دانشگاه کاشان

Complex Network Theory

Lecture 2-2

Basic network concepts and metrics

Instructor: S. Mehdi Vahidipour

(Vahidipour@kashanu.ac.ir)

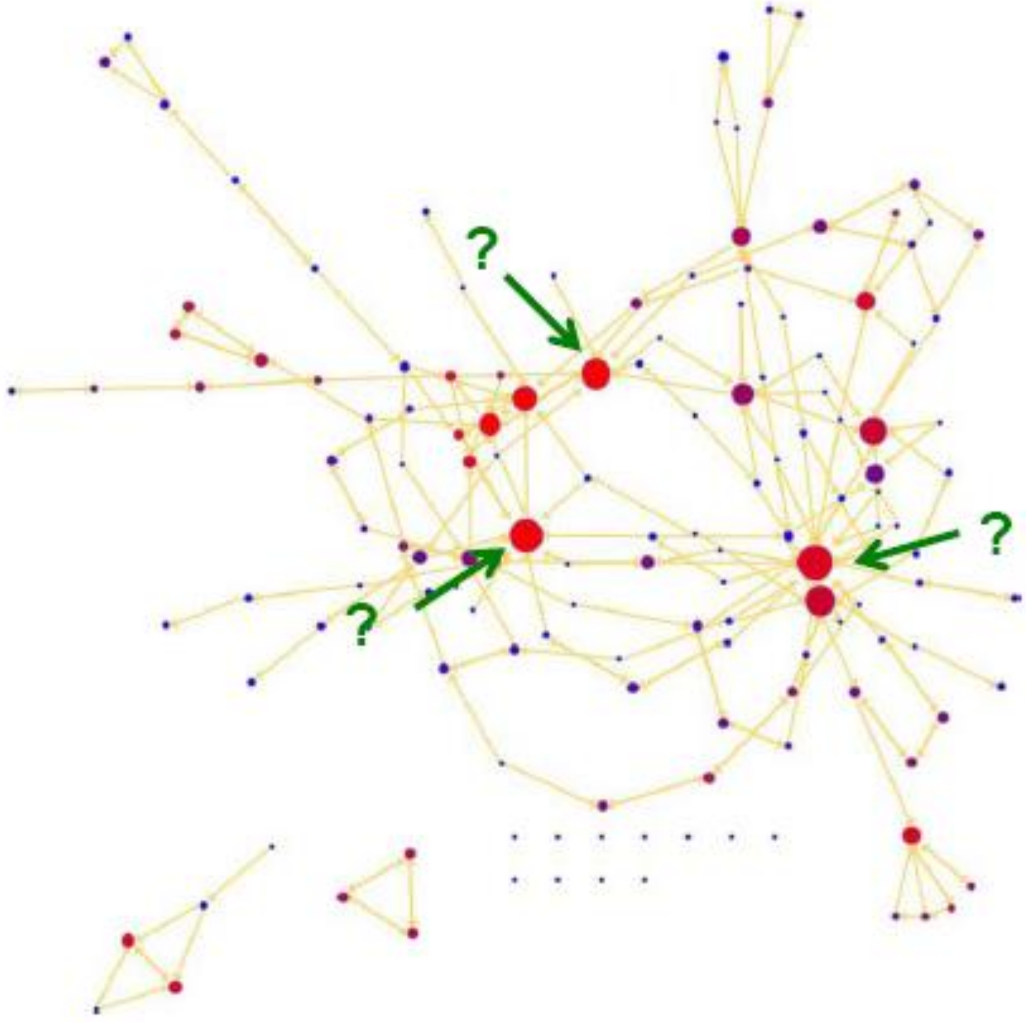
Spring 2018

Thanks A. Rezvanian

A. Barabasi, L. Adamic and J. Leskovec

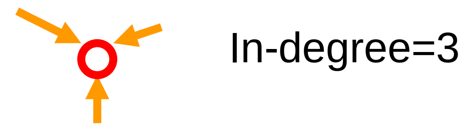
Who is most central?

- Who is most important?



