Introduction to Networks

Amanda Perofsky

Introduction to Biological Statistics Workshop

November 21, 2014

Overview

Mathematics of networks

The large-scale structure of networks

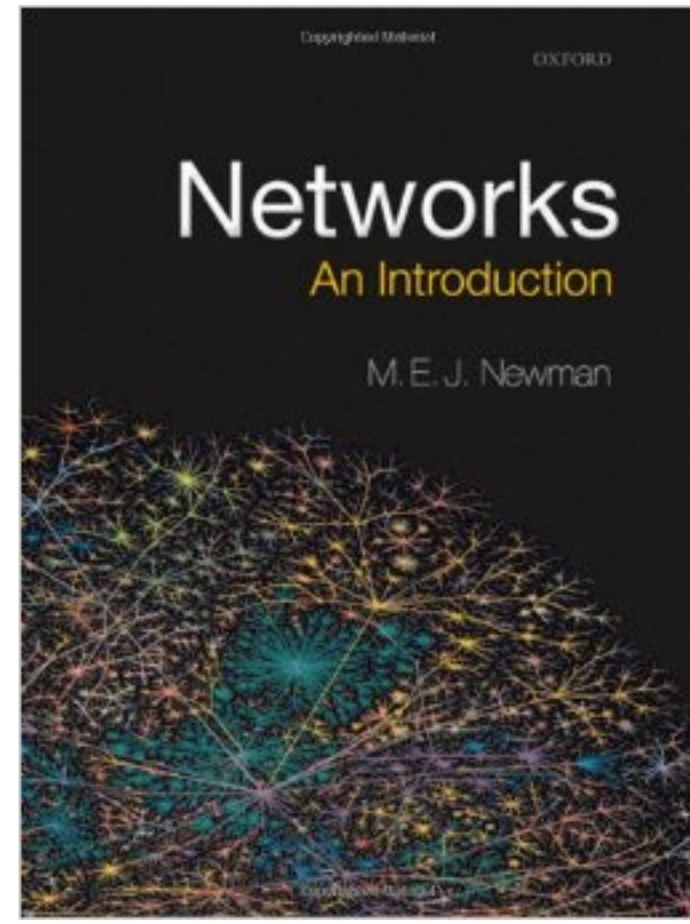
Network models

Centrality measures

R package: “igraph”

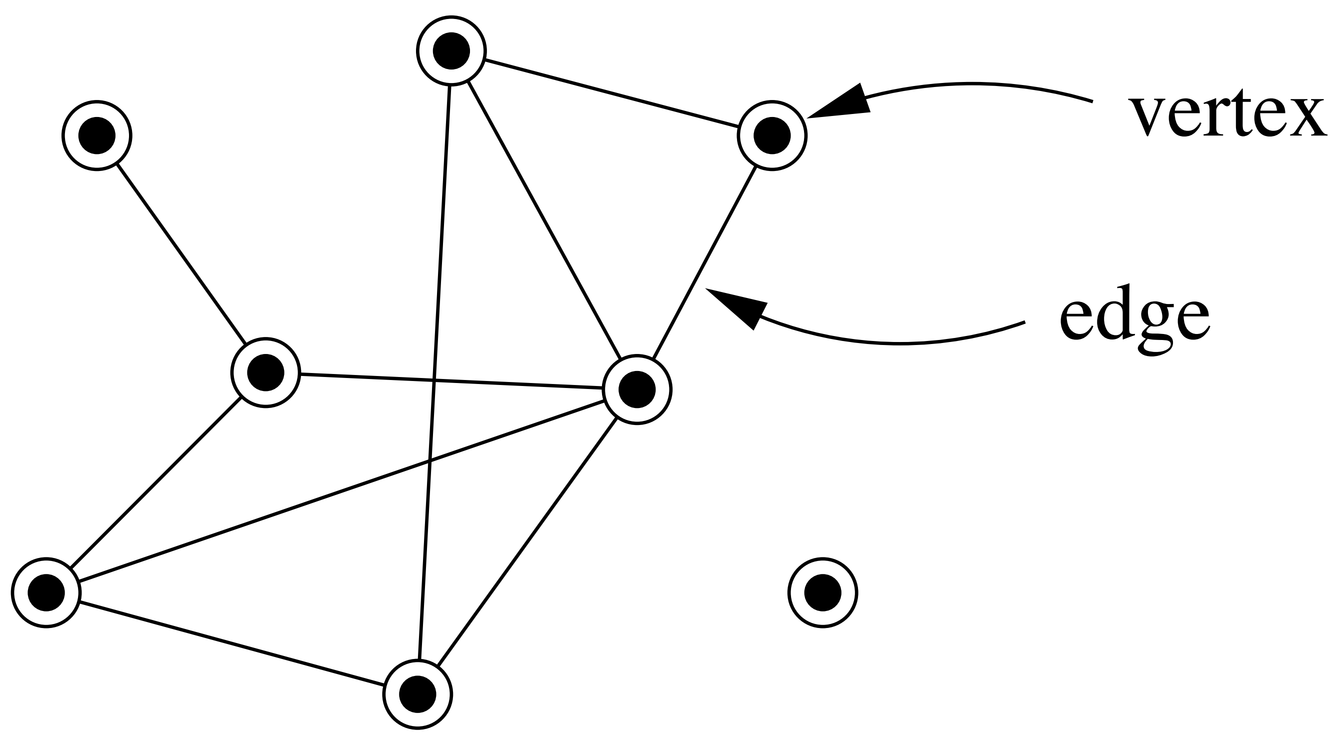
Acknowledgements

Dr. Lauren Ancel Meyers
Professor of Integrative Biology



Networks

A **network** is a collection of points, which we refer to as **vertices** or **nodes**, with connections between them, called **edges**.



In mathematics, these are called **graphs.**

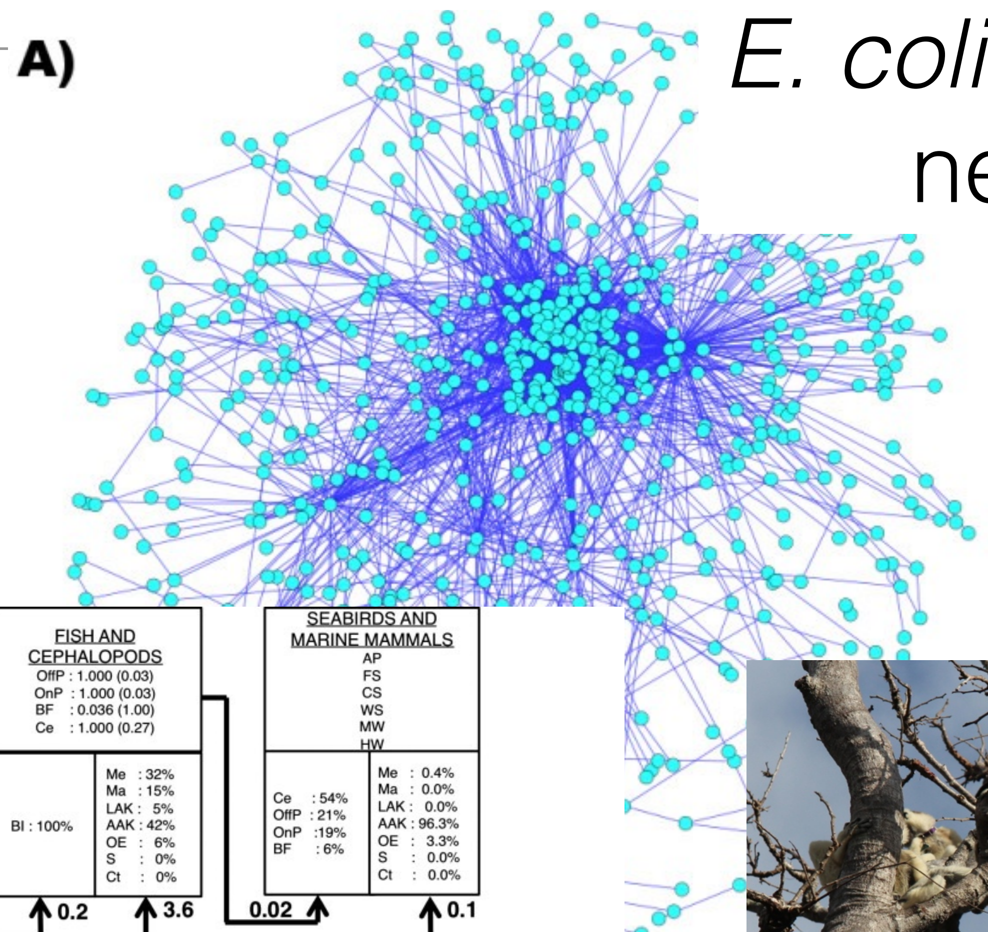
Why networks?

General yet powerful means of representing patterns of connections between the parts of a system

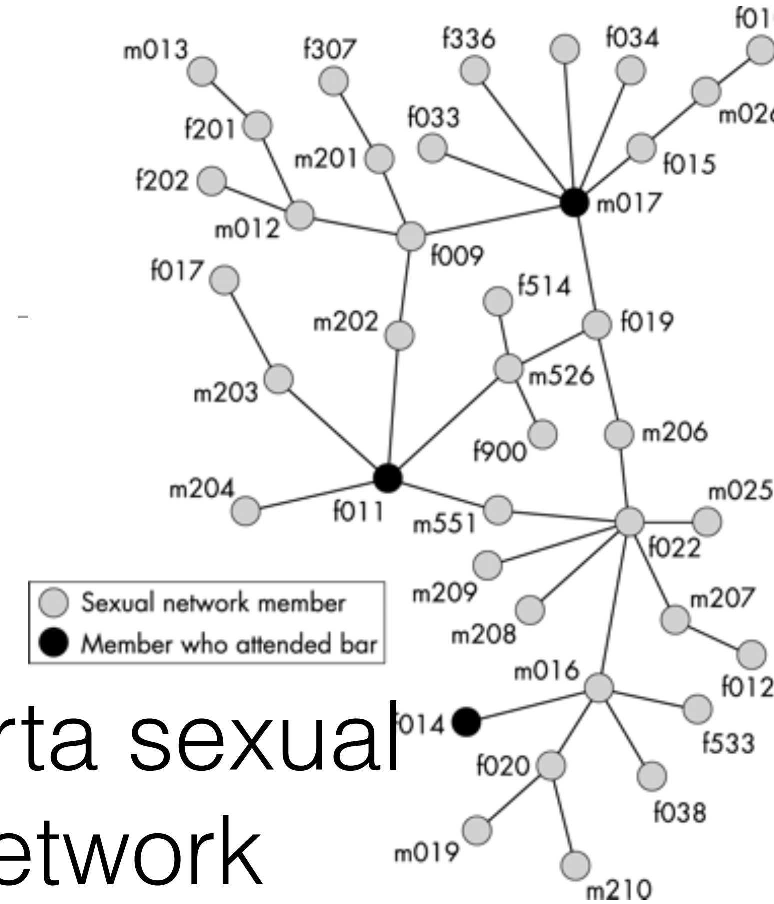
Mathematical, computational, and statistical framework for studying scientific systems:

- Can statistically characterize the structure of systems
- Can use models in an effort to understand how network properties arise in the first place
- Can add network processes on top of models of network structure to examine the interplay between structure and dynamics and to predict system behavior

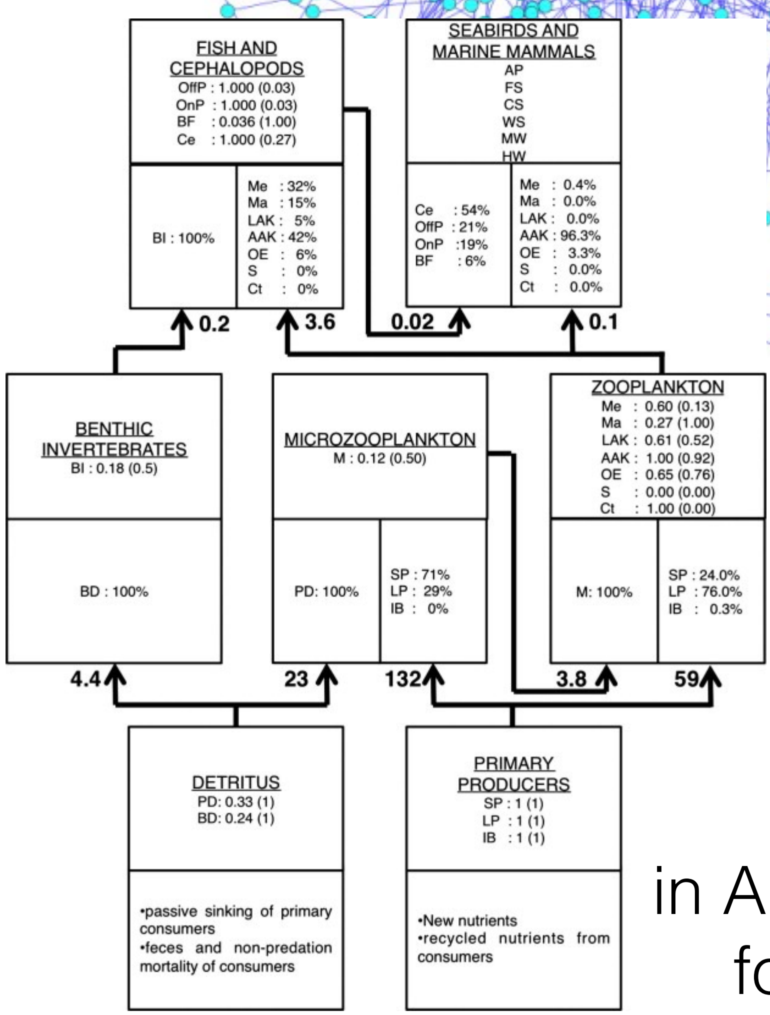
Biological Networks



E. coli metabolic network

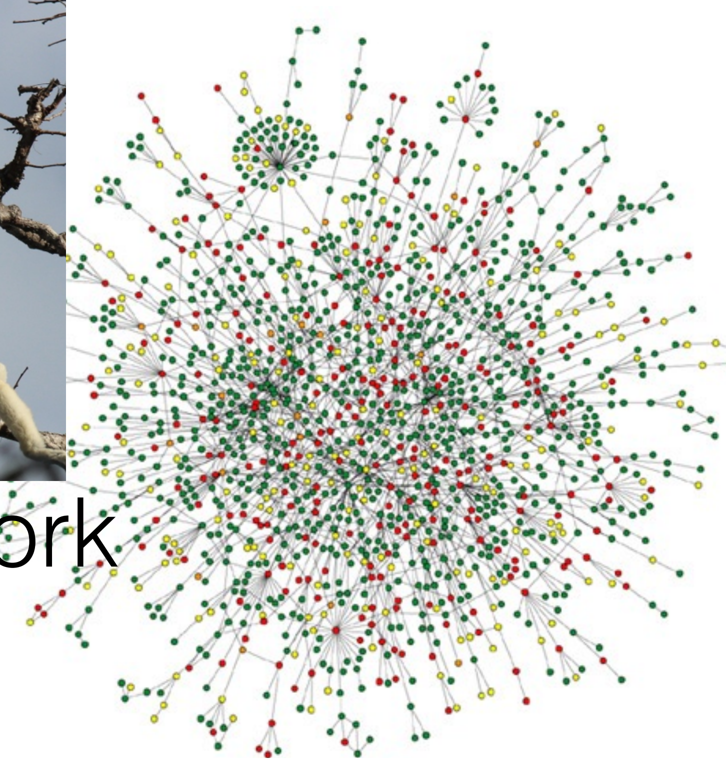


Alberta sexual network



sifaka social network

Energy flows
in Antarctic Peninsula
food web model



Saccharomyces protein-protein interaction network

Mathematics of Networks

Notation and Definitions

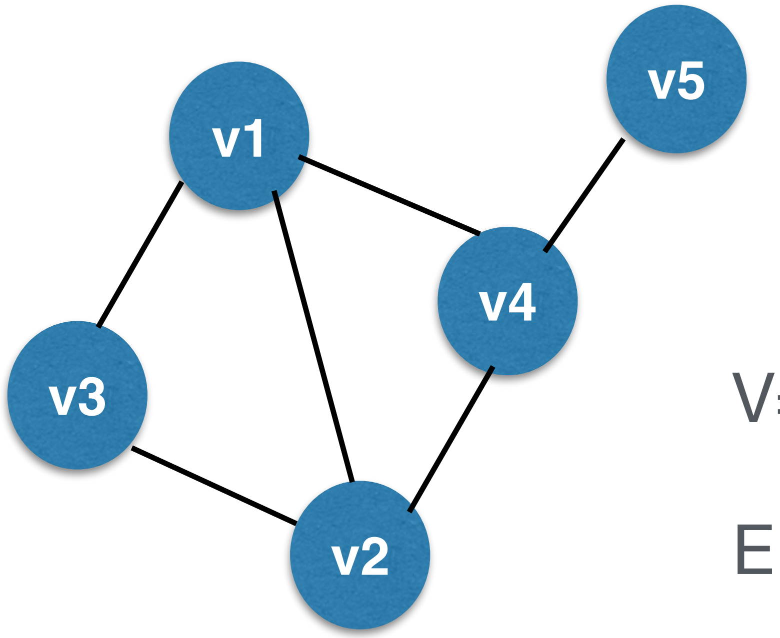
Consider an undirected network (graph) **G** with *n* vertices

G = (V, E)

V is the set of vertices

E is the set of edges

Edge **(u, v)** is the edge from the origin vertex *u* to destination vertex *v*



V= {v1, v2, v3, v4, v5}

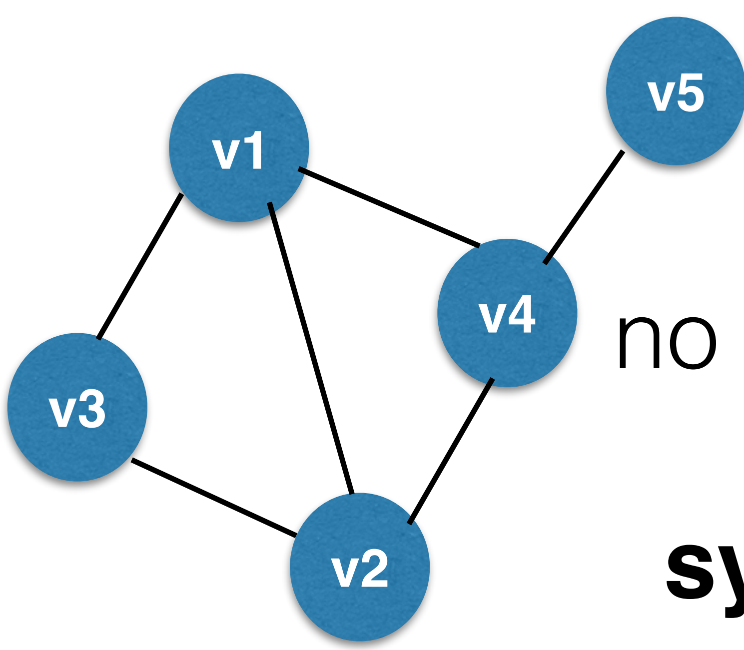
E = {(v1, v3), (v1, v4), (v1, v2), (v2, v3), (v2, v4), (v4, v5)}

The Adjacency Matrix

Edge list: E = {(v1, v3), (v1, v4), (v1, v2), (v2, v3), (v2, v4), (v4, v5)}

The adjacency matrix of a network with *n* vertices is the *n x n* matrix **A** in which:

$$A_{ij} = \begin{cases} 1 \text{ if } (i, j) \in E \\ 0 \text{ otherwise} \end{cases}$$



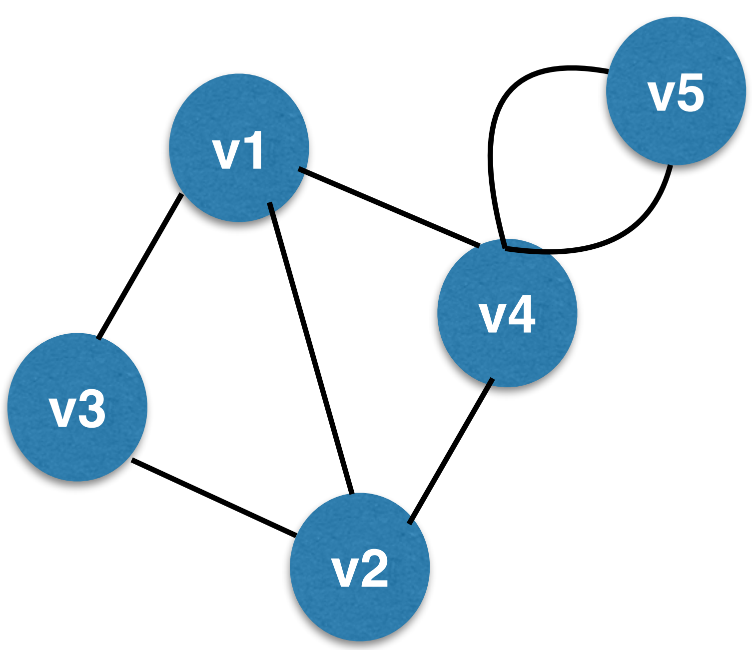
no self edges

symmetric

Undirected network

	v1	v2	v3	v4	v5
v1	0	1	1	1	0
v2	1	0	1	1	0
v3	1	1	0	0	0
v4	1	1	0	0	1
v5	0	0	0	1	0

Multi-edges



	v1	v2	v3	v4	v5
v1	0	1	1	1	0
v2	1	0	1	1	0
v3	1	1	0	0	0
v4	1	1	0	0	**2**
v5	0	0	0	**2**	0

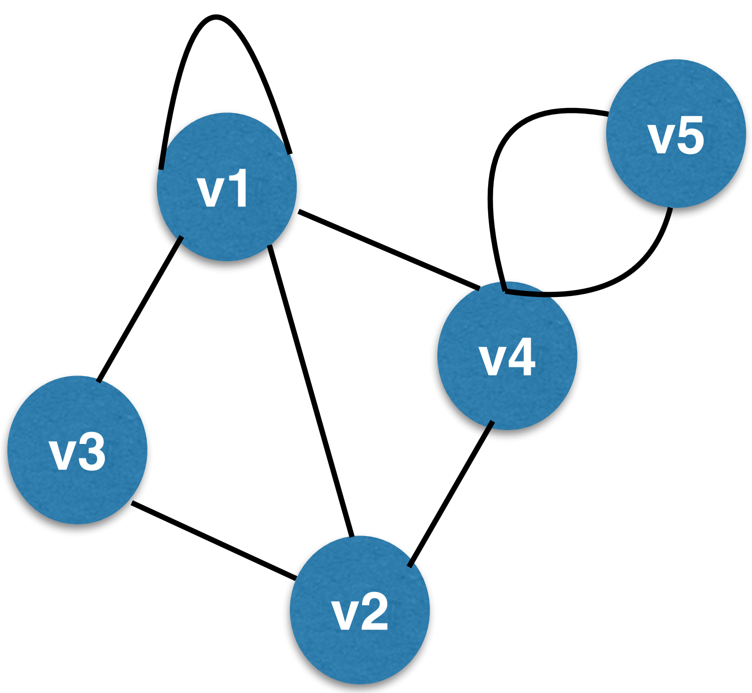
Self-edges

Set corresponding diagonal element A_{ii} to 2

Why 2 and not 1???

