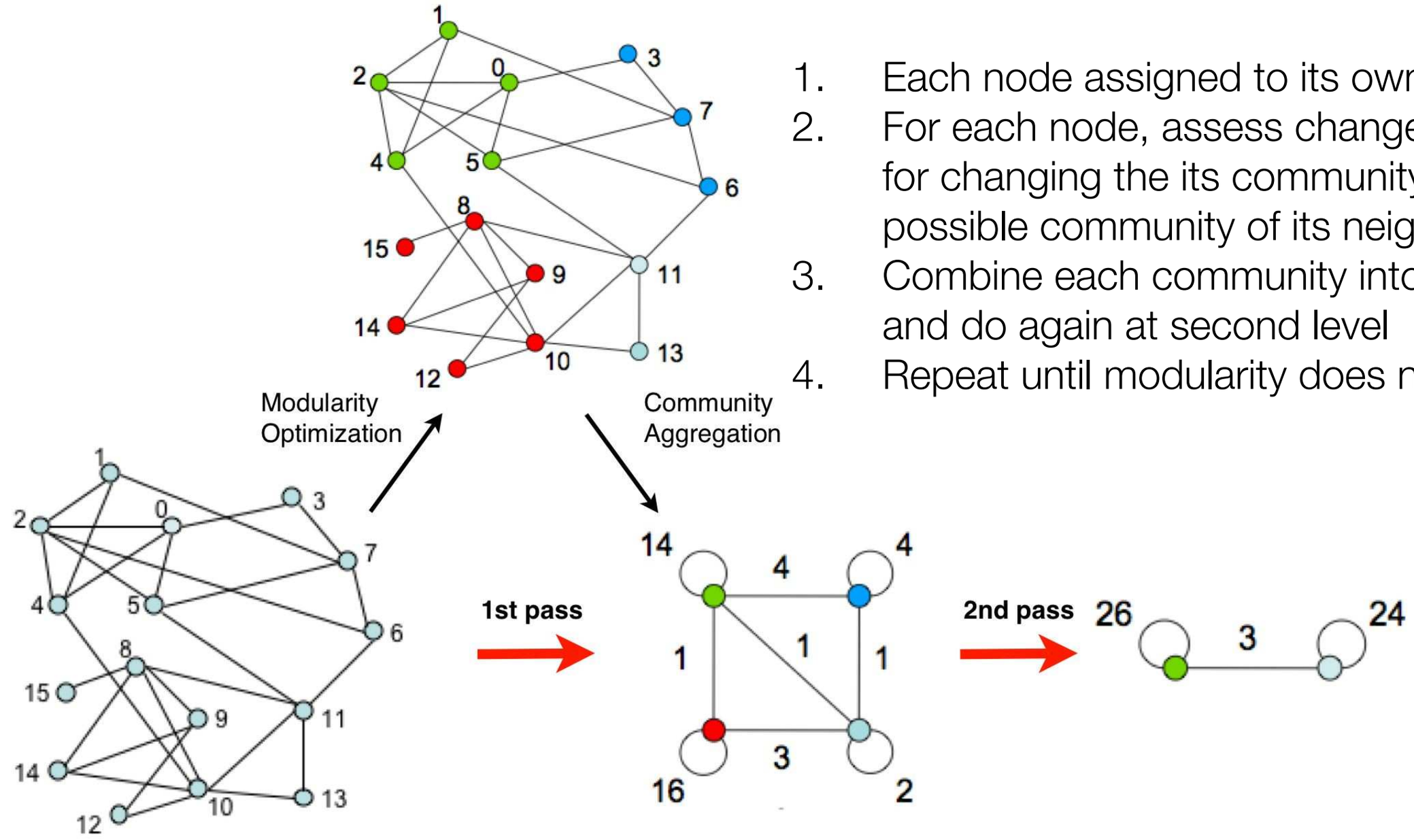


FIG. 12 Hierarchical optimization of modularity by Blondel et al. (Blondel *et al.*, 2008). The diagram shows two iterations of the method, starting from the graph on the left. Each iteration consists of a step, in which every vertex is assigned to the (local) cluster that produces the largest modularity increase, followed by a successive transformation of the clusters into vertices of a smaller (weighted) graph, representing the next higher hierarchical level. Reprinted figure with permission from Ref. (Blondel *et al.*, 2008). ©2008 by IOP Publishing and SISSA.

from Fortunato, 2010

Problems with modularity maximization

- Modularity landscapes can be relatively flat
  - i.e. there can be a large number of possible partitions with roughly equivalent modularity
-