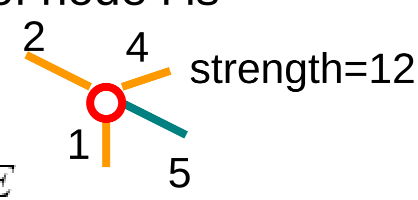
Nodes

- Node network properties
  - from immediate connections
    - In-degree (directed)
      how many directed edges (arcs) are incident on a node



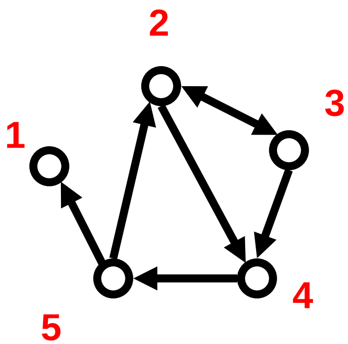
    - Out-degree (directed)
      how many directed edges (arcs) originate at a node

    - degree (in or out) - undirected
      number of edges incident on a node
  - In weighted networks instead of degree, strength of nodes are defined
  - If the weighted adjacency matrix is $W=(w_{ij})$, the strength of node i is defined as



    - $s_i = \sum_{j=1}^{n} W_{ij}$
  - Average degree (Avg. degree) $\bar{k} = \langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N}$

Node degree from matrix values



- Out-degree = $\sum_{j=1}^{n} A_{ij}$

example: out-degree for node 3 is 2, $\sum_{j=1}^{n} A_{3j}$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

- In-degree = $\sum_{i=1}^{n} A_{ij}$

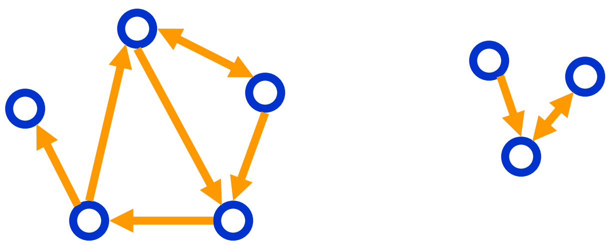
example: the in-degree for node 3 is 1, $\sum_{i=1}^{n} A_{i3}$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

- Average In-degree = Average Out-degree ?

Network metrics: degree sequence and distribution

- Degree sequence: An ordered list of the (in,out) degree of each node
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]

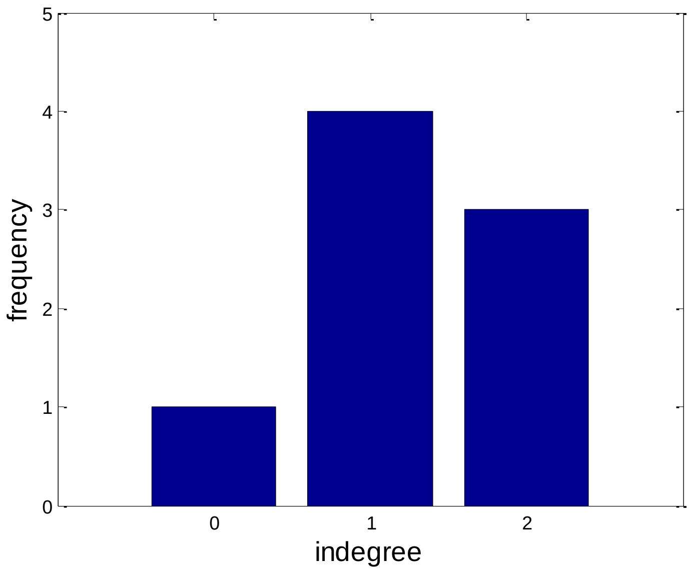


- **Degree distribution:** A frequency count of the occurrence of each degree
- **Degree distribution P(k):** Probability that a randomly chosen node has degree **k**

  N_k = # nodes with degree **k**
- **Normalized** histogram (PDF):

  $P(k) = N_k / N$

  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]