Need to count both ends of every edge

Non self-edges appear twice in the adjacency matrix



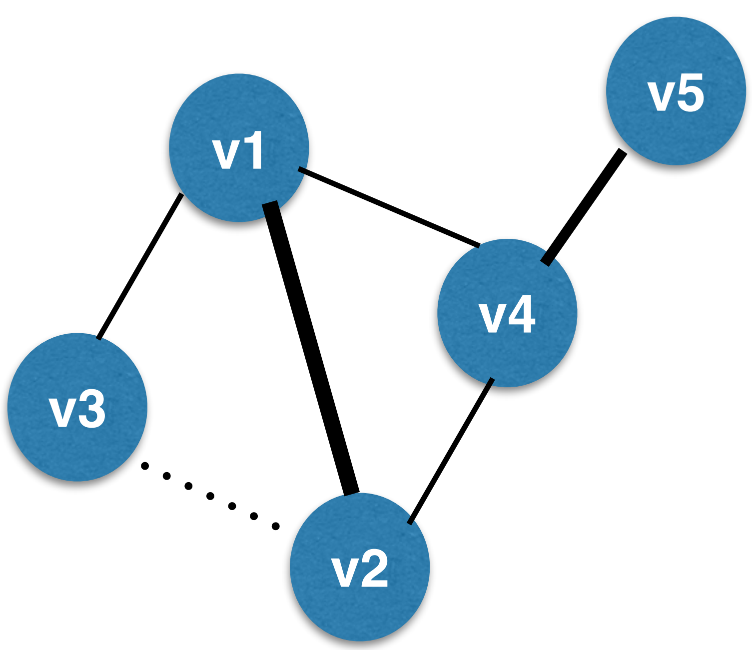
	v1	v2	v3	v4	v5
v1	**2**	1	1	1	0
v2	1	0	1	1	0
v3	1	1	0	0	0
v4	1	1	0	0	**2**
v5	0	0	0	**2**	0

Weighted Networks

Many networks have edges that form simple presence/absence connections between vertices

However, in some situations, it is useful to represent edges as having a strength, weight, or value (e.g., energy flow in predator-prey interactions, frequency of contact between individuals in a social network)

Values can be positive or negative



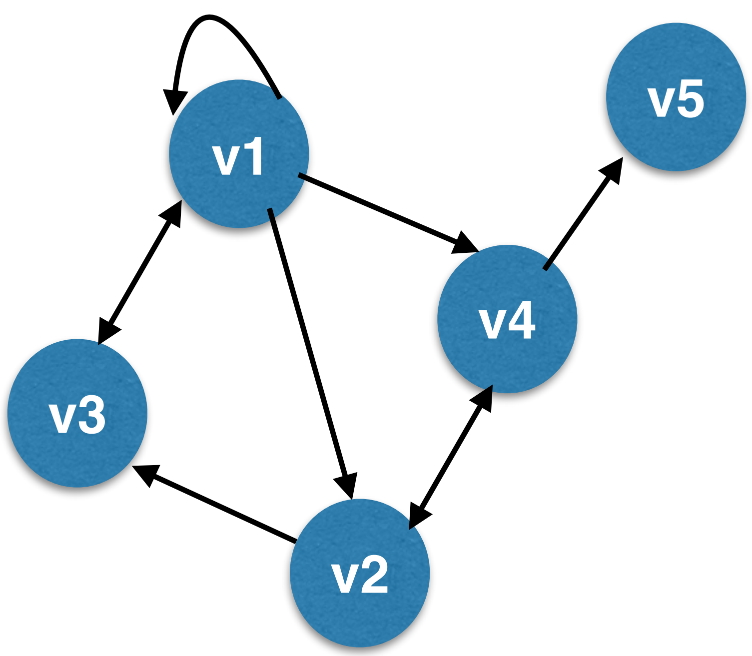
	v1	v2	v3	v4	v5
v1	0	**4**	1	1	0
v2	**4**	0	**-0.5**	1	0
v3	1	**-0.5**	0	0	0
v4	1	1	0	0	**2**
v5	0	0	0	**2**	0

Directed Networks (Digraphs)

Networks in which each edge has a direction, pointing from one vertex to another

Self-edges are given a value of 1

$$A_{ij} = \begin{cases} 1 \text{ if there's an edge from } j \text{ to } i \\ 0 \text{ otherwise} \end{cases}$$



vertex *j*

vertex *i*

	v1	v2	v3	v4	v5
v1	1	0	1	0	0
v2	1	0	0	1	0
v3	1	1	0	0	0
v4	1	1	0	0	0
v5	0	0	0	1	0

asymmetric

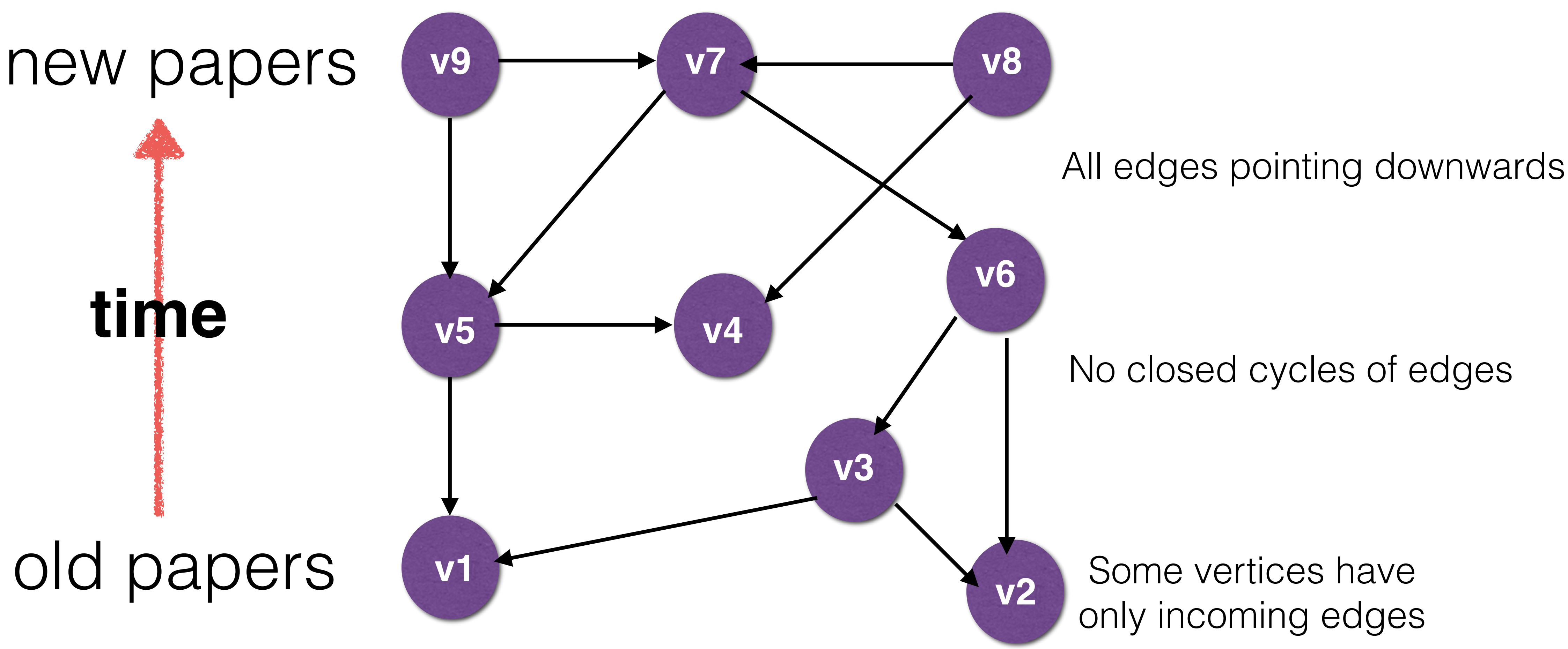
Directed Acyclic Graphs (DAGs)

A **cycle** is a path that starts and ends at the same vertex.

Acyclic directed networks have no cycles (i.e., there is no closed loop of edges with the arrow on each of the edges pointing the same way around the loop).

A self-edge counts as a cycle; therefore, acyclic networks have no self-edges.

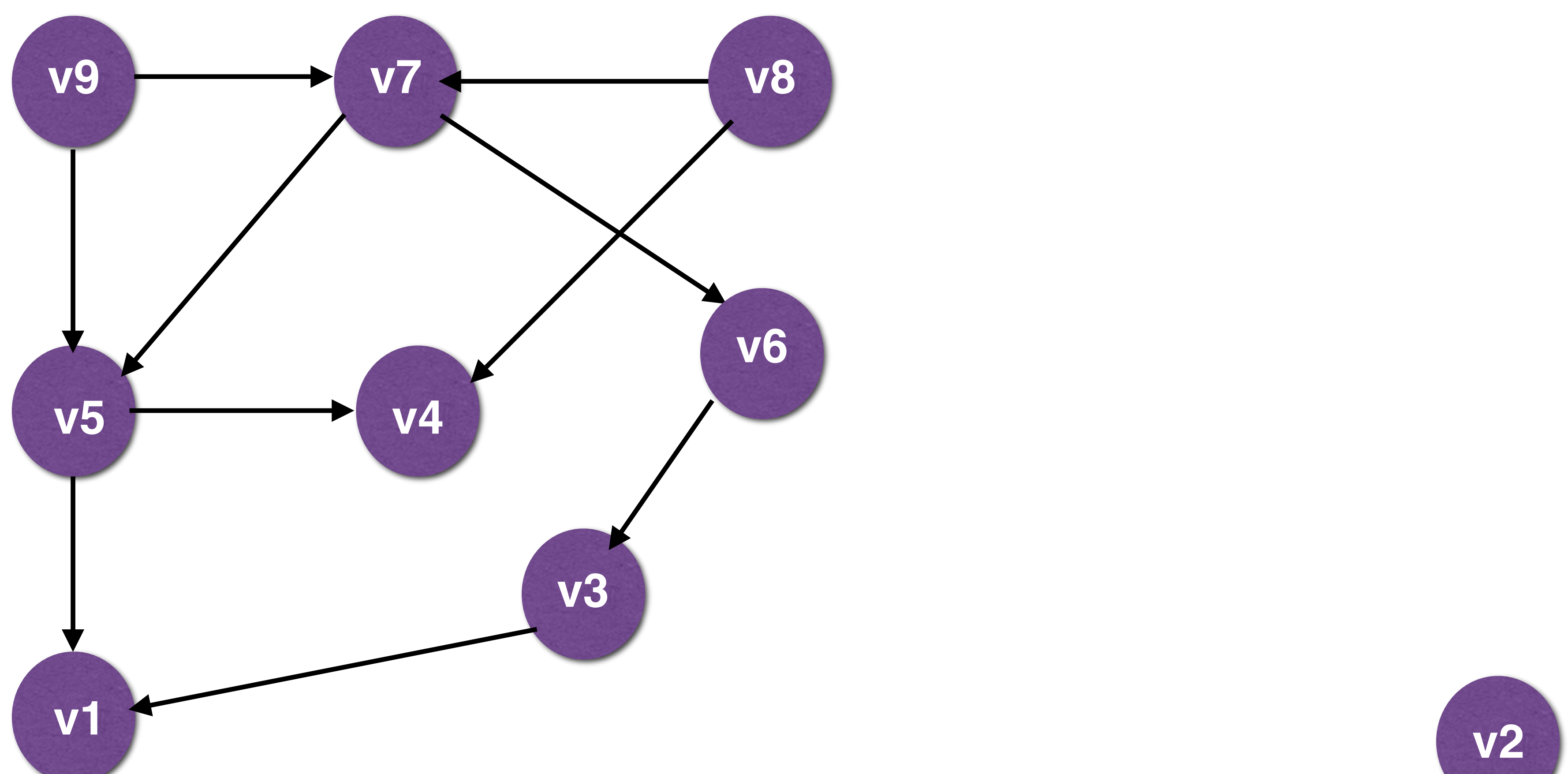
Example: network of citations between papers, gene ontology, epidemiology



Directed Acyclic Graphs

To determine whether a network is cyclic:

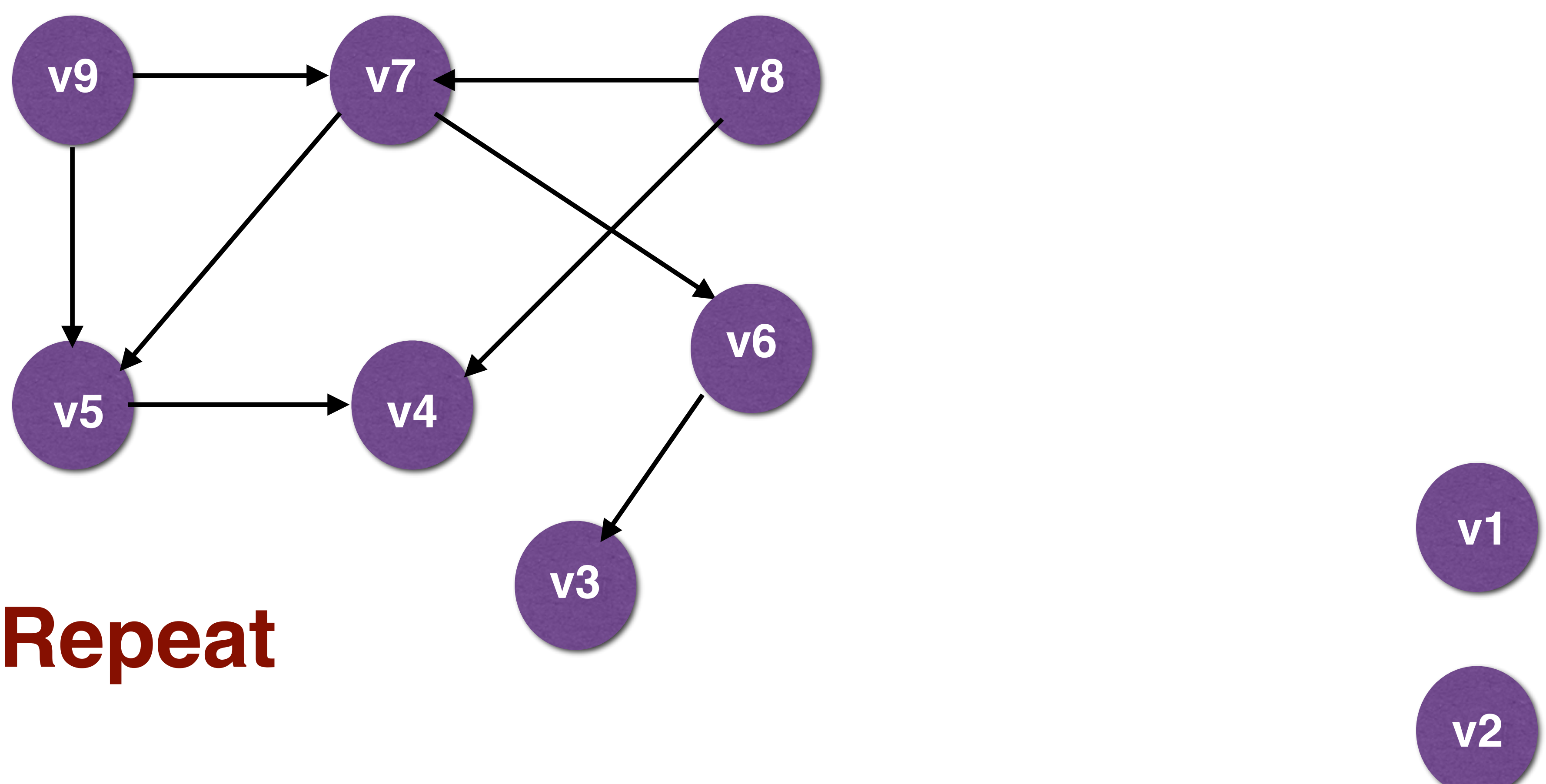
1. Find a vertex with no outgoing edges
2. If no such vertex exists, the network is cyclic. Otherwise, if such a vertex does exist, remove it and all its ingoing edges from the network.
3. If all vertices have been removed, the network is acyclic. Otherwise go back to step 1.



Directed Acyclic Graphs

To determine whether a network is cyclic:

1. Find a vertex with no outgoing edges
2. If no such vertex exists, the network is cyclic. Otherwise, if such a vertex does exist, remove it and all its ingoing edges from the network
3. If all vertices have been removed, the network is acyclic. Otherwise go back to step 1.

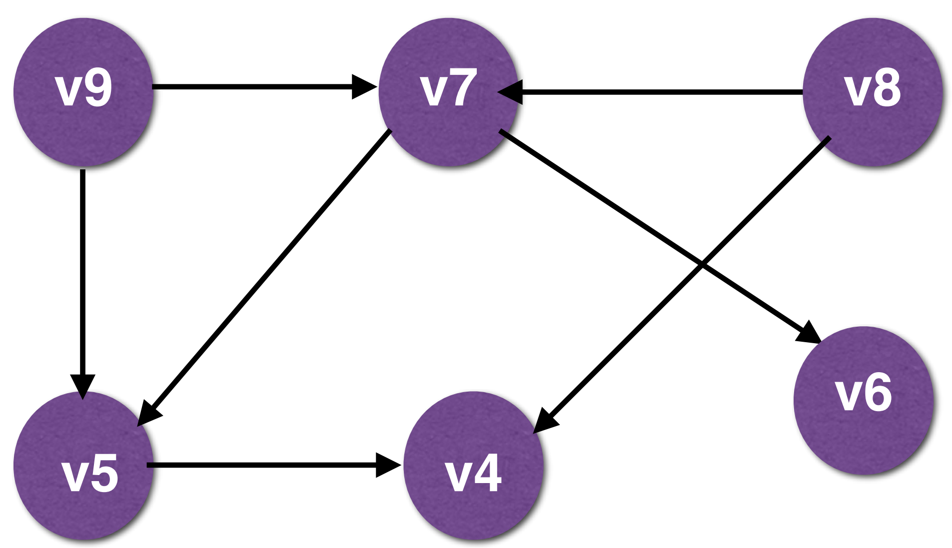


Repeat

Directed Acyclic Graphs

To determine whether a network is cyclic:

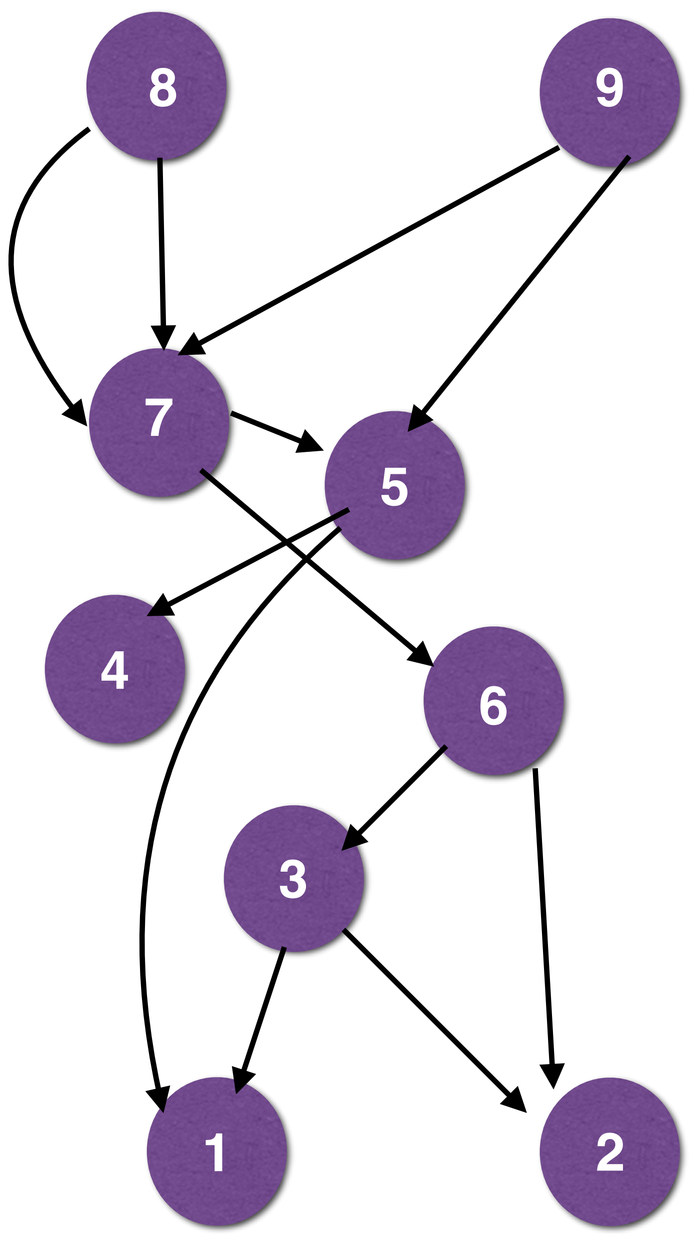
1. Find a vertex with no outgoing edges
2. If no such vertex exists, the network is cyclic. Otherwise, if such a vertex does exist, remove it and all its ingoing edges from the network
3. If all vertices have been removed, the network is acyclic. Otherwise go back to step 1.



Repeat

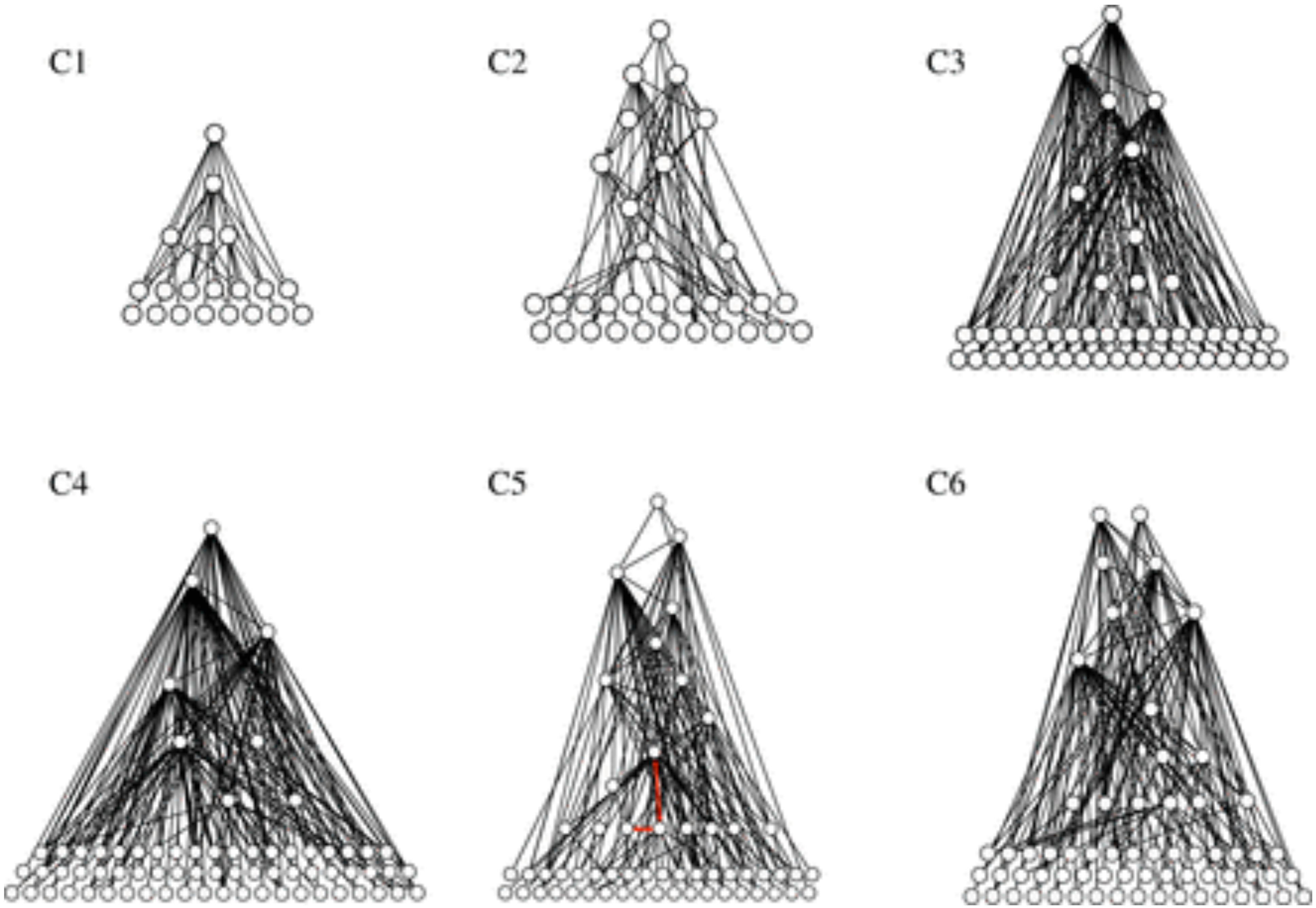
Directed Acyclic Graphs

For every acyclic directed network, there exists at least one labeling of the vertices such that the adjacency matrix **A** will be strictly upper triangular.