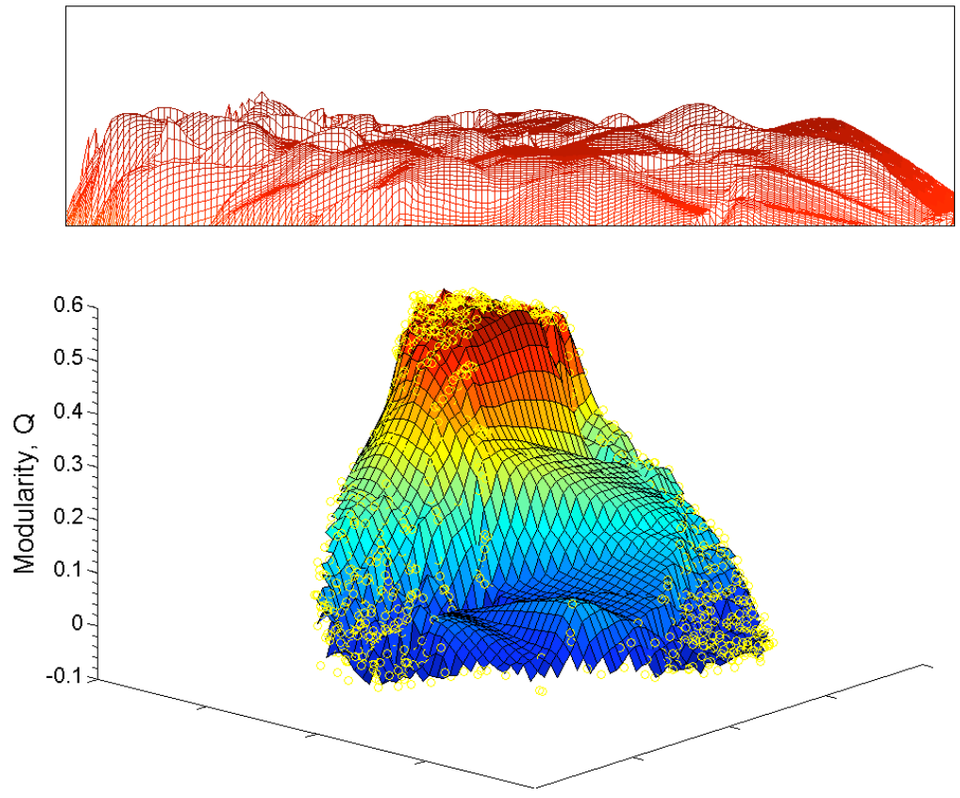


FIG. 16 Low-dimensional visualization of the modularity landscape for the metabolic network of the spirochete *Treponema pallidum*. The big degeneracy of suboptimal high-modularity partitions is revealed by the plateau (whose shape is detailed in the inset), which is large and very irregular. Modularity values in the plateau are very close to the absolute maximum, although they may correspond to quite different partitions. Reprinted figure with permission from Ref. (Good *et al.*, 2009).

from Fortunato, 2010

Problems with modularity maximization

- Modularity landscapes can be relatively flat
  - i.e. there can be a large number of possible partitions with roughly equivalent modularity
- Resolution limit
  - Find the partition with maximum modularity
- Below some limit, it is not possible to say whether the clusters are single communities or whether they are combinations of smaller communities
- i.e. modularity optimization is biased towards finding larger communities

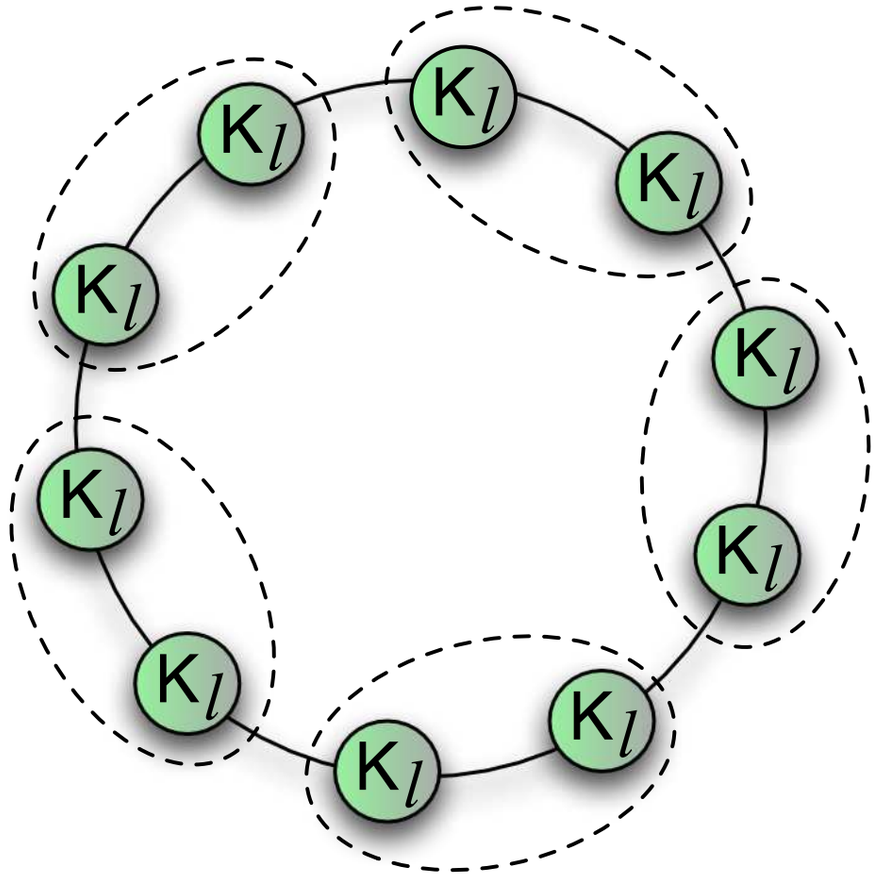


FIG. 15 Resolution limit of modularity optimization. The natural community structure of the graph, represented by the individual cliques (circles), is not recognized by optimizing modularity, if the cliques are smaller than a scale depending on the size of the graph. In this case, the maximum modularity corresponds to a partition whose clusters include two or more cliques (like the groups indicated by the dashed contours). Reprinted figure with permission from Ref. (Fortunato and Barthélemy, 2007). ©2007 from the National Academy of Science of the USA.

from Fortunato, 2010