Network metrics: Density

- The **maximum number of edges** in an undirected graph on ***N*** nodes is

$$E_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

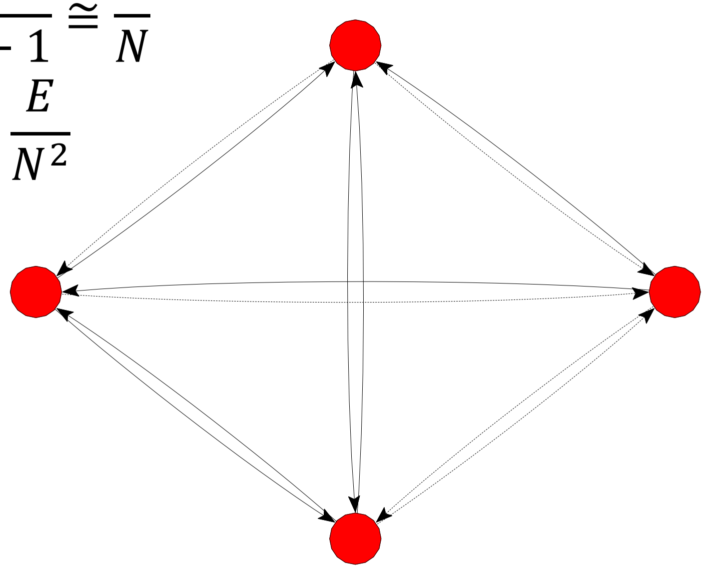
- A graph with the number of edges ***$E = E_{max}$*** is a **complete graph**

- *density* of a graph:

$$\rho = \frac{E}{E_{max}} = \frac{2E}{N(N-1)} = \frac{\overline{K}}{N-1} \cong \frac{\overline{k}}{N}$$

$$\rho = \frac{E}{E_{max}} = \frac{2E}{N(N-1)} \approx \frac{E}{N^2}$$

For example, out of 12 possible connections, this graph has 7, giving it a density of 7/12 = 0.583



Most real-world networks are sparse

$$E \ll E_{max} \quad (\text{or } \bar{k} \ll N\text{-}1)$$

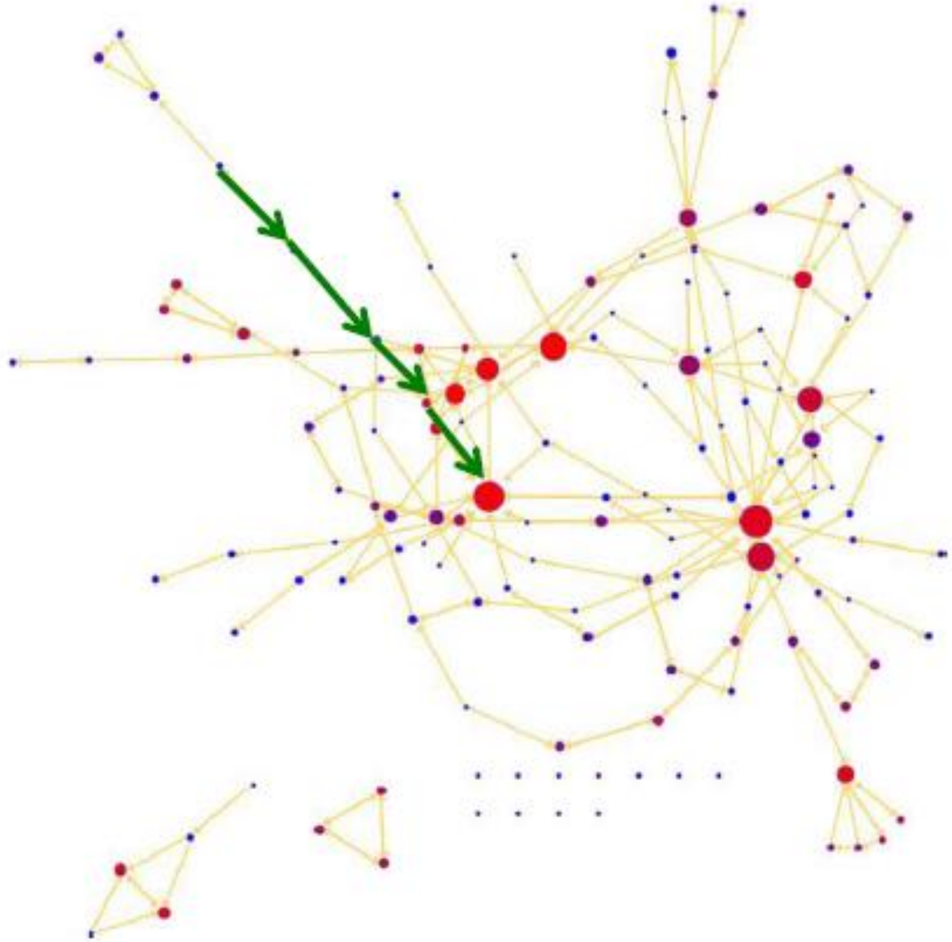
WWW (Stanford-Berkeley):	N=319,717	$\langle k \rangle$=9.65
Social networks (LinkedIn):	N=6,946,668	$\langle k \rangle$=8.87
Communication (MSN IM):	N=242,720,596	$\langle k \rangle$=11.1
Coauthorships (DBLP):	N=317,080	$\langle k \rangle$=6.62
Internet (AS-Skitter):	N=1,719,037	$\langle k \rangle$=14.91
Roads (California):	N=1,957,027	$\langle k \rangle$=2.82
Proteins (S. Cerevisiae):	N=1,870	$\langle k \rangle$=2.39