All eigenvalues are zero

	v1	v2	v3	v4	v5	v6	v7	v8	v9
v1	0	0	1	0	1	0	0	0	0
v2	0	0	1	0	0	1	0	0	0
v3	0	0	0	0	0	1	0	0	0
v4	0	0	0	0	1	0	0	1	0
v5	0	0	0	0	0	0	1	0	1
v6	0	0	0	0	0	0	1	0	0
v7	0	0	0	0	0	0	0	1	1
v8	0	0	0	0	0	0	0	0	0
v9	0	0	0	0	0	0	0	0	0



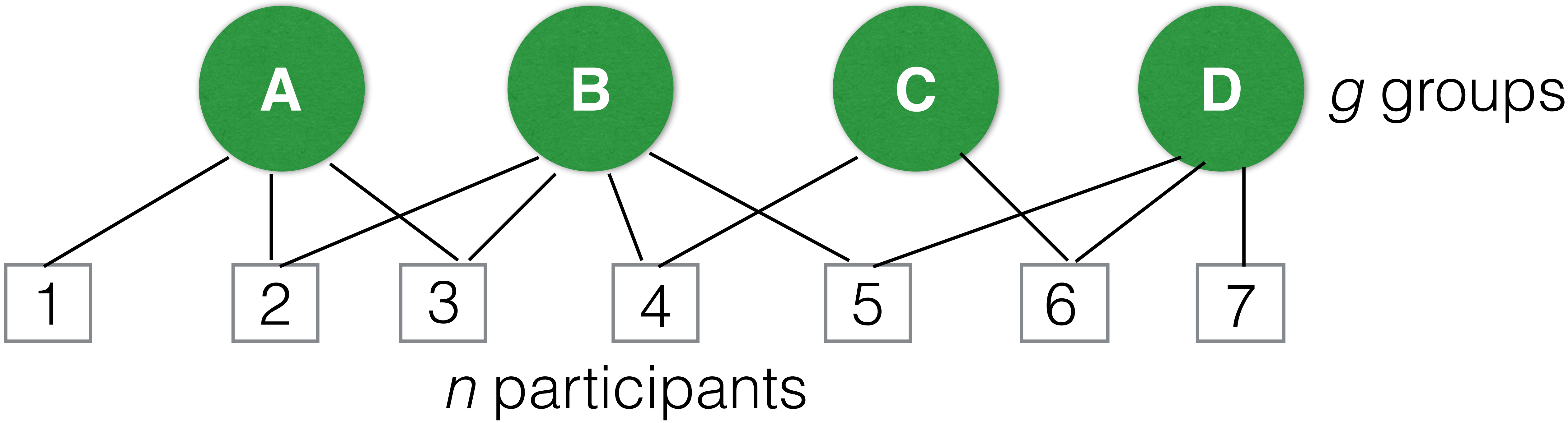
Observed dominance hierarchies in ant networks (approximate DAGs). Workers are aligned by rank.

Shimoji, Hiroyuki, et al. "Global network structure of dominance hierarchy of ant workers." Journal of The Royal Society Interface 11.99 (2014): 20140599.

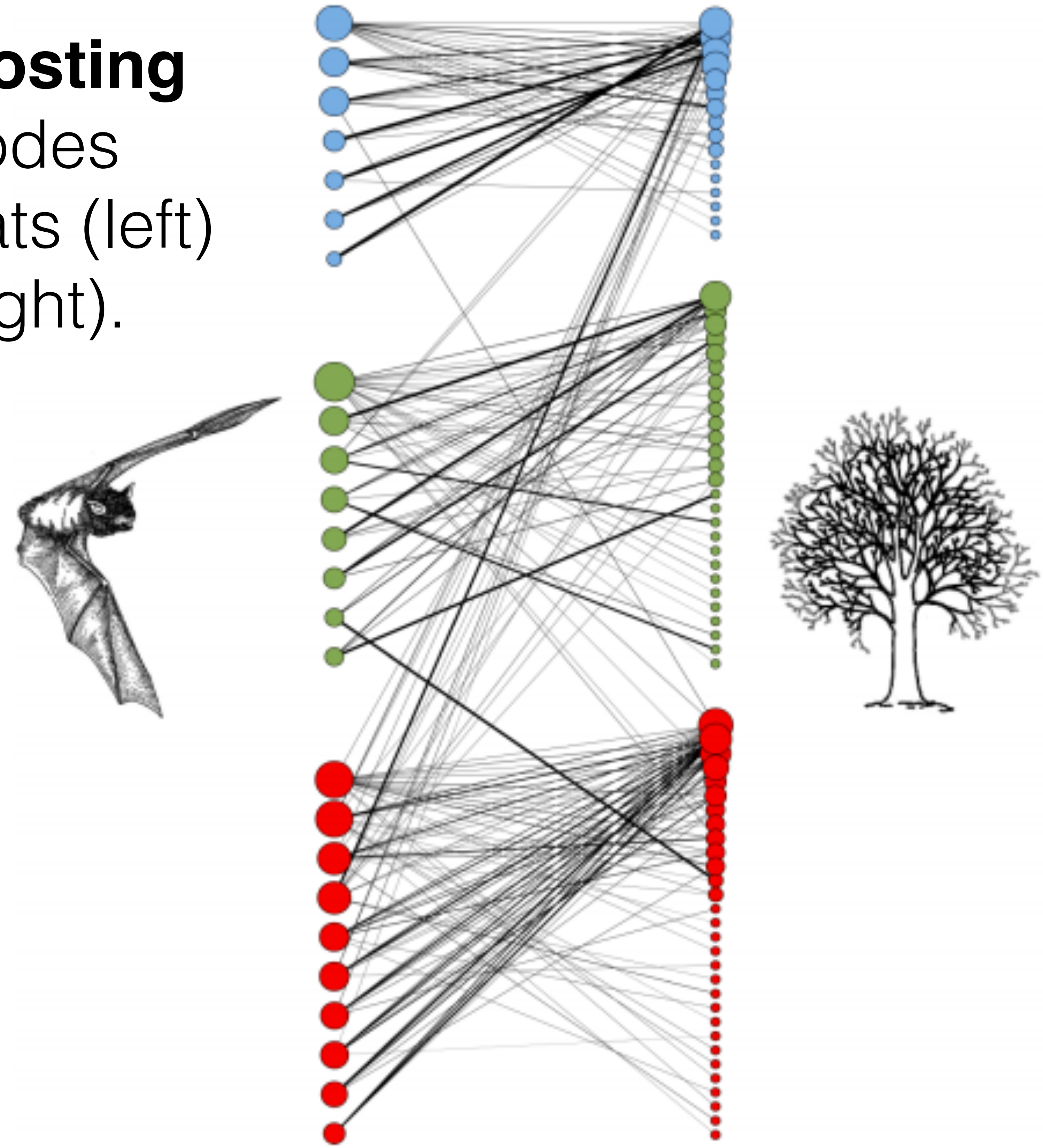
Bipartite networks

Bipartite networks contain two different types of vertices, and the edges run only between vertices of unlike types.

Examples: group membership, actor-film, co-authorship, metabolic reactions



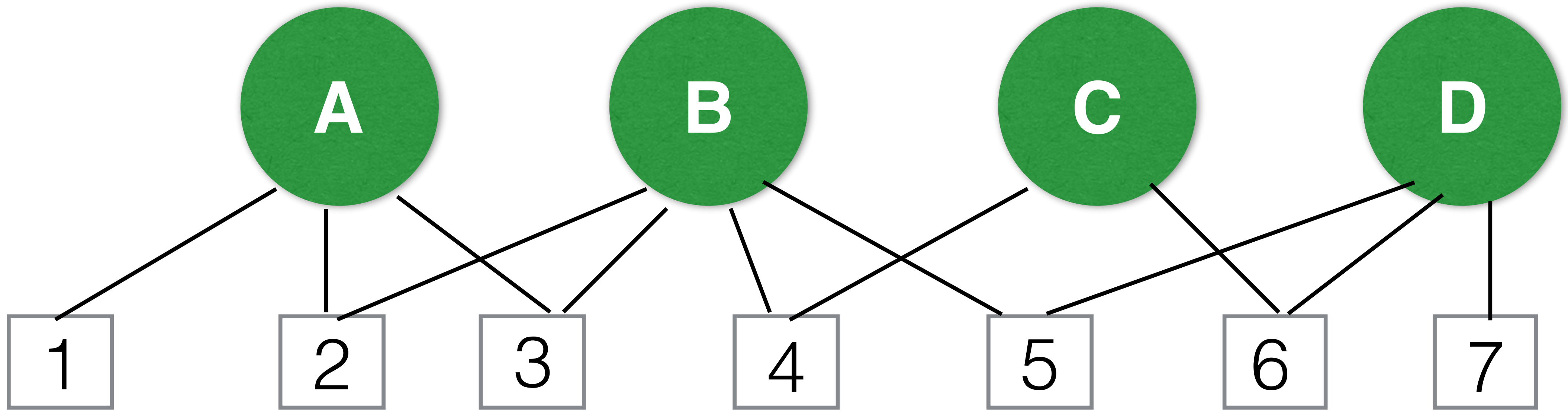
Bipartite roosting network. Nodes represent bats (left) and trees (right).



Miguel A. Fortuna, Ana G. Popa-Lisseanu, Carlos Ibáñez, and Jordi Bascompte 2009. The roosting spatial network of a bird-predator bat. Ecology 90:934–944. http://dx.doi.org/10.1890/08-0174.1

Bipartite networks

The **incidence matrix *B*** for a bipartite network is a g x n matrix with elements Bij:

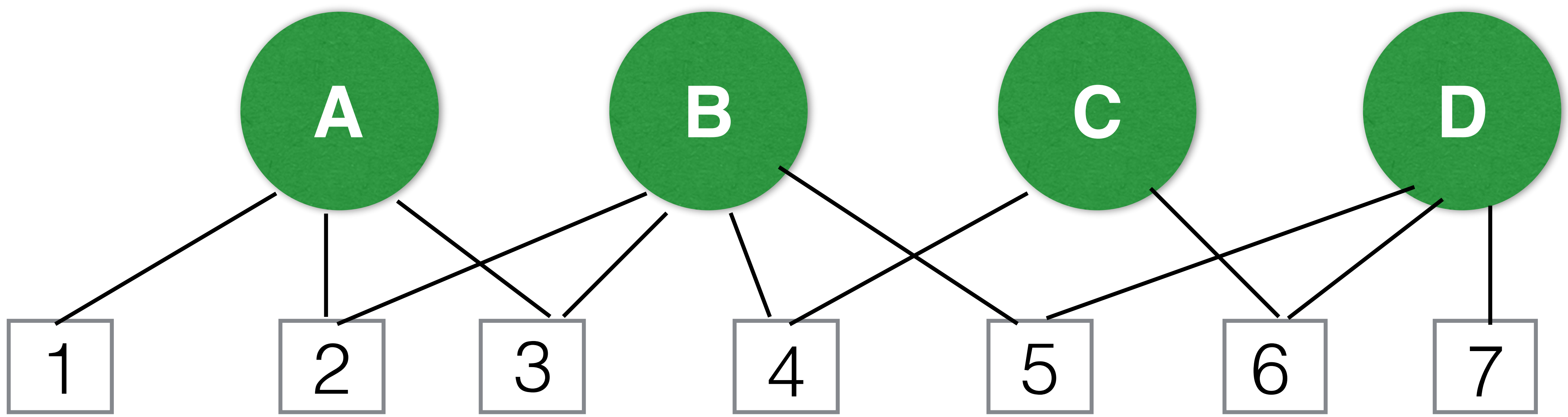


$$B_{ij} = \begin{cases} 1 \text{ if participant } j \text{ belongs to } i \\ 0 \text{ otherwise} \end{cases}$$

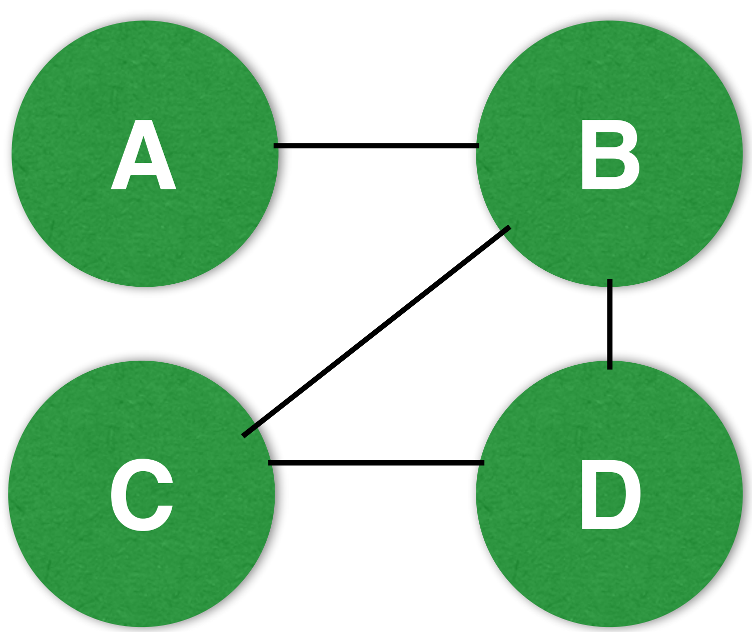
	1	2	3	4	5	6	7
A	1	1	1	0	0	0	0
B	0	1	1	1	1	0	0
C	0	0	0	1	0	1	0
D	0	0	0	0	1	1	1

One-mode projections

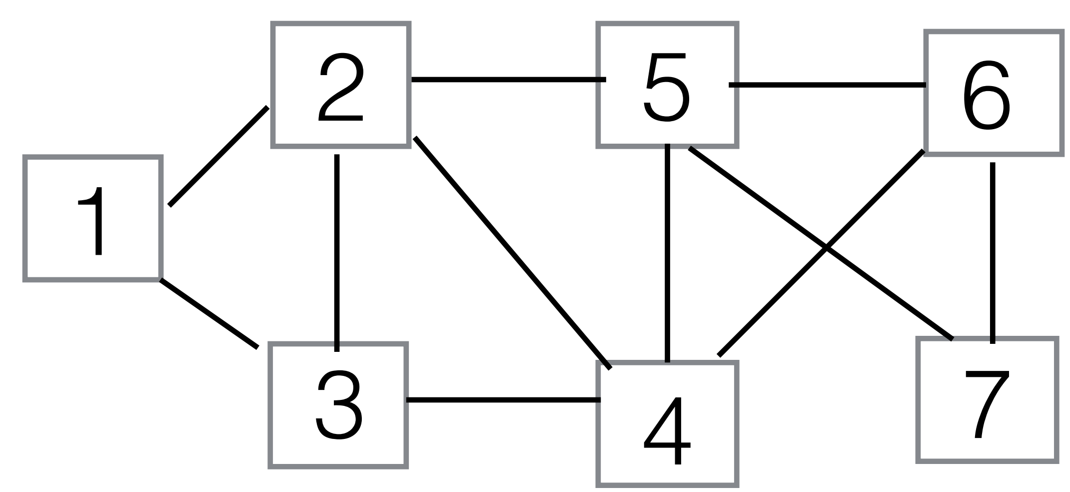
If we want to work with direct connections of vertices of just one type:



Shared participants



Common membership



Adjacency matrix from the incidence matrix

B=

	1	2	3	4	5	6	7
A	1	1	1	0	0	0	0
B	0	1	1	1	1	0	0
C	0	0	0	1	0	1	0
D	0	0	0	0	1	1	1

B^T=

	A	B	C	D
1	1	0	0	0
2	1	1	0	0
3	1	1	0	0
4	0	1	1	0
5	0	1	0	1
6	0	0	1	1
7	0	0	0	1

$$P = B^T B = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Adjacency matrix from the incidence matrix

$$P = B^T B = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

set diagonal to zero

set non-zero items to 1

$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

unweighted

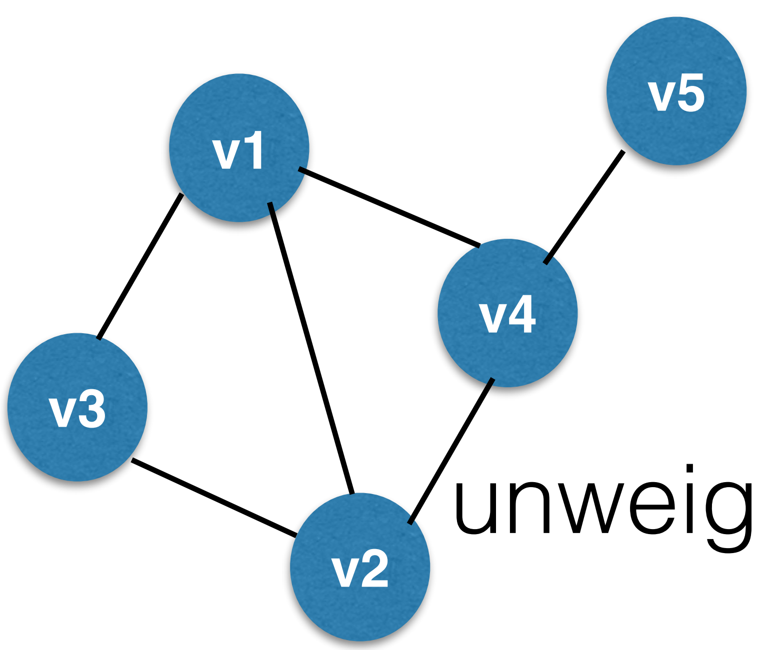
Degree

The **degree** of a vertex is the number of edges connected to it.

How can we compute the degree of a node from the adjacency matrix?

$$k_i = \sum_{j=1}^{n} A_{ij}$$

$$k_3 = \sum_{j=1}^{n} A_{3j} = A_{31} + A_{32} + A_{33} + A_{34} + A_{35} = 2$$



unweighted

	v1	v2	v3	v4	v5
v1	0	1	1	1	0
v2	1	0	1	1	0
v3	1	1	0	0	0
v4	1	1	0	0	1
v5	0	0	0	1	0

symmetric

Degrees and edges

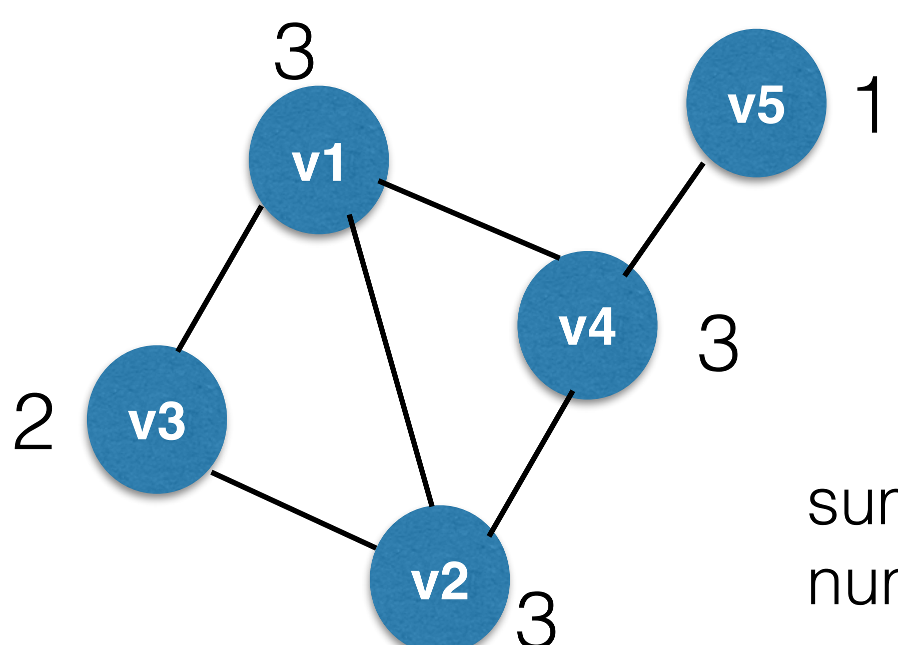
m: number of edges

n: number of vertices

k_i: degree of vertex *i*

c: average degree

What is the relationship between the sum of degrees and the number of edges in the graph?



$$2m = \sum_{i=1}^{n} k_i$$

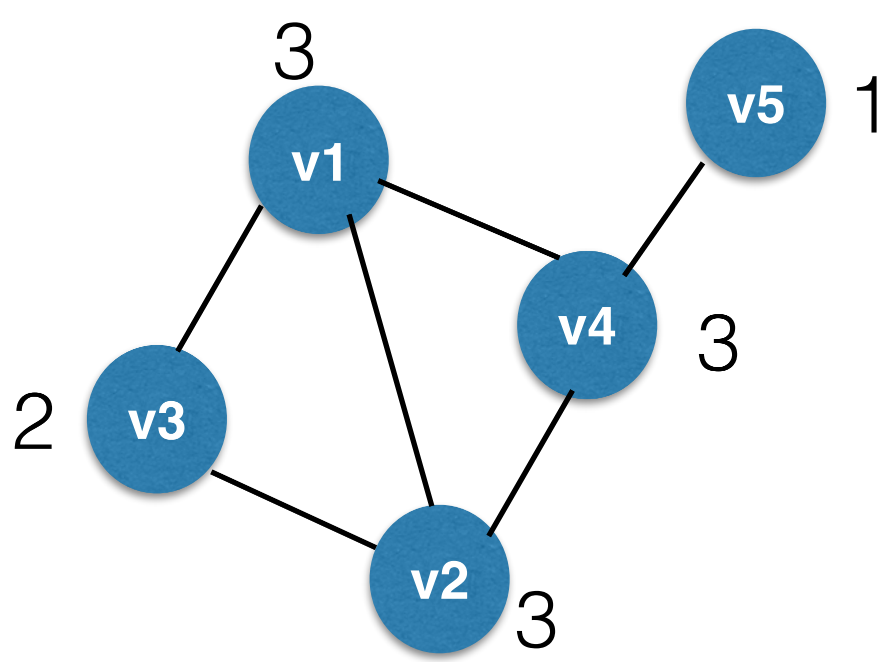
sum of degrees: 3+3+2+3+1=12
number of edges = 6
sum of degrees= 2 x # edges