(Source: *Leskovec et al., Internet Mathematics, 2009*)

Consequence: Adjacency matrix is filled with zeros!

(**Density of the matrix (E/N^2):** WWW=1.51×10^{-5}, MSN IM = 2.27×10^{-8})

How far apart are nodes?

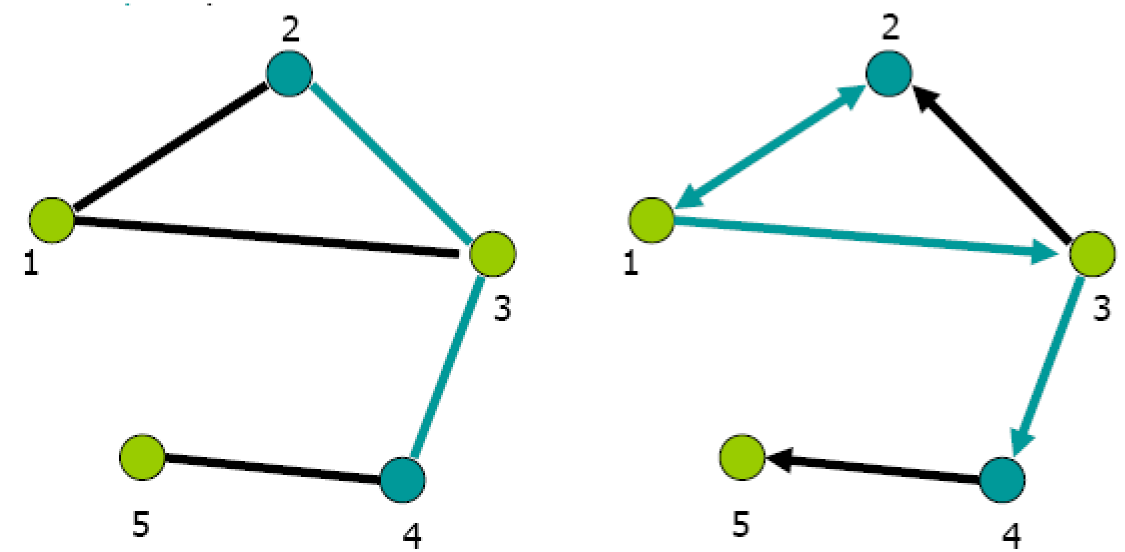


Paths

- Path from node *i* to node *j*: a sequence of edges (directed or undirected from node *i* to node *j*)

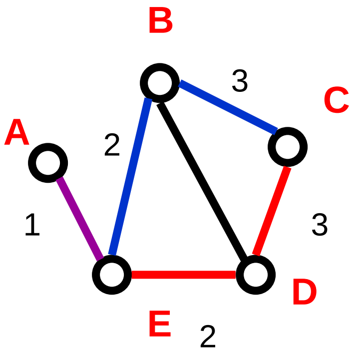
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

  - path length: number of edges on the path (unweighted networks)
  - nodes *i* and *j* are connected
  - Cycle (loop): a path that starts and ends at the same node
  - Self-loop: a path from a node to itself

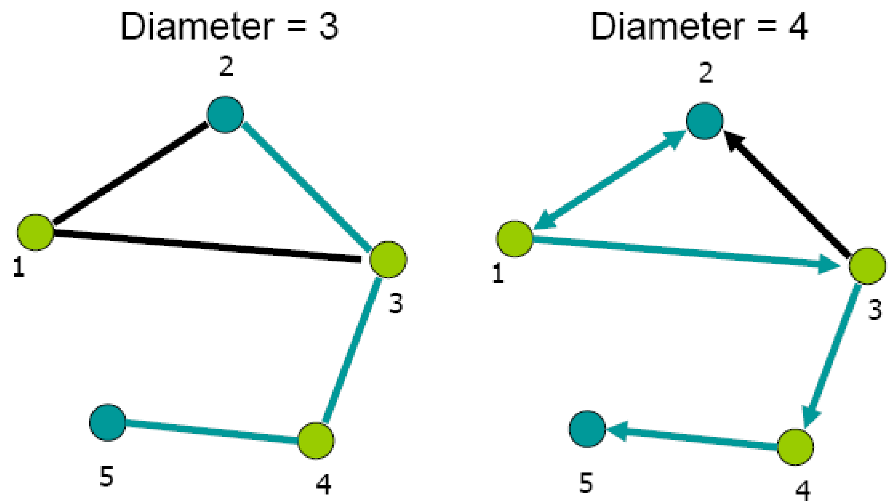


Network metrics: shortest paths

- **Shortest path** (also called a geodesic path, BFS path)
  - The shortest sequence of links connecting two nodes
  - Not always unique
  - A and C are connected by 2 shortest paths
    - A – E – B - C
    - A – E – D - C



- **Diameter**: the largest geodesic distance in the graph (Maximum shortest path)
  - The distance between A and C is the maximum for the graph: 3



- Caution: some people use the term 'diameter' to be the average shortest path distance, in this class we will use it only to refer to the maximal distance

Network metrics: shortest paths

- **Average path length** for a connected graph (component) or a strongly connected (component of a) directed graph

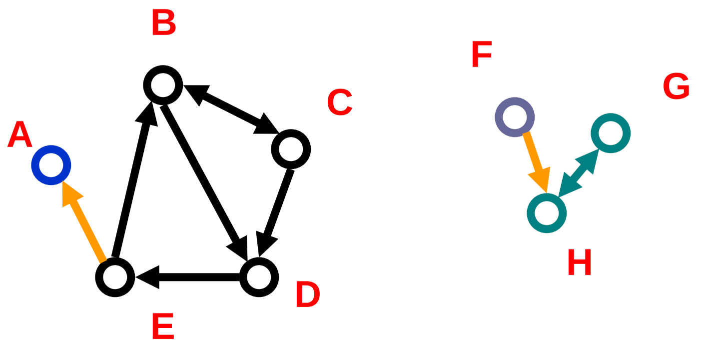
$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{ij}$$

where h_{ij} is the distance from node i to node j

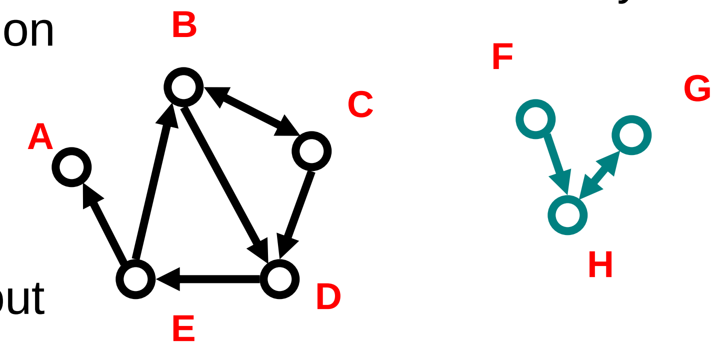
  - Many times we compute the average only over the connected pairs of nodes (we ignore “infinite” length paths)