Degrees and edges

What is the relationship between the sum of degrees and the number of edges in the graph?

$$m = \left(\frac{1}{2}\right)\sum_{i=1}^{n} k_i = \left(\frac{1}{2}\right)\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}$$

What is the average degree of the network?



$$c = \frac{1}{n}\sum_{i=1}^{n} k_i$$

$$= \left(\frac{1}{5}\right)(3+3+2+3+1) = 2.4$$

$$c = \frac{2m}{n}$$

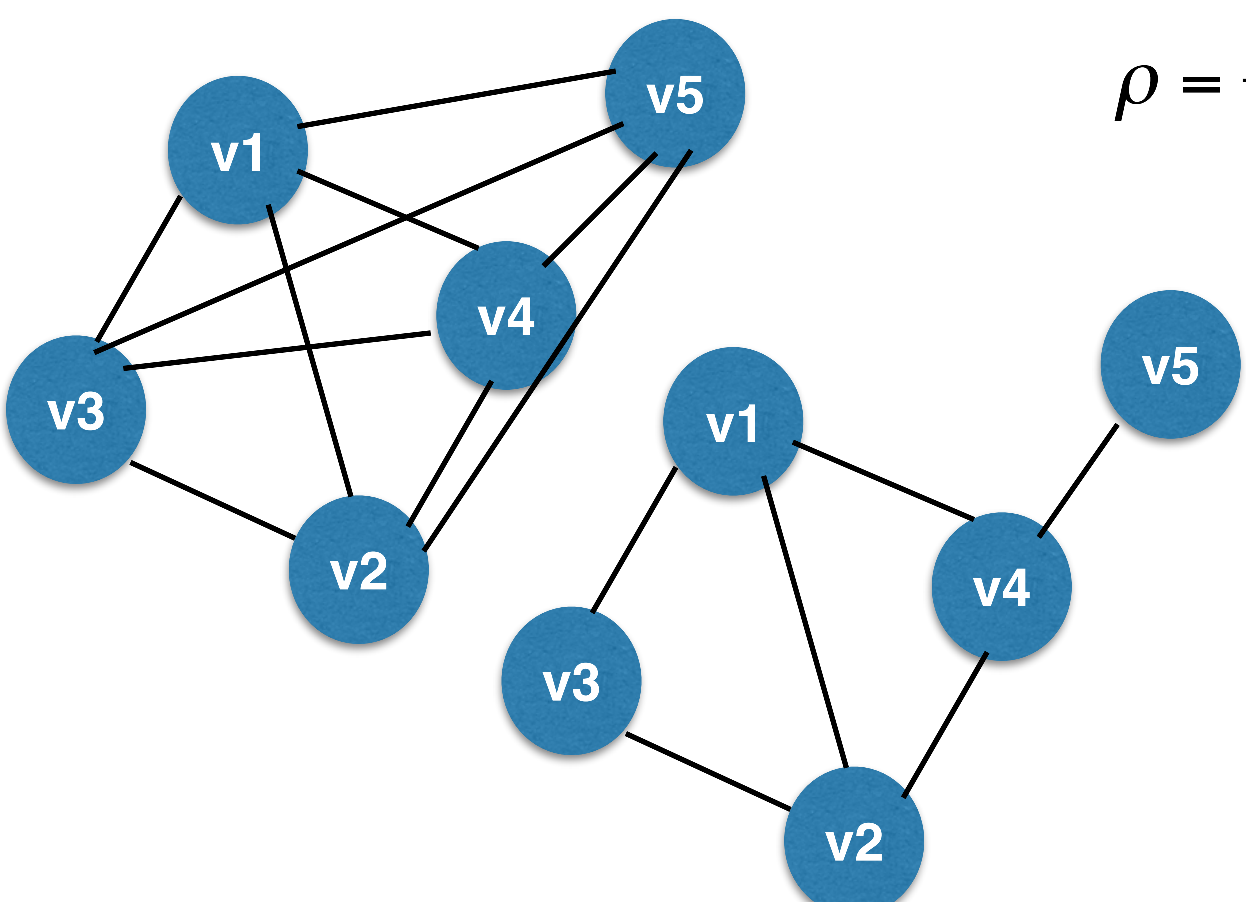
Degrees and edges

What is the **maximum number of edges** in a graph (no multi-edges or self edges)?

$$\binom{n}{2} = \frac{1}{2}n(n-1)$$

$$\binom{5}{2} = \frac{1}{2}5(5-1) = 10$$

The **density** of a graph is the fraction of all possible edges actually present.



$$\rho = \frac{m}{\binom{n}{2}} = \frac{m}{\frac{1}{2}n(n-1)} = \frac{2m}{n(n-1)} = \frac{12}{20} = 0.6$$

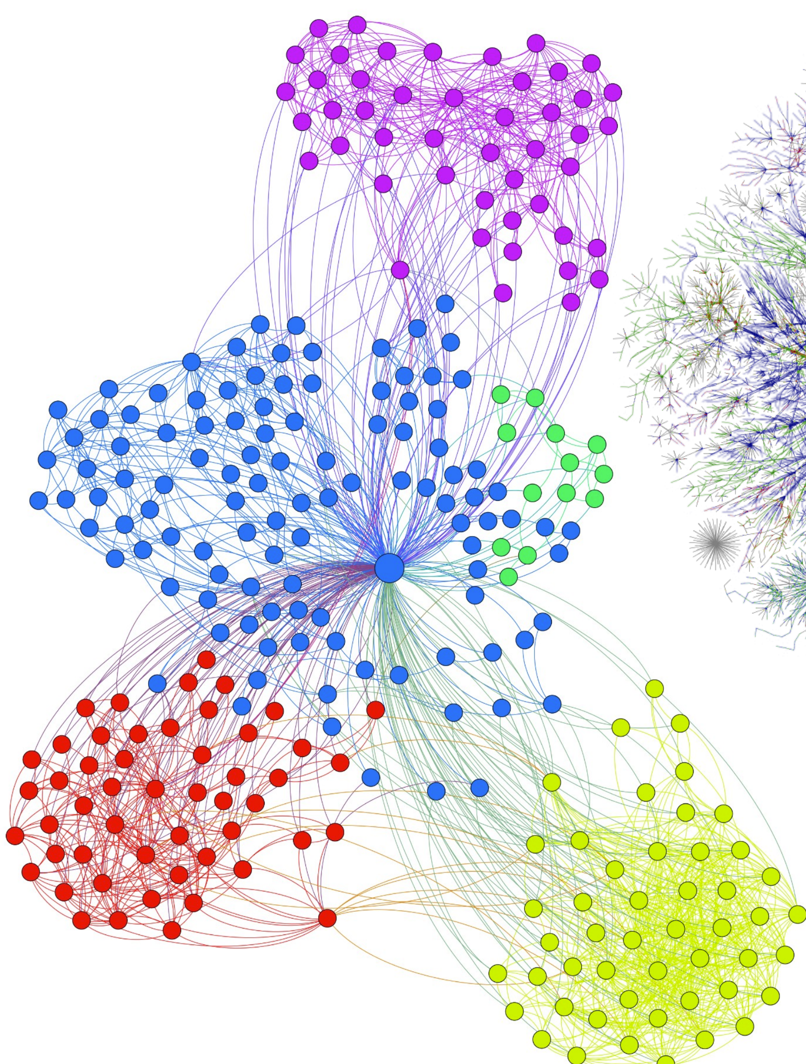
mean degree

$$\rho = \frac{2m}{n(n-1)} = \frac{c}{n-1} = 0.6$$

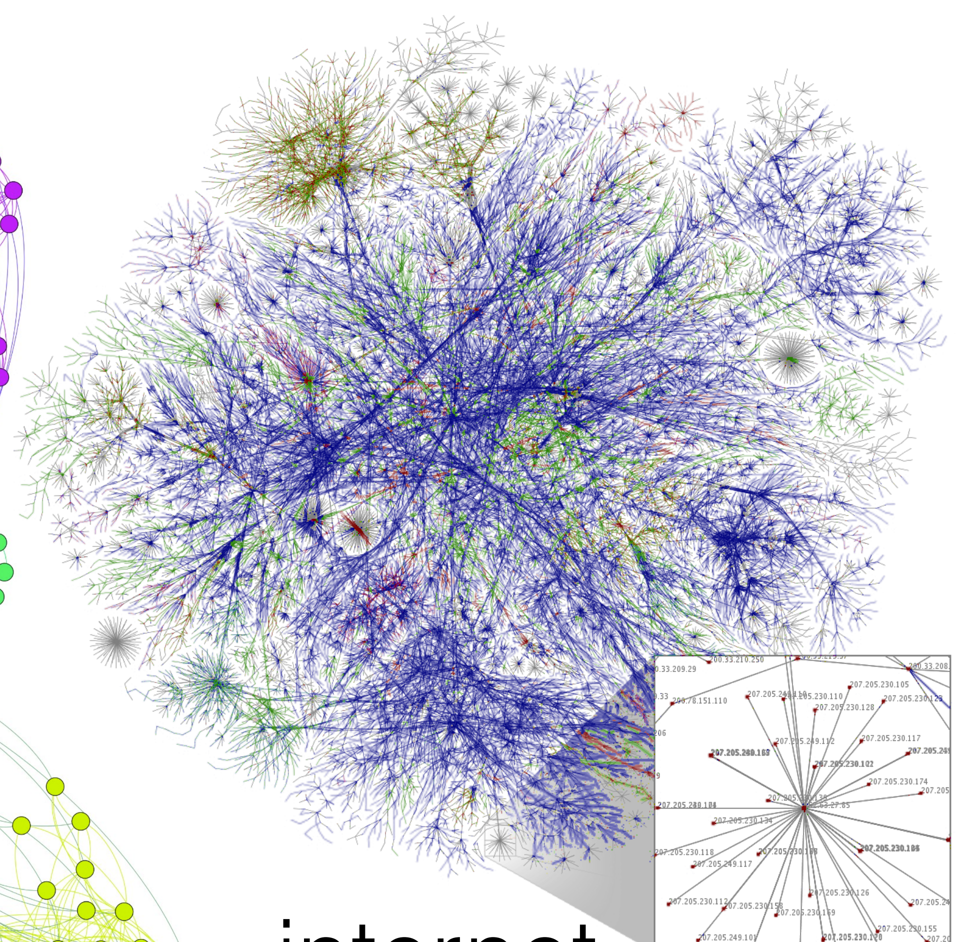
Dense vs. Sparse Networks

A network for which the density ρ tends to a constant as n —> ∞ is **dense.**

A network in which ρ —> 0 as n —> ∞ is **sparse** (the case for most networks).

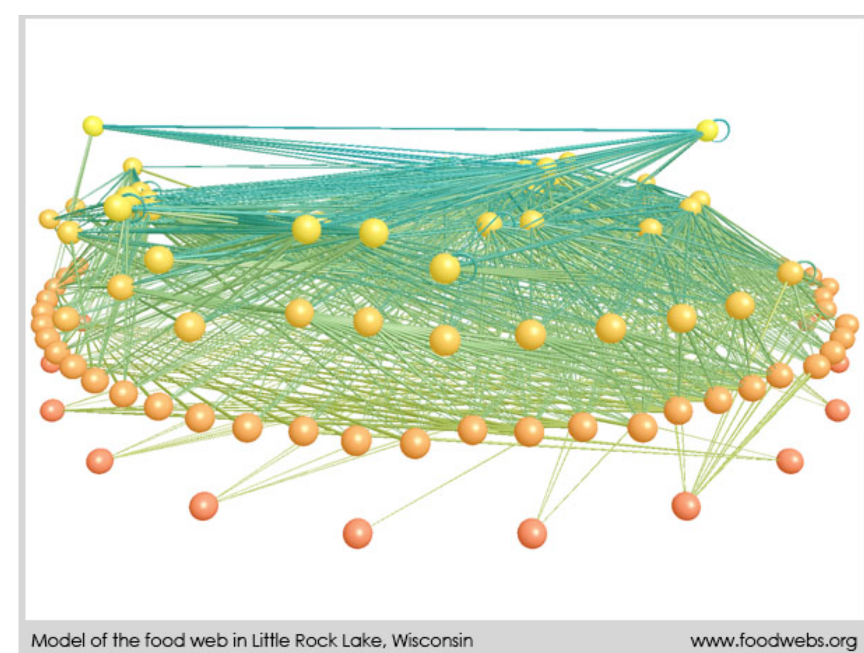


friendship



internet

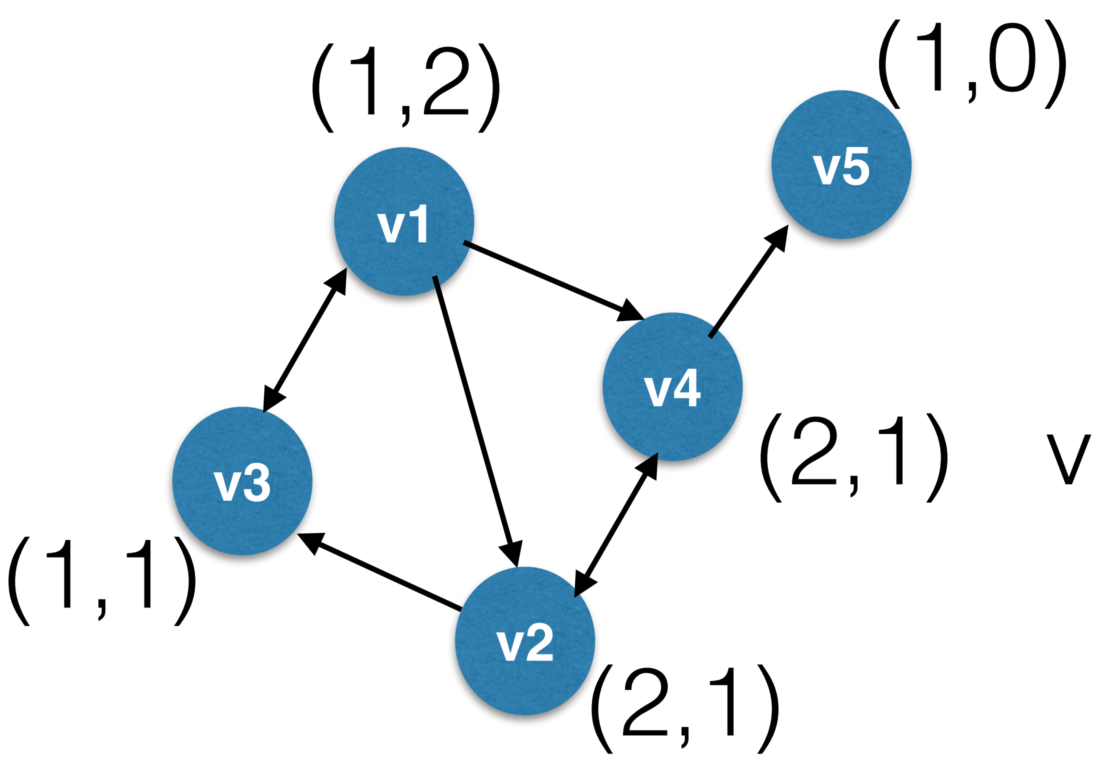
Food webs: density tends to be constant regardless of size



Degree in directed networks

Vertices in directed graphs have an **in-degree** and **out-degree.**

What is the relationship between the sum of the in-degrees and the sum of the out-degrees?



vertex *j*

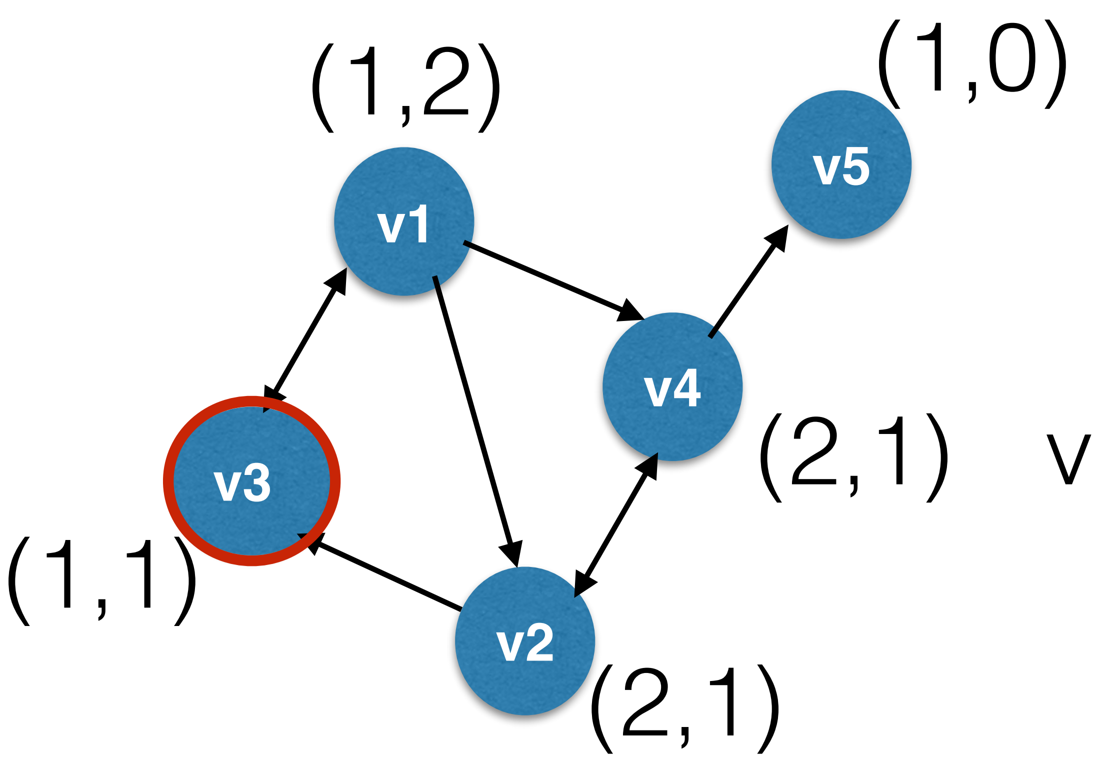
vertex *i*

	v1	v2	v3	v4	v5
v1	0	0	1	0	0
v2	1	0	0	1	0
v3	1	1	0	0	0
v4	1	1	0	0	0
v5	0	0	0	1	0

Degree in directed networks

$$k_i^{\text{in}} = \sum_{j=1}^{n} A_{ij}$$ sum the corresponding row

$$k_j^{\text{out}} = \sum_{i=1}^{n} A_{ij}$$ sum the corresponding column



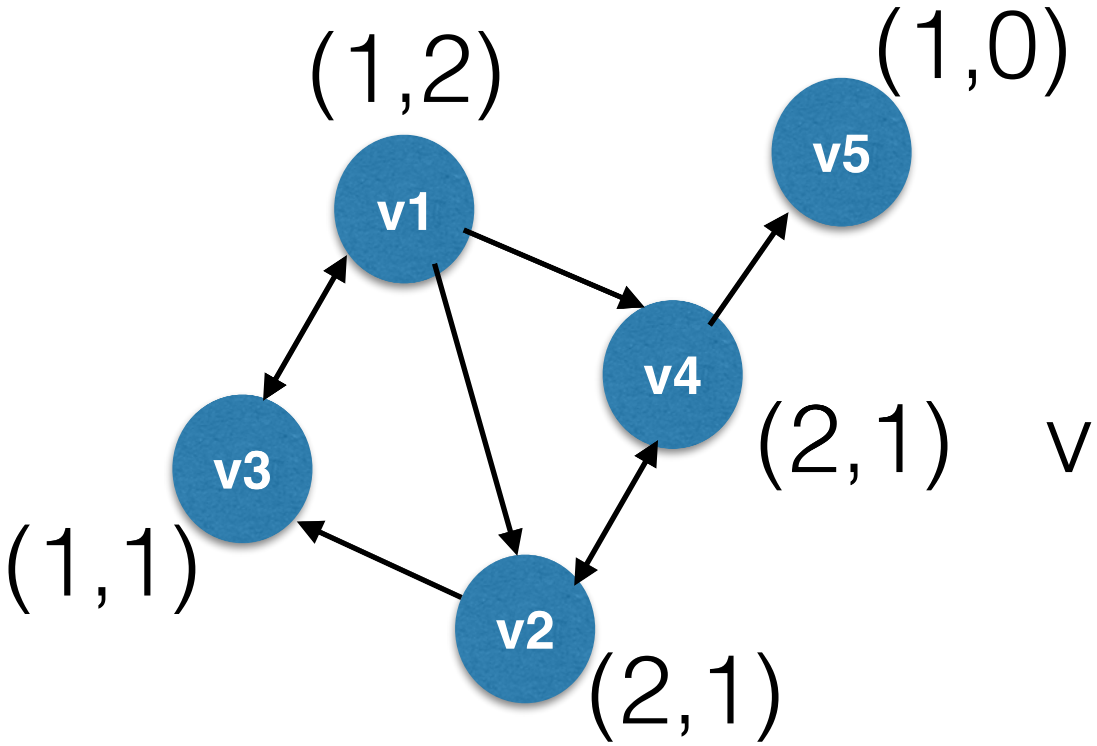
vertex *j*

vertex *i*

	v1	v2	v3	v4	v5
v1	0	0	1	0	0
v2	1	0	0	1	0
v3	1	1	0	0	0
v4	1	1	0	0	0
v5	0	0	0	1	0

Mean degree in directed networks

$$c_{\text{in}} = c_{\text{out}} = \frac{2m}{n}$$



vertex *j*

vertex *i*

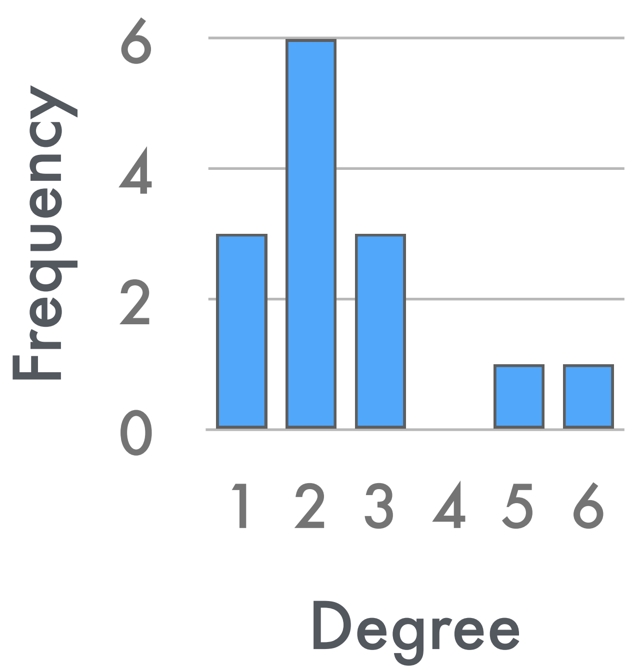
	v1	v2	v3	v4	v5
v1	0	0	1	0	0
v2	1	0	0	1	0
v3	1	1	0	0	0
v4	1	1	0	0	0
v5	0	0	0	1	0

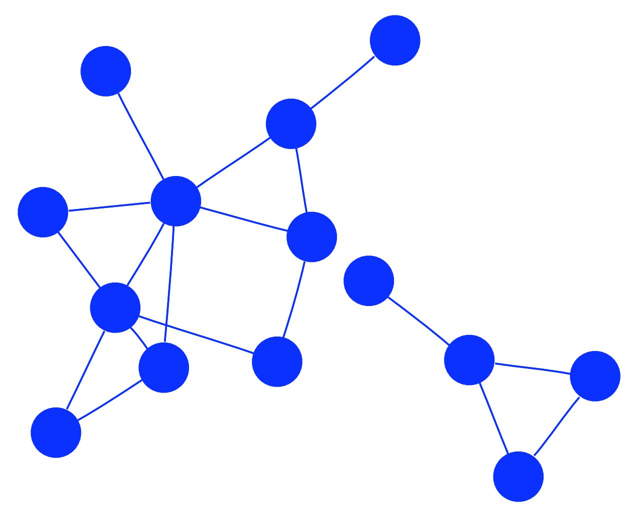
Degree Distribution

The **degree distribution** of a network is the number or fraction of vertices with each possible degree.

p_k = fraction of nodes in the network with degree k

p_k is also the probability that a randomly chosen node has degree k





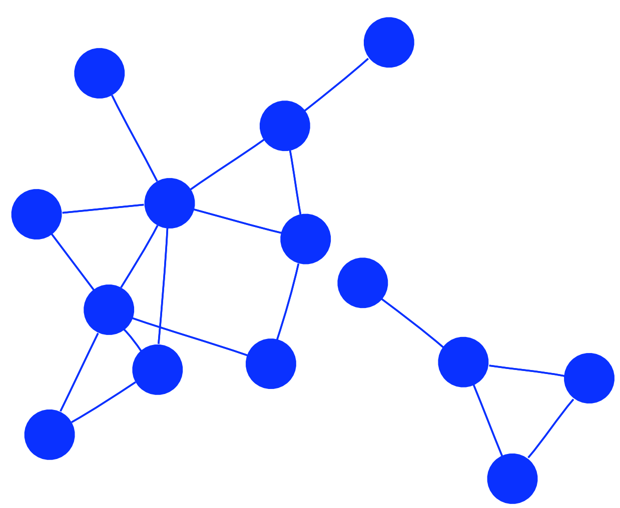
Degree sequence:
(1,1,1,2,2,2,2,2,2,3,3,3,5,6)

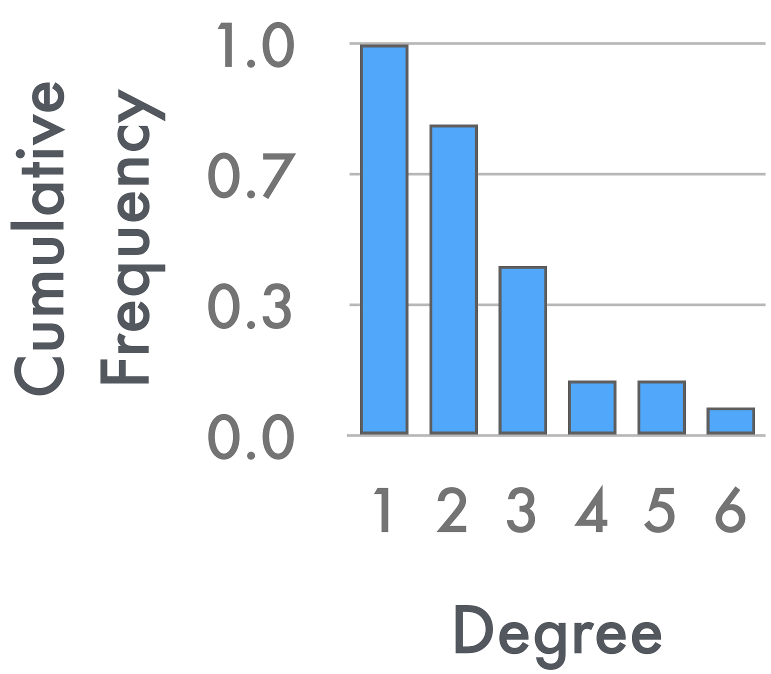
Degree	Number of nodes	Fraction of nodes
1	3	0.21
2	6	0.43
3	3	0.21
4	0	0.00
5	1	0.07
6	1	0.07
Total	14	1

Cumulative Degree Distribution

The **cumulative degree distribution** P_k gives the fraction of vertices with degree greater than or equal to k.

P_k is also the probability that a randomly chosen vertex has degree <u>at least</u> k





Degree	Number of nodes	Fraction of nodes	Cumulative Frequency
1	3	0.21	1.00
2	6	0.43	0.79
3	3	0.21	0.43
4	0	0.00	0.14
5	1	0.07	0.14
6	1	0.07	0.07
Total	14	1	

Large Scale Structure of Networks

Connectivity

Connectivity is the number of independent paths between a pair of vertices.

Edge-independent paths are paths between the same pair of vertices that share no edges.

Vertex-independent paths are paths between the same pair of vertices that share no other vertices (other than the starting and ending vertices).

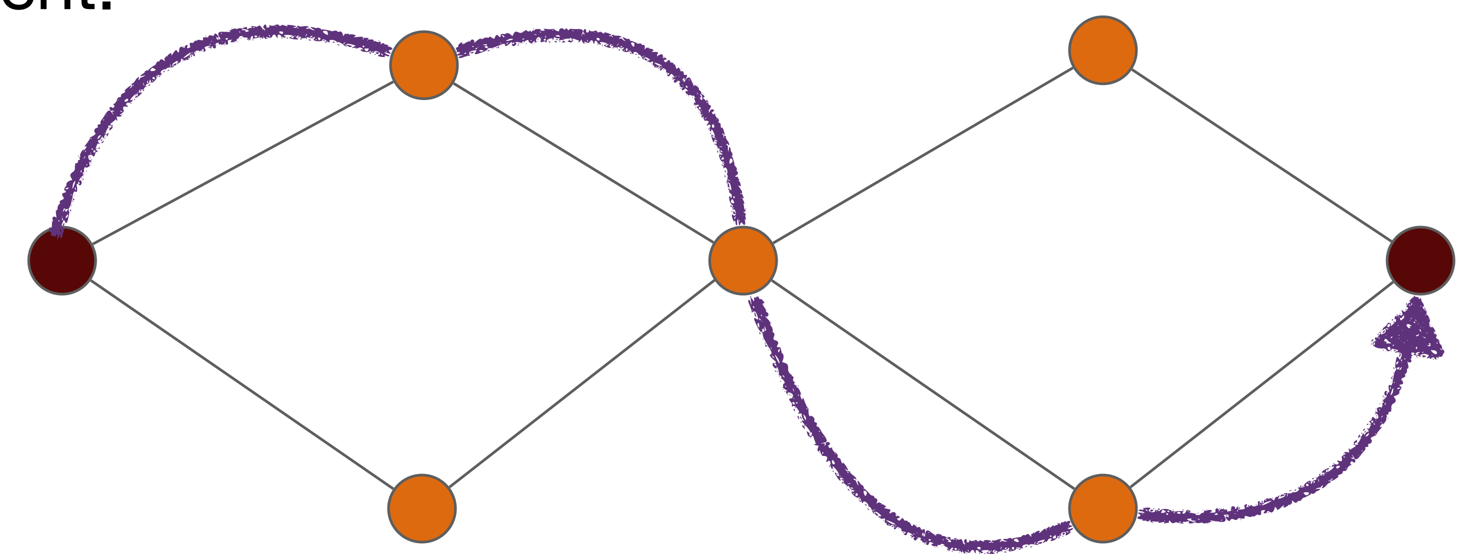
A **cut set** is the set of vertices (or edges) whose removal will disconnect a specified pair of vertices.

Menger's theorem: If there is no cut set of size less than n between a given pair of vertices, then there are at least n independent paths between the same vertices. Applies to both edges and vertices.

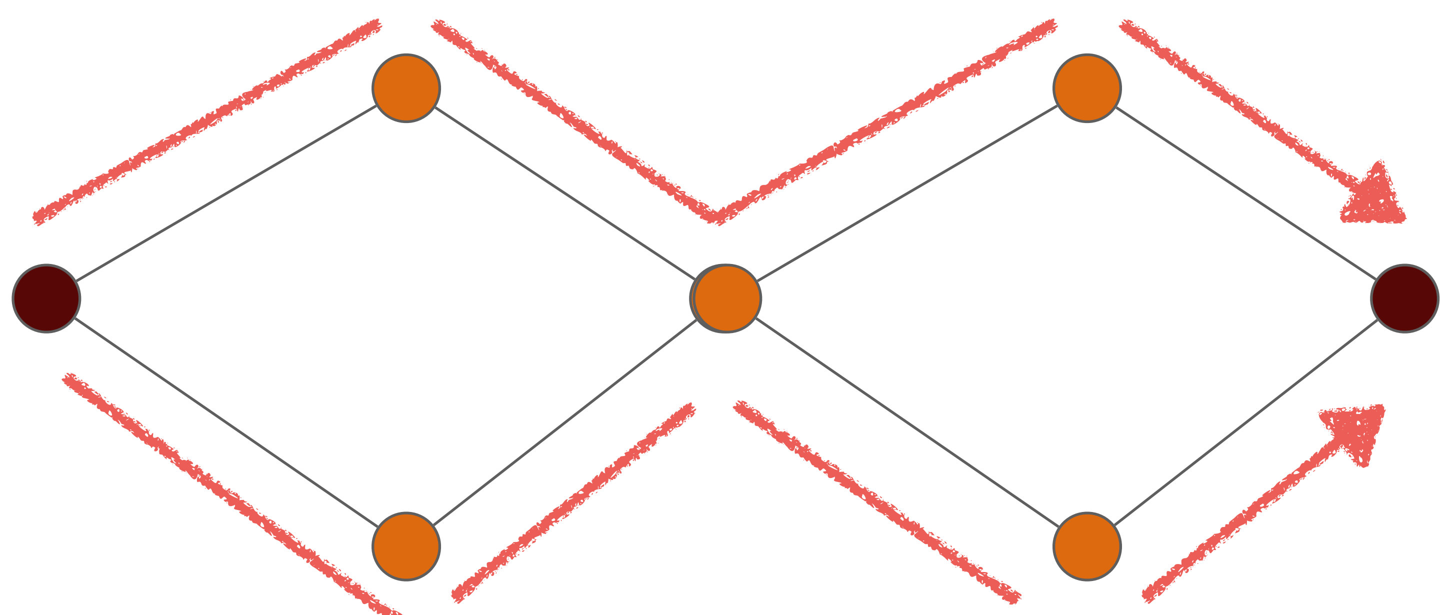
Connectivity

If two paths are vertex-independent, then they are also edge-independent.

If two paths are edge-independent, they <u>are not</u> necessarily vertex-independent.



One vertex-independent path



Two edge-independent paths

Paths

A **path** between two vertices is a continuous sequence of edges, where each successive edge begins where the previous one left off.

A **shortest path** (or **geodesic**) between two vertices is the minimum number of edges you have to travel across to move from one vertex to another. There may be multiple different geodesics, all of the same length.

The length of a shortest path between (*u, v*) is called the **geodesic distance** or **graph distance**.

The **diameter** of a network is the length of the longest shortest path between two vertices in the network.

The **average path length** is the average shortest path between all pairs of vertices.

Average Path Length

d_{ij} denotes the **geodesic distance** from vertex i to vertex j.

The **mean geodesic** is:

$$L = \frac{1}{\frac{n(n+1)}{2}} \sum_{i \geq j} d_{ij}$$

When analyzing disconnected networks, the **harmonic mean** of the geodesics (**global efficiency**) is:

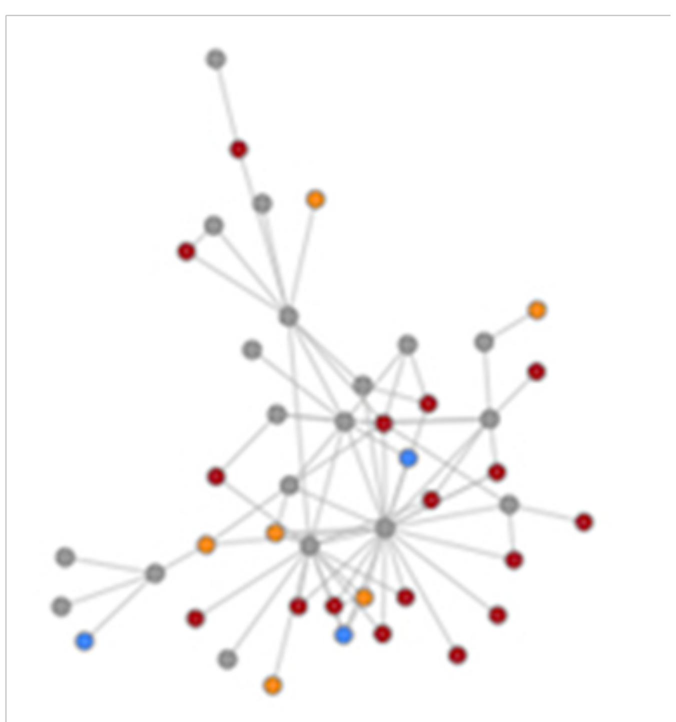
$$L^{-1} = \frac{1}{\frac{n(n+1)}{2}} \sum_{i \geq j} \frac{1}{d_{ij}}$$

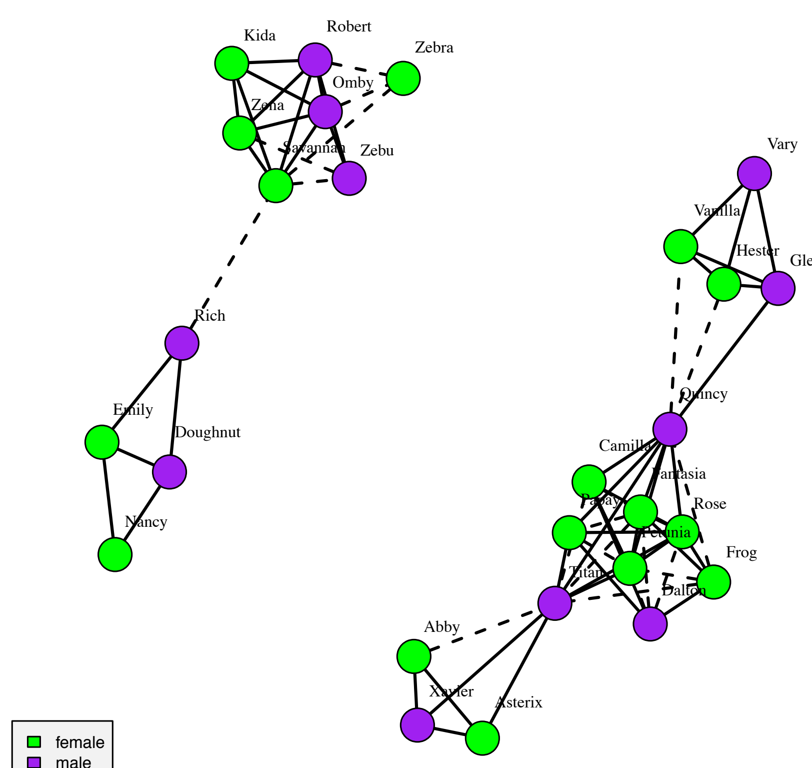
Components

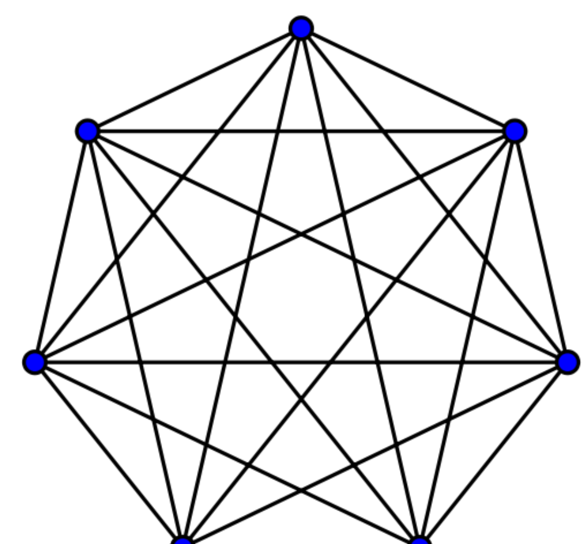
A **connected** network is one in which all pairs of vertices can be connected by a path.

It is possible for there to be no path at all between a given pair of vertices. A **disconnected network** consists of disjoint **connected components** (subgroups).

A **complete** network is one in which there are edges connecting every pair of vertices.





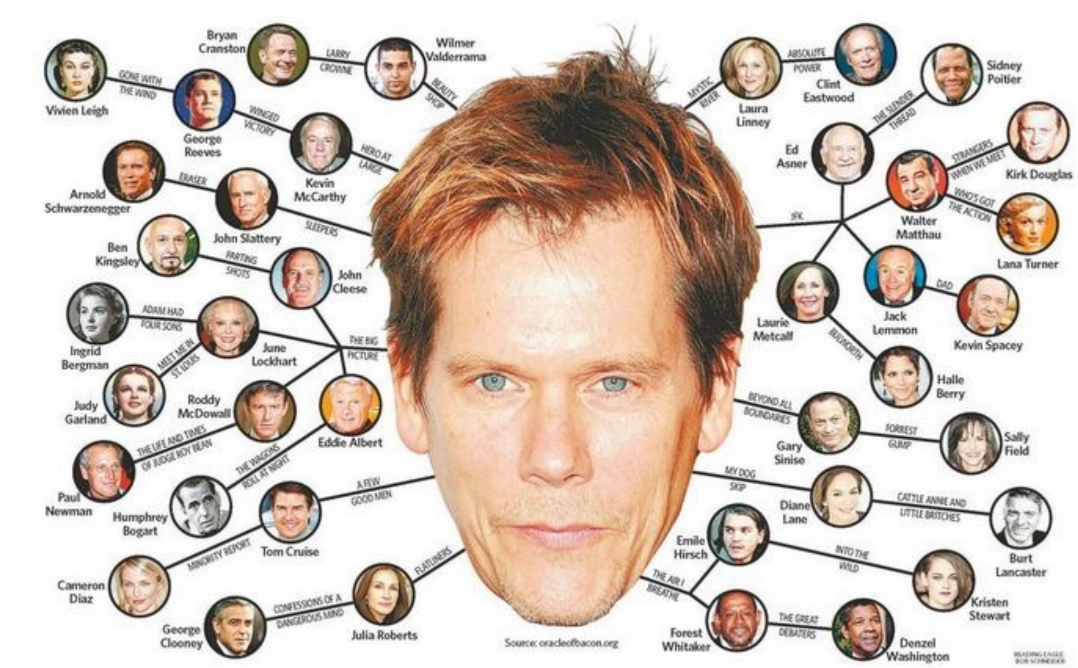


The Small World Problem

What is the average distance between two people?

Stanley Milgram's experiment (1967):

- 296 arbitrarily-selected letter "senders" in Boston and Omaha
- Ask "sender" to generate acquaintance chains to target a person in Boston ("the small world method")
- Mean number of intermediaries= 5.2 ("six degrees of separation")
- 48% of chains passed through 3 people



Small world effect: most pairs of vertices in most networks are connected by a short path.

Groups of Vertices

Many networks divide naturally into groups (*e.g.*, biochemical networks divide into functional modules, networks of people divide into social groups)

A **clique** is a maximal subset of vertices in an undirected network such that every member of the set is connected by an edge to every other. The occurrence of a clique in a sparse network indicates indicates a highly cohesive group. Cliques can overlap each other.

A **k-plex** relaxes the above requirement- a **k-plex** of size n is a maximal subset of n vertices such that each vertex is connected to at least n-k of the others. If k = 2, this is a clique.

A **k-core** is a maximal set of vertices such that each is connected to at least k others in the subset.

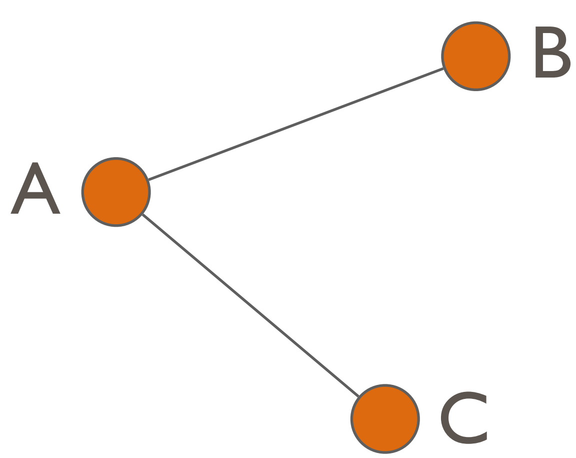
Network Clustering

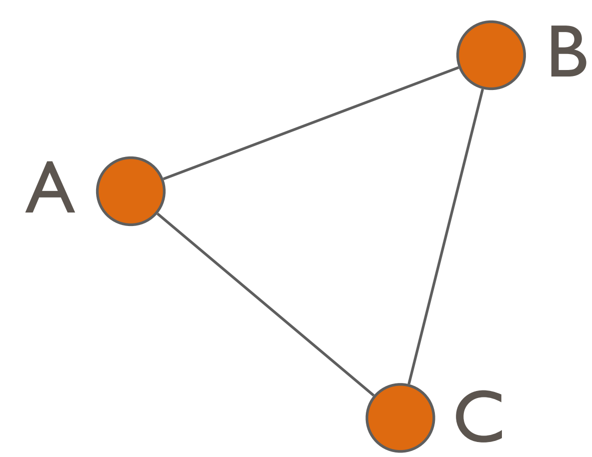
Network clustering (or **transitivity**) is the probability that two neighbors of a vertex will also connect to each other.

Networks with high transitivity are considered to have **local structure**.

A **connected triple** is a set of three nodes A, B, and C, such that A is connected to both B and C.

A **triangle** is a set of three nodes A, B, and C, such that all three are connected to each other.





Clustering coefficient

The **clustering coefficient** of a network is the fraction of triples that have their third edge filled to form a triangle:

$$C = \frac{3 \times \text{the number of triangles in the network}}{\text{number of connected triples of vertices}}$$

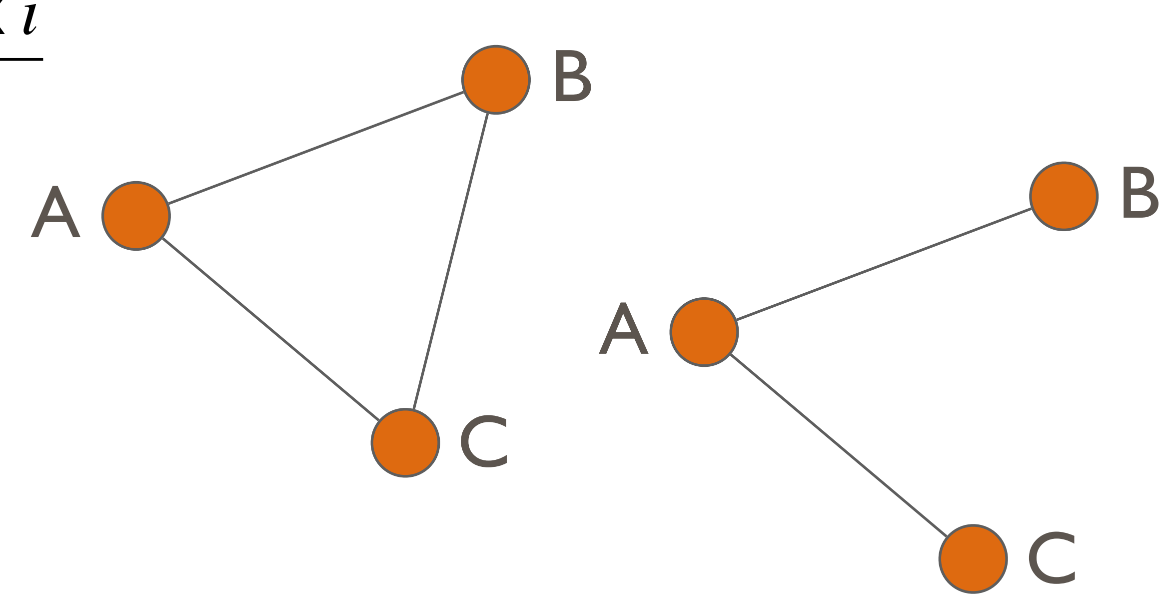
An alternative clustering coefficient starts by calculating the clustering at each node:

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}$$

$C_i = 0$ for nodes with degree 0 or 1

$$C_{WS} = \frac{1}{n} \sum_i C_i$$

weights low degree vertices more heavily



Graph Partitioning and Community Detection

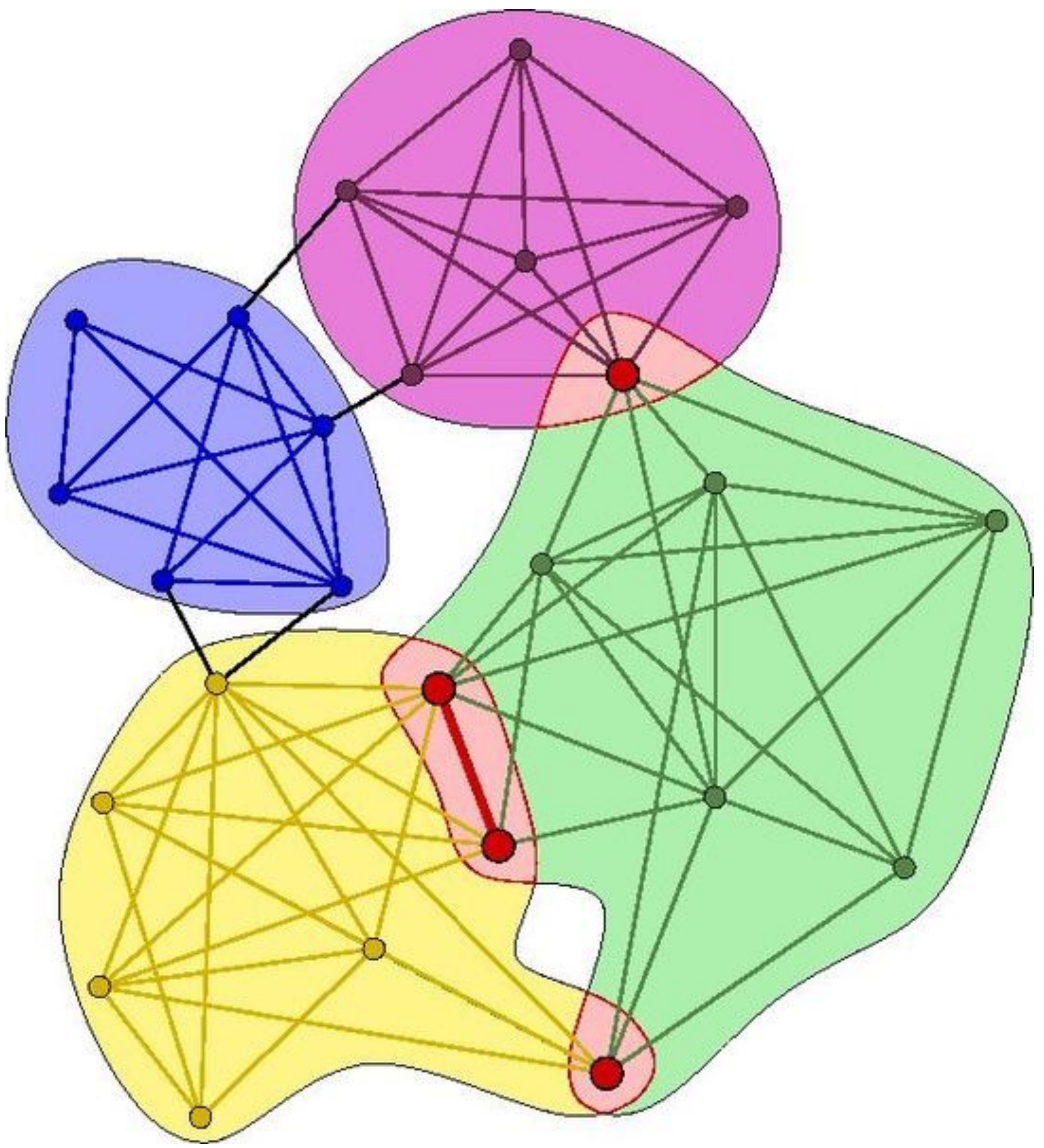
Graph partitioning divides the vertices of a network into a given number of non-overlapping groups of given sizes such that the number of edges between groups is minimized.

- Performed as a way to divide up network into smaller and more manageable pieces.

Community detection finds the natural fault lines along with a network separates. The group sizes and numbers are unspecified.

- Used as a tool to understand the structure of a network.

A network has **modularity** or **community structure** if its vertices fall into groups which have high densities of edges within them, and lower densities of edges between them.



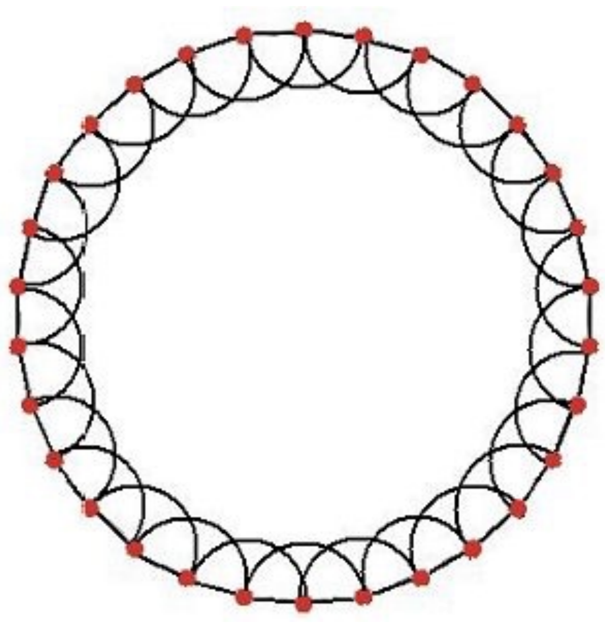
Network Models

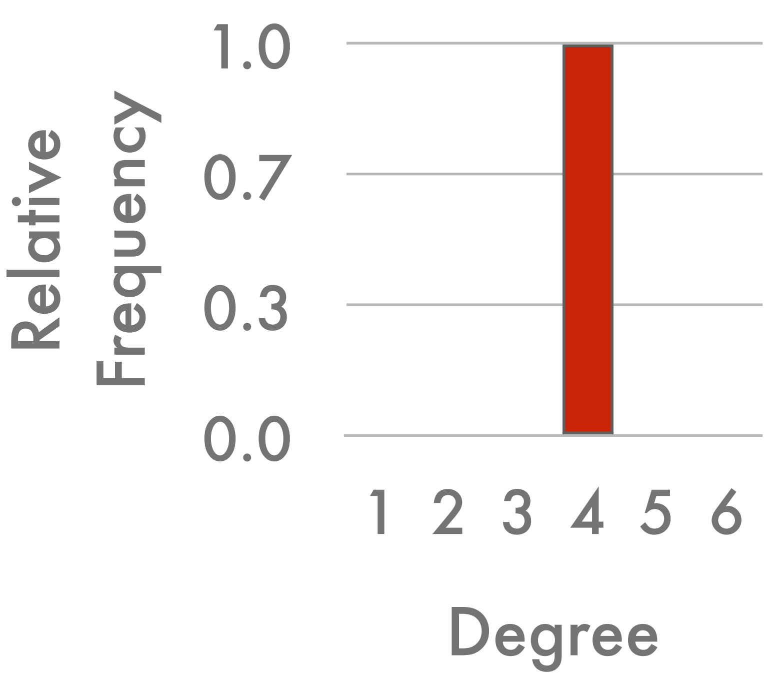
Lattice networks

Homogeneous degree distributions (**regular graphs**)

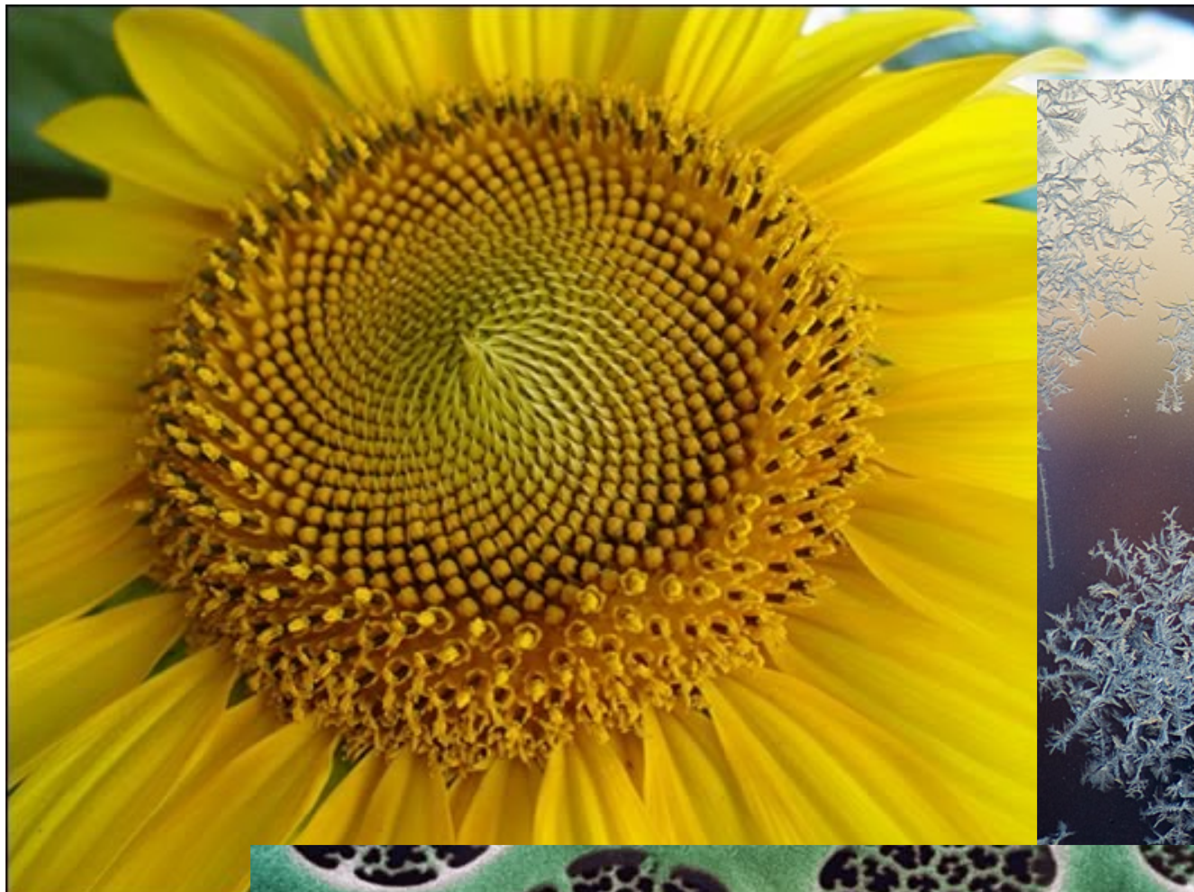
Regular graphs in which all vertices have degree *k* are called **k-regular**

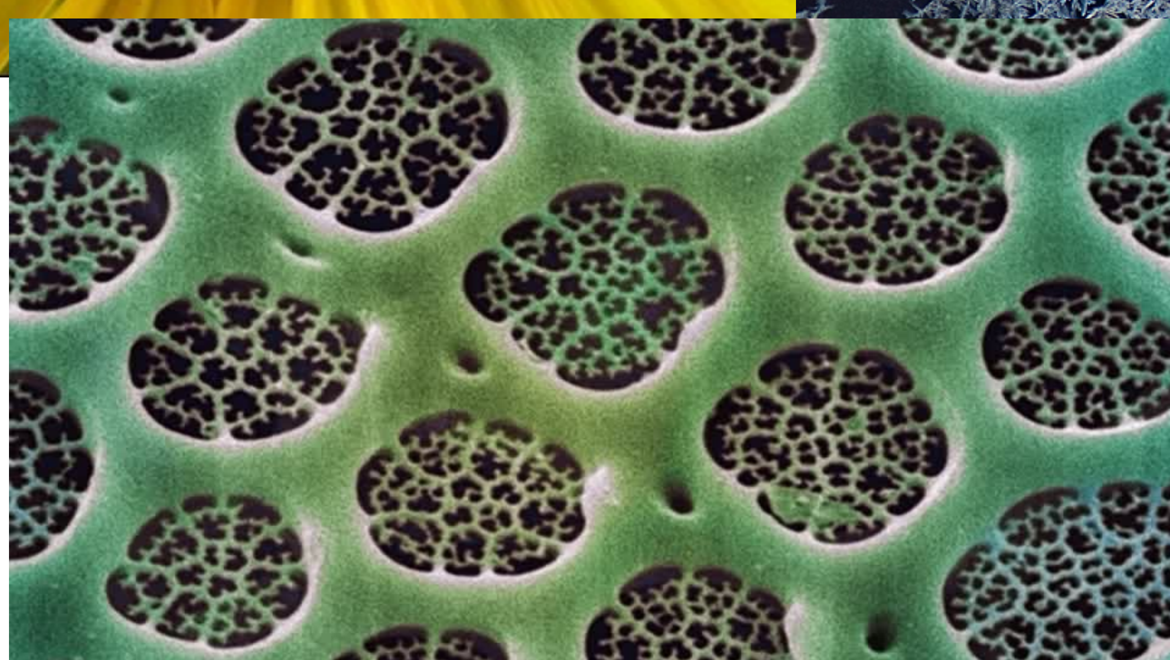
Spatially determined - edges link nearby vertices.





Lattices in nature





Erdös-Rényi random network

1. Create *n* vertices

2. For each pair of vertices *i* and *j*, create an edge (*i, j*) with probability *p*. The vertices will remain unconnected with probability 1-*p*.

The expected number of edges: $m = \binom{n}{2} p$

Each node has a degree between 0 and n-1

$$\Pr\{\text{degree } k\} = \Pr\{Y = k\} = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

binomial distribution

Erdös-Rényi random network

The structure of the network depends on *p.*

Random connections are non-spatial.

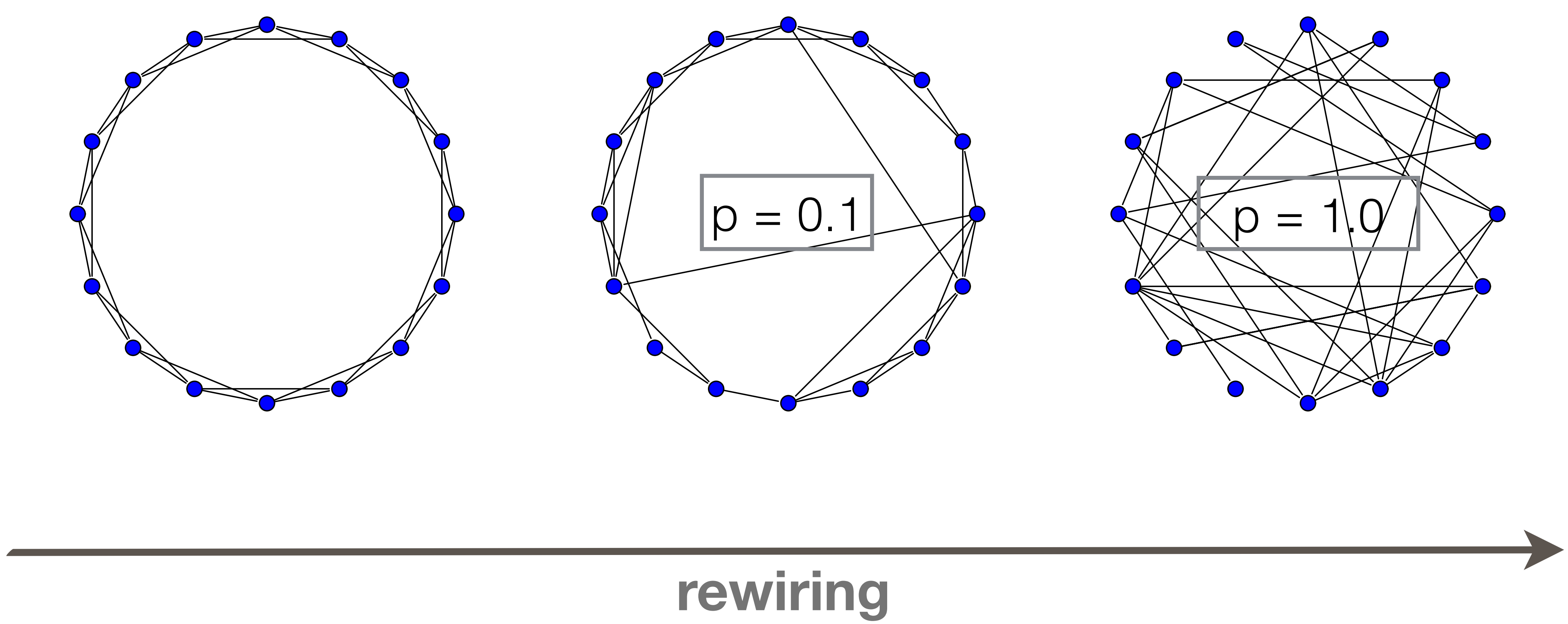


Figure by L.A. Meyers

Poisson Degree Distribution

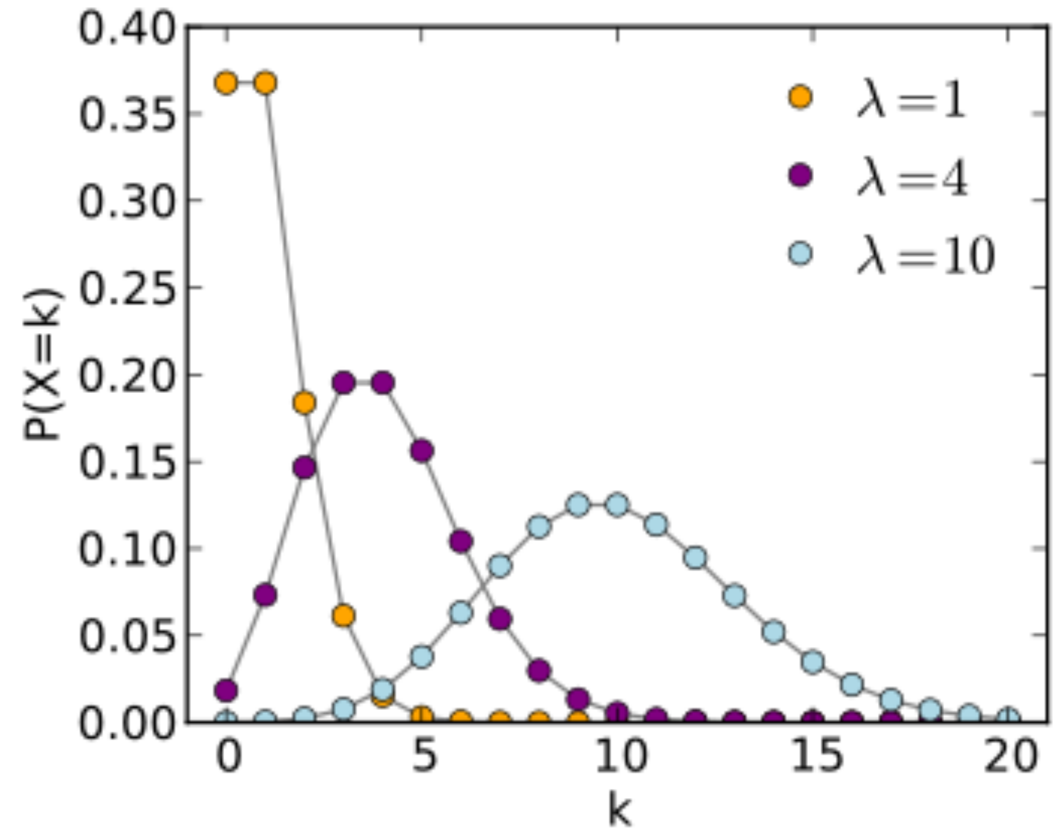
Erdös-Rényi networks are also called **Poisson random graphs**.

For large networks (large *n*) and low *p*, the binomial degree distribution becomes a Poisson degree distribution.

Generally, you can use the Poisson to approximate the binomial when the probability of the rare event $p \leq 0.05$ and the number of trials $n \geq 20$.

$$\Pr\{\text{degree } k\} \approx \frac{e^{-\lambda}\lambda^k}{k!}$$

where $\lambda = (n-1)p$ is the average degree of the network



Power Laws

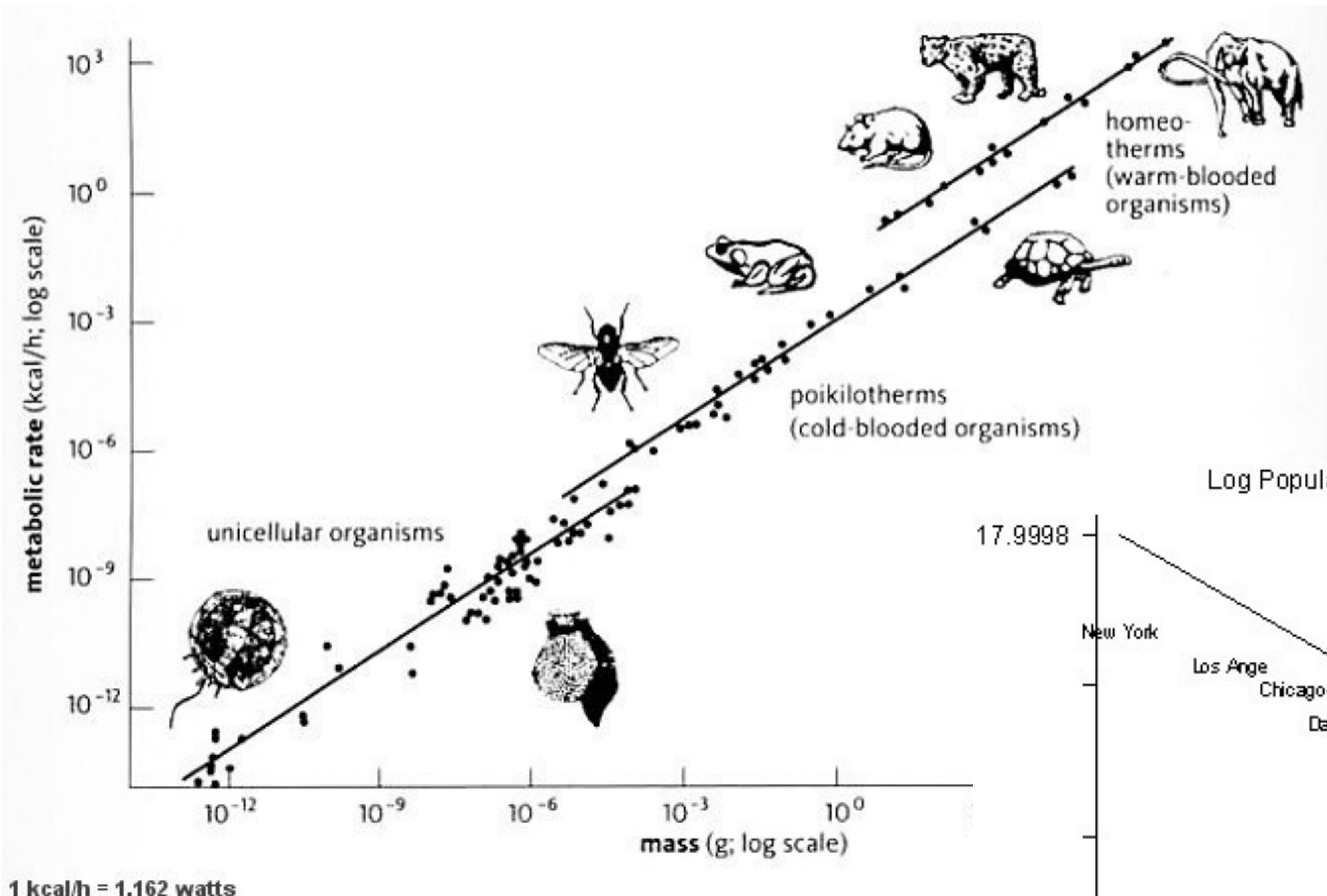
A power law function is of the form $f(x) = \beta \cdot x^{-\alpha}$

Get a linear function by taking the log of both sides:

$\log(f(x)) = \beta \cdot \log(x_{-\alpha}) = -\alpha \cdot \log(x) + \log(\beta)$

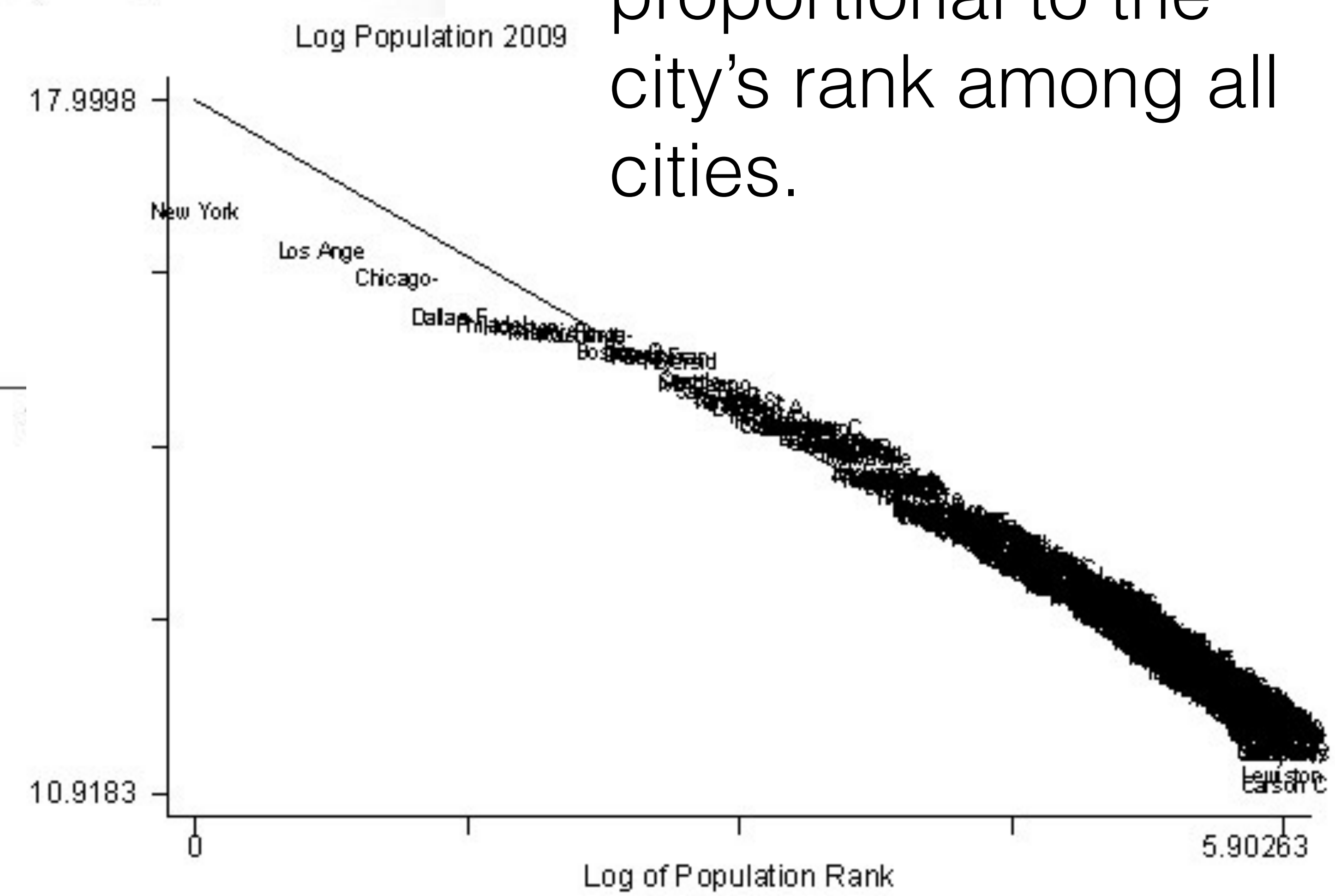
$y = \alpha \cdot x + b$

Power Laws



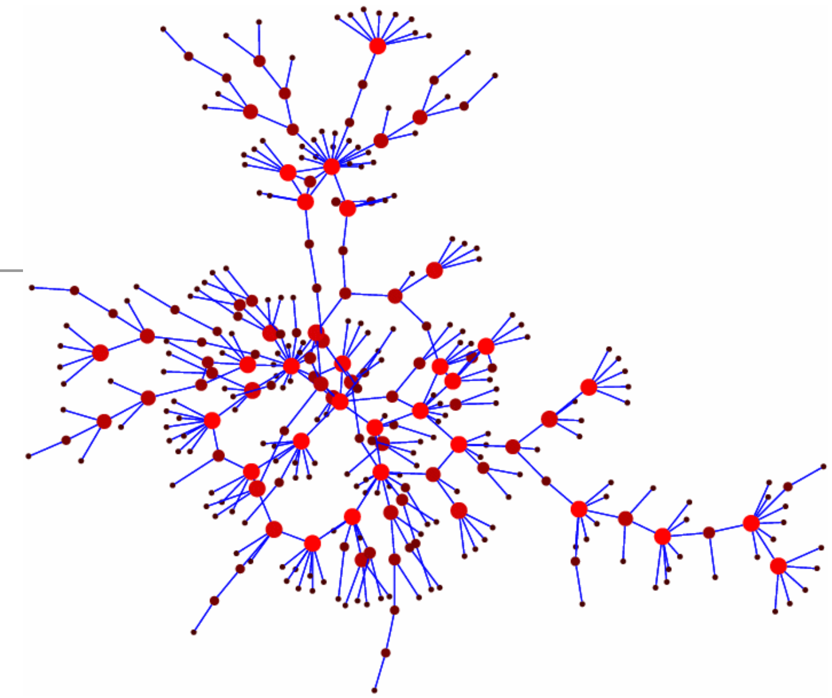
Kleiber's Law: an animal's metabolic rate scales to the ¾ power of the animal's mass.

Zipf's Law: the number of people in a city is inversely proportional to the city's rank among all cities.



http://economix.blogs.nytimes.com/2010/04/20/a-tale-of-many-cities/

Scale Free Networks

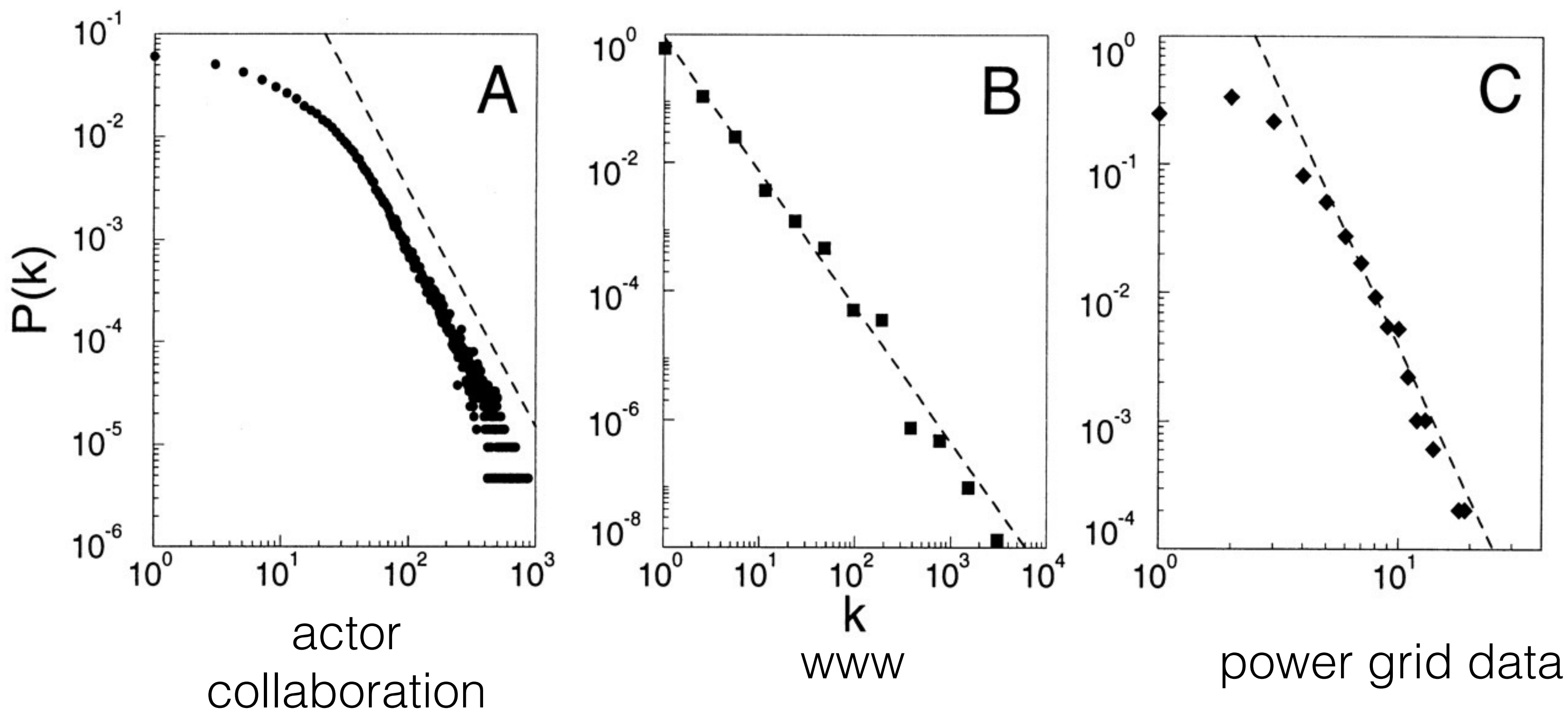


Scale free networks have power law degree distributions.

$$p_k \sim k^{-\gamma}$$

They are also called **power law networks**.

The vast majority of vertices have very low degree (**spokes**) while a small number of vertices have high degree (**hubs**).



Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." Science 286.5439 (1999): 509-512.

Scale free networks

Empirical networks show deviations from strict mathematical degree distributions.

Quick test for scale free network: make a log-log plot of the CDF and look for a straight line.

Scale free networks are often only power law in the tail of the distribution (for high values).

Can estimate the **exponent** of the power law: $\alpha = 1 + \frac{N}{\sum_{i=k_{\min}}^{k_{\max}} \ln \frac{k_i}{k_{\min} - \frac{1}{2}}}$

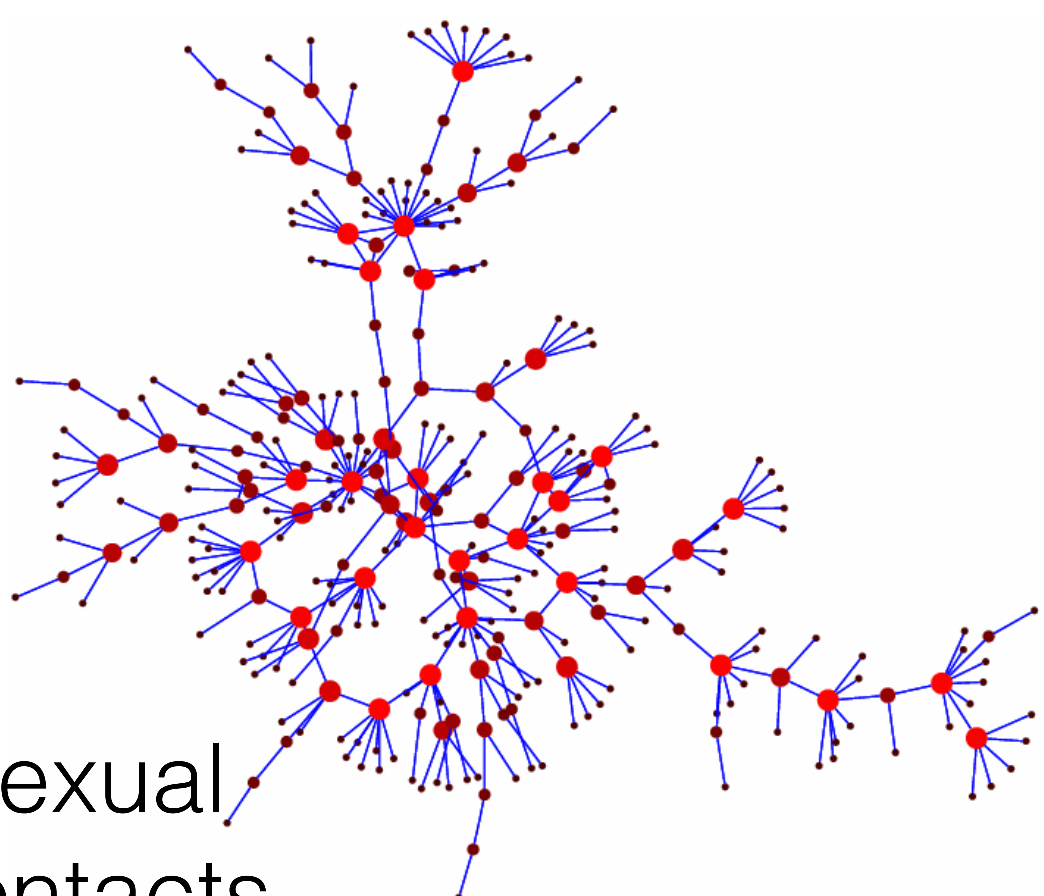
where k_{min} is the minimum degree for which a power law holds and k_{max} is the highest degree of the network.

Why are networks scale free?

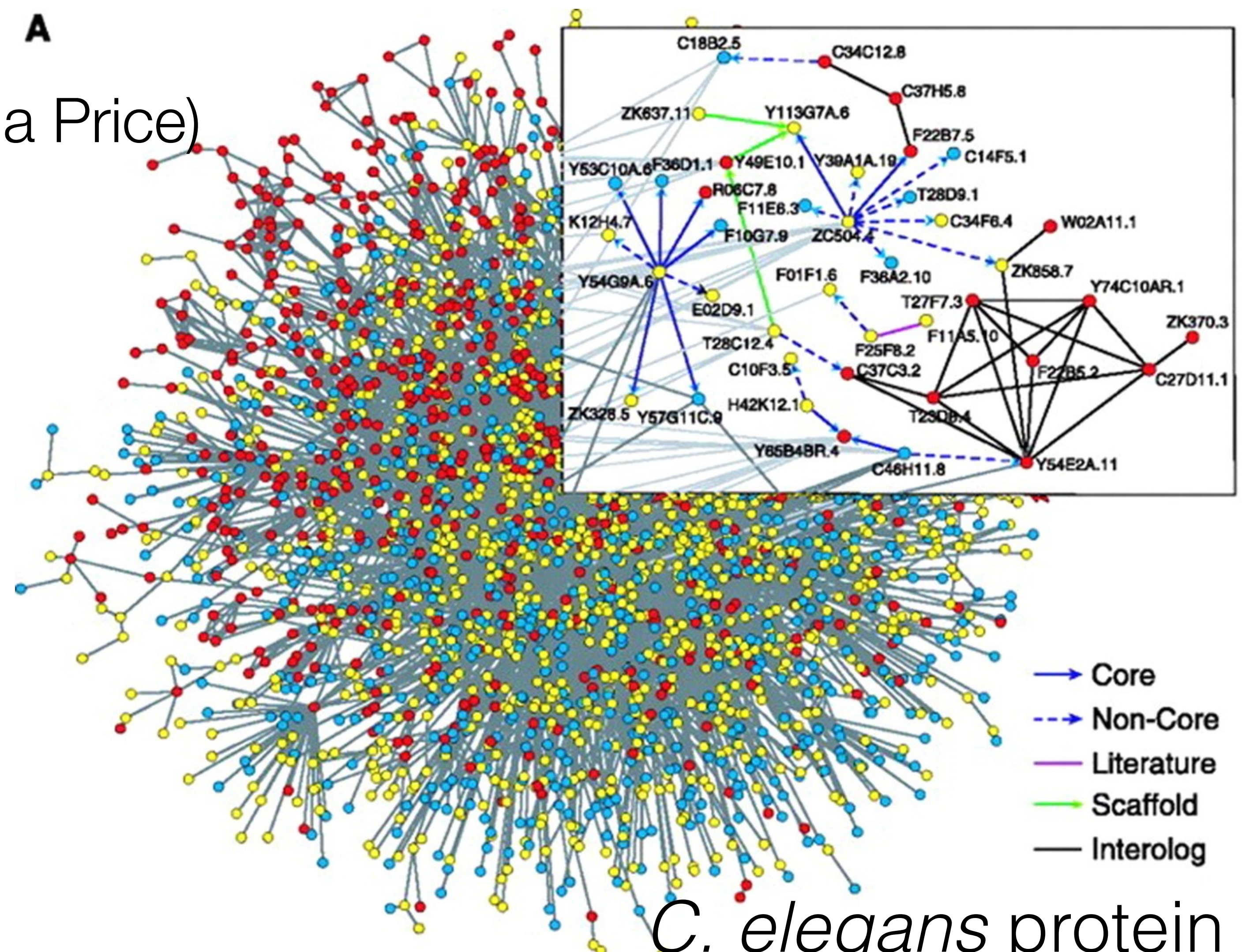
What natural processes potentially give rise to networks with power law degree distributions?

“The rich get richer” (Herbert Simon)

Cumulative advantage (Derek de Solla Price)



sexual contacts



C. elegans protein interaction network

Barabási-Albert Model (1999)

The **Barabási-Albert Model** of preferential attachments describes a simple and realistic process that produces scale free networks.

Growth: the network grows by adding vertices as a function of time.

Preferential attachment: edges are attached to existing vertices chosen at random weighted by the degree of each vertex.

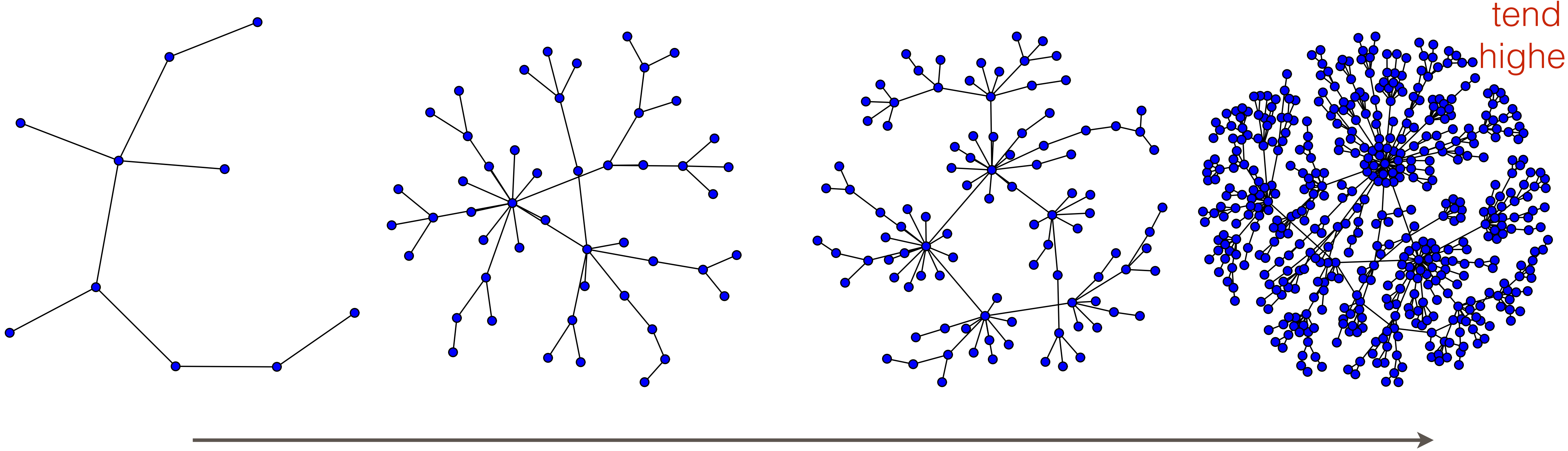


Figure by L.A. Meyers

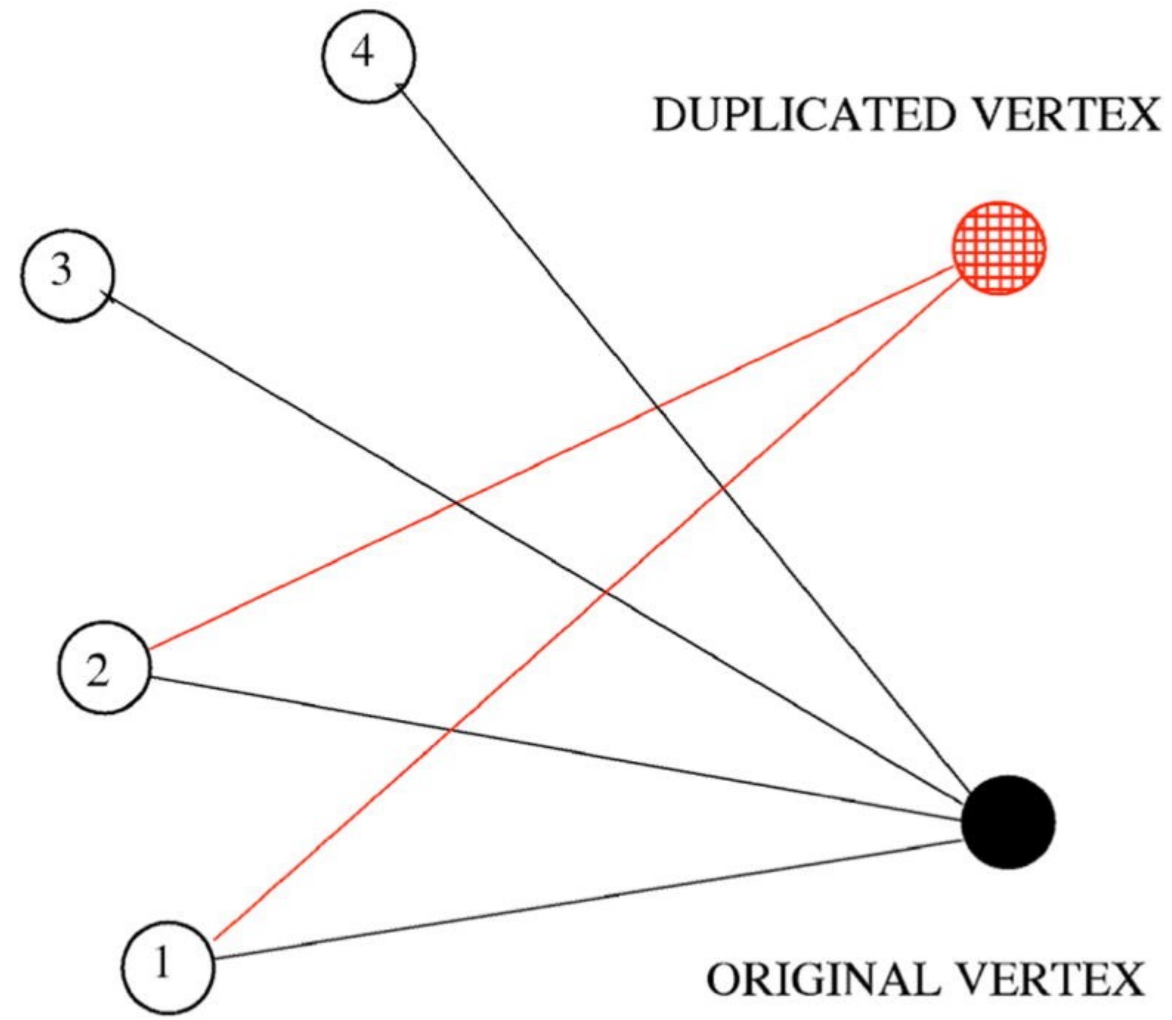
Duplication-divergence model

1. Duplication: A randomly chosen target node (~gene) is duplicated (i.e., its replica is introduced and connected to each neighbor of the target node). This represents the creation of new proteins that are initially identical to the old ones.

2. Divergence (new protein "survives" by acquiring a new function): Each link emanating from the replica is activated with retention probability σ. If at least one link is established, the replica is preserved; otherwise the attempt is considered as a failure and the network does not change.

Each node has at least one link and the network remains connected throughout the evolution.

Generates networks with power law distributions with exponent y_{DD}=2.5 (observed for yeast protein interaction network).



Ispolatov I, Krapivsky PL, Yuryev A. Duplication-divergence model of protein interaction network. Physical review. E, Statistical, nonlinear, and soft matter physics 2005;71(6 Pt 1):061911. doi:10.1103/PhysRevE.71.061911.

The Small World Model

Watts and Strogatz (1998) developed a simple model for the coexistence of clustering and small average path length.

Start with a one-dimensional ring lattice with *n* nodes where every node is connected to all nodes *k* or fewer steps away.

Rewire the network: For each edge, move one end to a new random location with probability p_r.

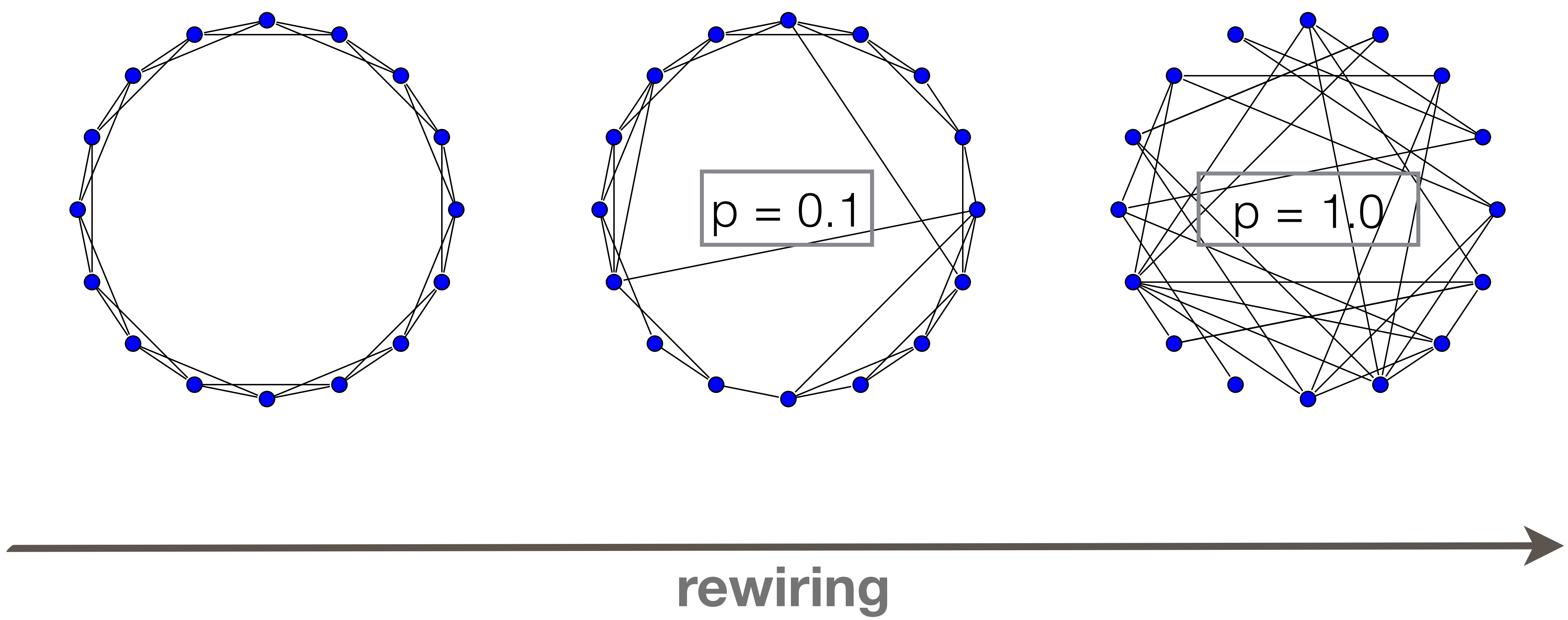


Figure by L.A. Meyers

The Small World Model

Clustering is unaffected by the addition of a few shortcuts

Average path length decreases dramatically with a few shortcuts

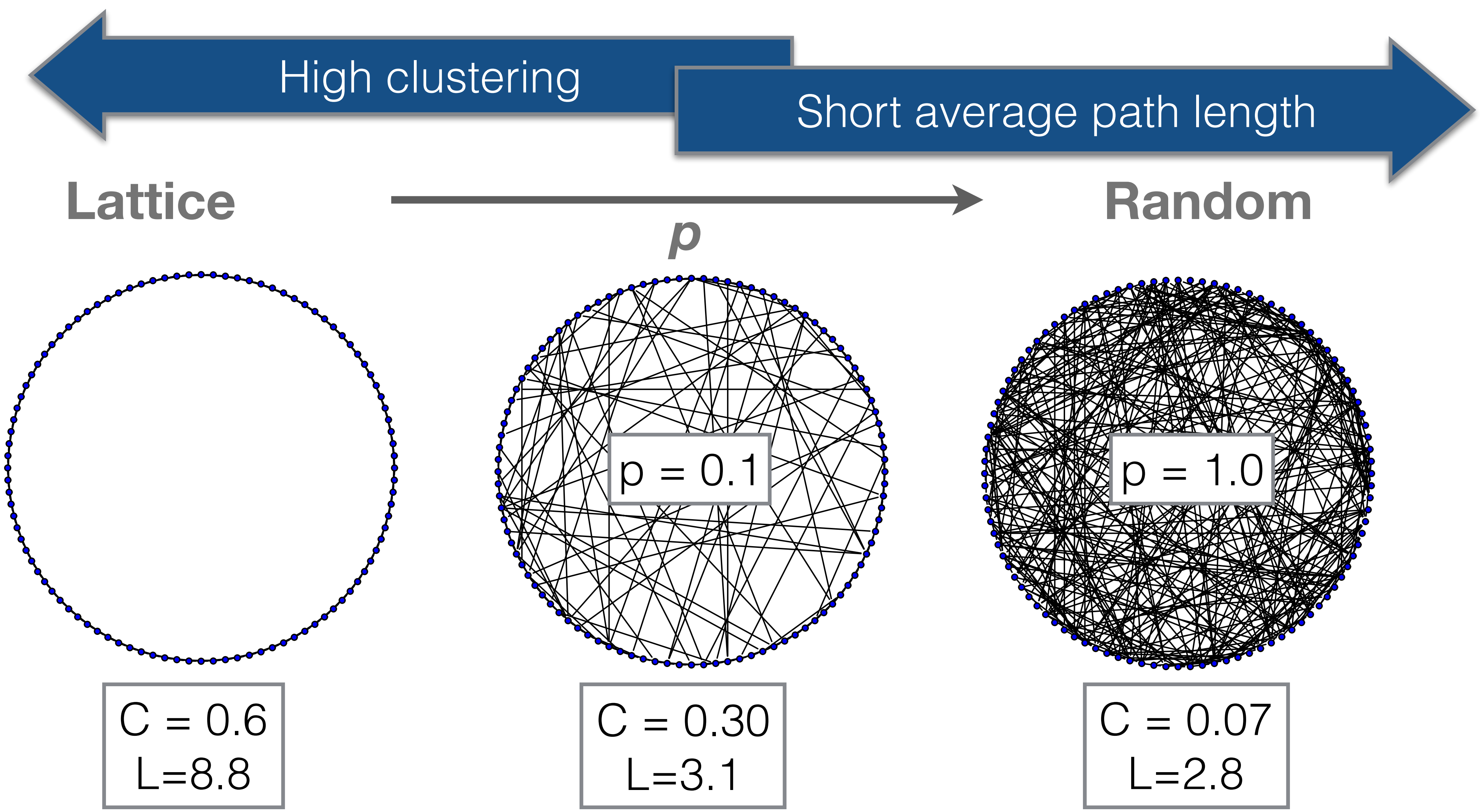


Figure by L.A. Meyers

Network Centrality Measures

Centrality

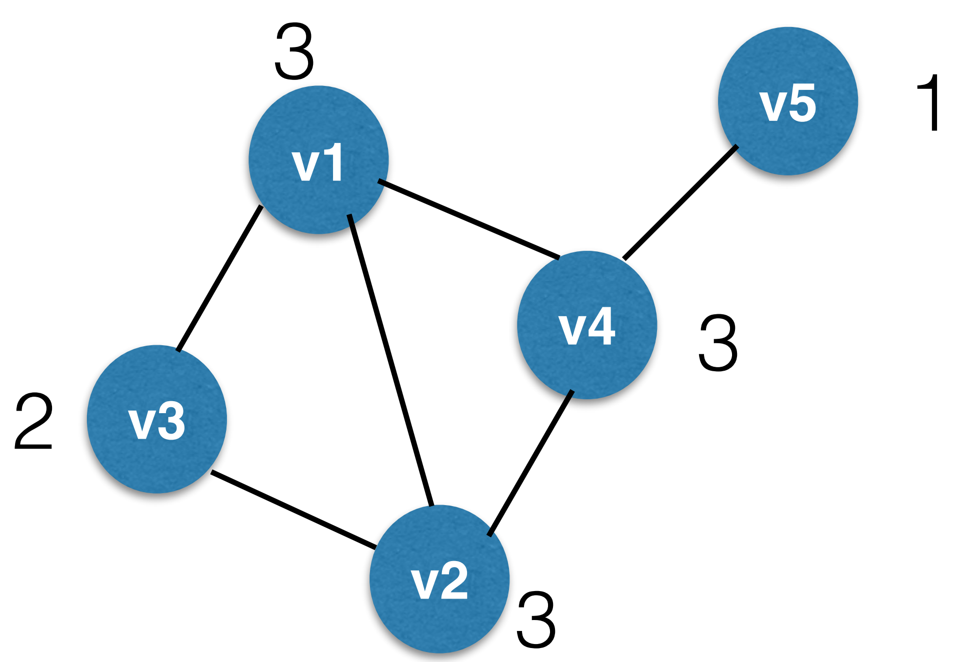
- What is the most important protein in a metabolic network?
- Which individuals should we target for vaccination?
- Which are the keystone species in an ecosystem?
- Which individuals have the most influence in a social network?
- Which papers have the greatest scientific impact?

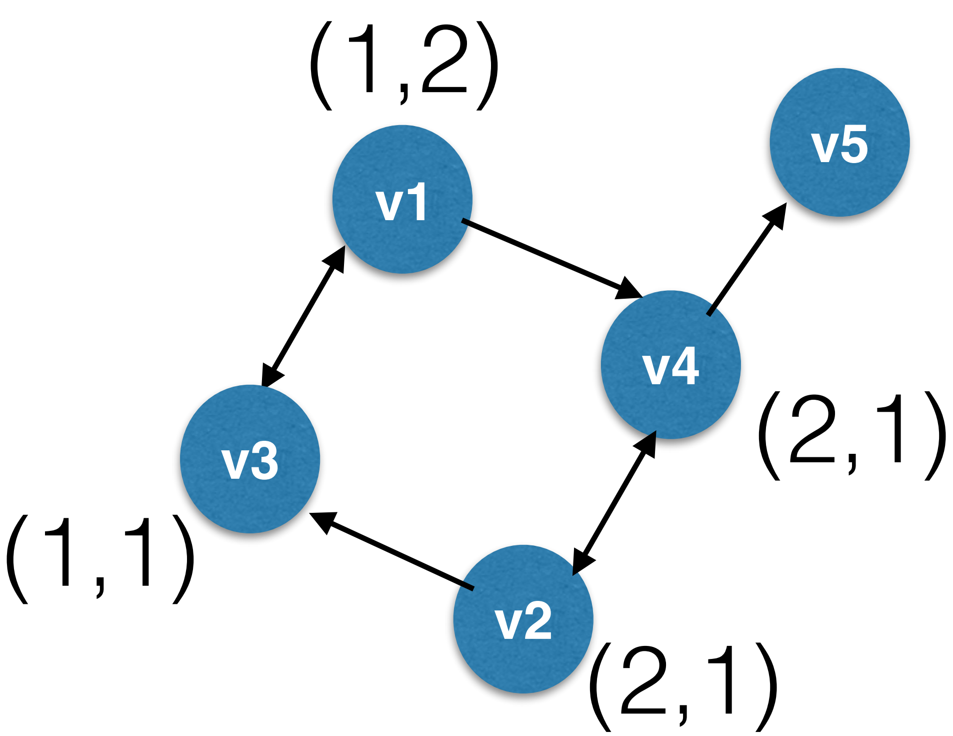
Centrality Rules

- We cannot compare centrality measures for different networks.
- We cannot compare different kinds of centrality measures on the same network.

Degree Centrality

- **Degree** is the number of edges connected to a vertex.
- Vertices have an **in-degree** and an **out-degree** in directed networks.





Eigenvector Centrality

A vertex's importance can be increased by having connections to other vertices that are *themselves* important.

The **eigenvector centrality** of a vertex is proportional to the sum of the eigenvector centralities of its neighbors.

Works best in the case of undirected networks.

leading eigenvalue of A

$$A\vec{c} = \kappa_1 \vec{c}$$

vector of all vertex centralities

$$c_i = \frac{1}{\kappa_1} \sum_j A_{ij} c_j$$

centrality of vertex *i*

Eigenvector Centrality Issues

A directed network has an asymmetric adjacency matrix, and thus has two sets of eigenvectors (and two leading eigenvalues). Typically use the right eigenvectors- represents other vertices pointing towards each vertex.

Vertices with only out-degree have centrality zero.

Only vertices in strongly connected components can have non-zero eigenvector centrality (thus, this measure is useless for acyclic networks).

Solutions: **Katz centrality** (each vertex gets a small amount of centrality "for free"), **PageRank centrality** (variation of Katz; centrality derived from neighbors is proportional to their centrality *divided by their out-degree*).

Closeness Centrality

Closeness centrality measures the mean distance from a vertex to other vertices.

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}}$$

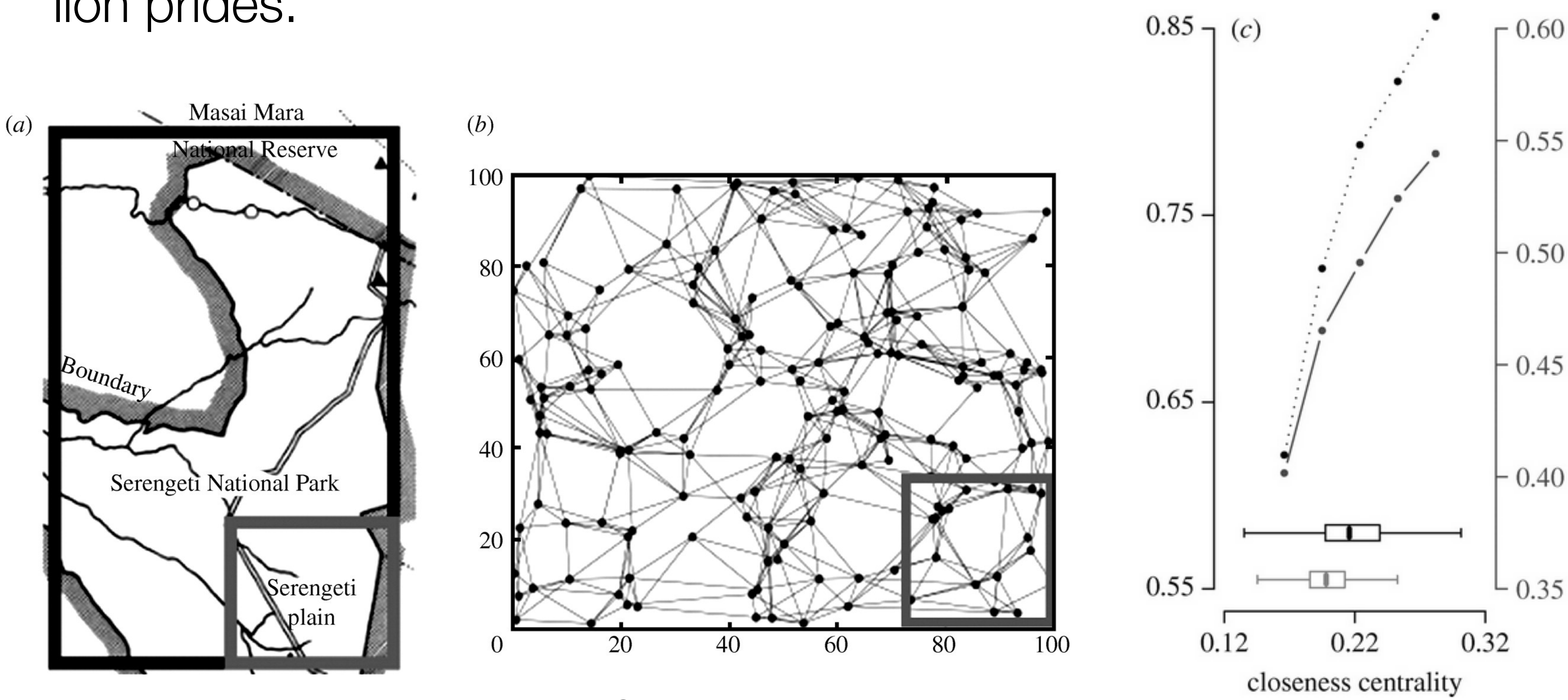
inverse of mean geodesic distance

number of vertices divided by sum of geodesic distances from *i* to *j*

Issues: CC values tend to span a small dynamic range from largest to smallest; small fluctuations in network structure can change the order of the values substantially; component size influences CC (use harmonic mean distance for disconnected networks)

Closeness Centrality

Epidemiological risk correlates with closeness centrality for Serengeti lion prides.



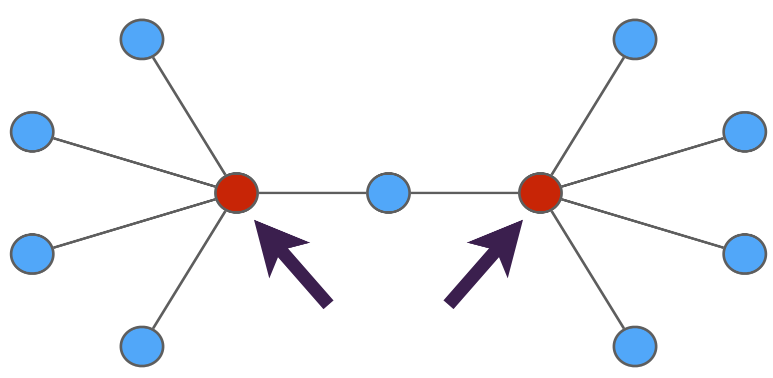
The ecosystem and study area in Serengeti National Park and a simulated lion population based on estimates of territory locations and adjacencies from Serengeti Lion Project data.

Craft, Meggan E., et al. "Distinguishing epidemic waves from disease spillover in a wildlife population." Proceedings of the Royal Society B: Biological Sciences (2009): rspb-2008.

Betweenness Centrality

Betweenness centrality measures the extent to which a vertex lies on paths between other vertices.

Example: a social network with messages, information, or disease passed from one person to another. If “messages” take the shortest path and are passed at the same rate, the number passing through each vertex is proportional to the number of geodesics the vertex lies on.



Betweenness Centrality

σ_{ij} number of shortest paths between vertices i and j

$\sigma_{ij}(v)$ number of those paths that pass through vertex v

$$\delta_{ij}(v) = \sigma_{ij}(v)/\sigma_{ij}$$

$$C(v) = \sum_{i \neq v} \sum_{j > i, j \neq v} \delta_{ij}(v)$$

Vertices with high BC have considerable influence by virtue of their control over “information” passing between others.

The removal of these vertices from the network will also cause the most disruption.