Network metrics: connected components

- **Connected graph**: a graph where every pair of nodes is connected
- **Disconnected graph**: a graph that is not connected
- **Connected Components:** subsets of vertices that are connected
- **Strongly connected components:** Each node within the component can be reached from every other node in the component by following directed links.
  - Strongly connected components
    - B C D E
    - A
    - G H
    - F



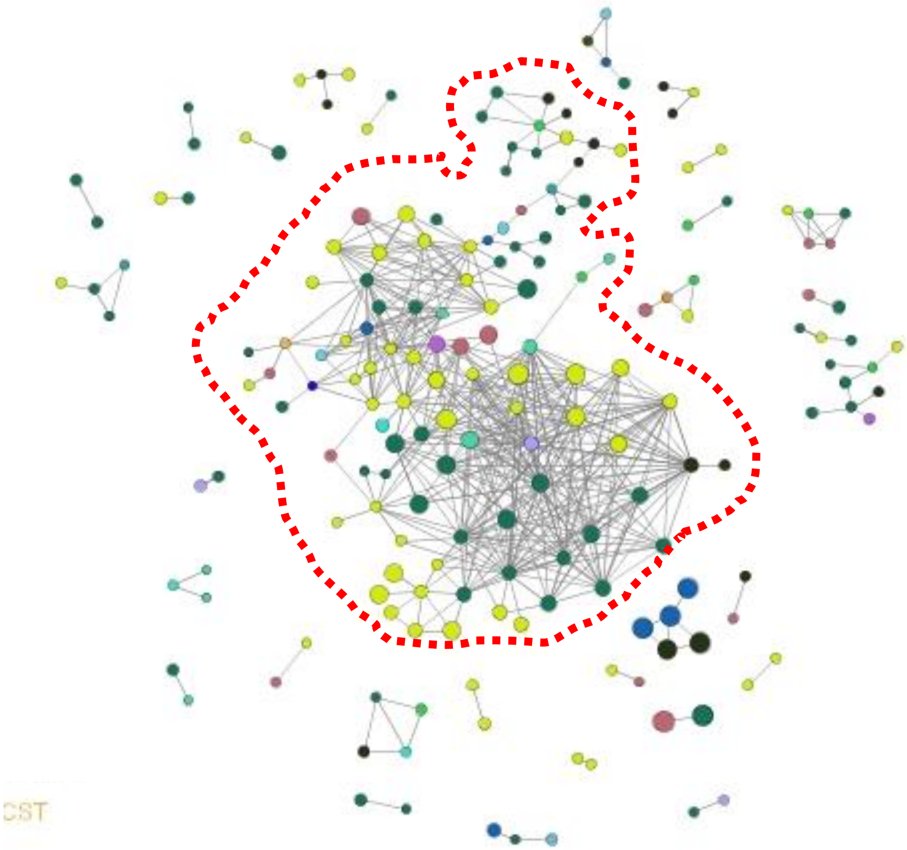
- **Weakly connected components**: every node can be reached from every other node by following links in either direction
  - Weakly connected components
    - A B C D E
    - G H F



- In undirected networks one talks simply about 'connected components'

Giant components and the web graph

- **Largest Connected Component**: the connected component with the largest number of nodes
- if the largest component encompasses a significant fraction of the graph, it is called the **giant component**



The bowtie model of the web

- The Web is a directed graph:
  - webpages link to other webpages
- The connected components tell us what set of pages can be reached from any other just by surfing (no ‘jumping’ around by typing in a URL or using a search engine)
- Broder et al. 1999 – crawl of over 200 million pages and 1.5 billion links.
- SCC – 27.5%
- IN and OUT – 21.5%
- Tendrils and tubes – 21.5%
- Disconnected – 8%

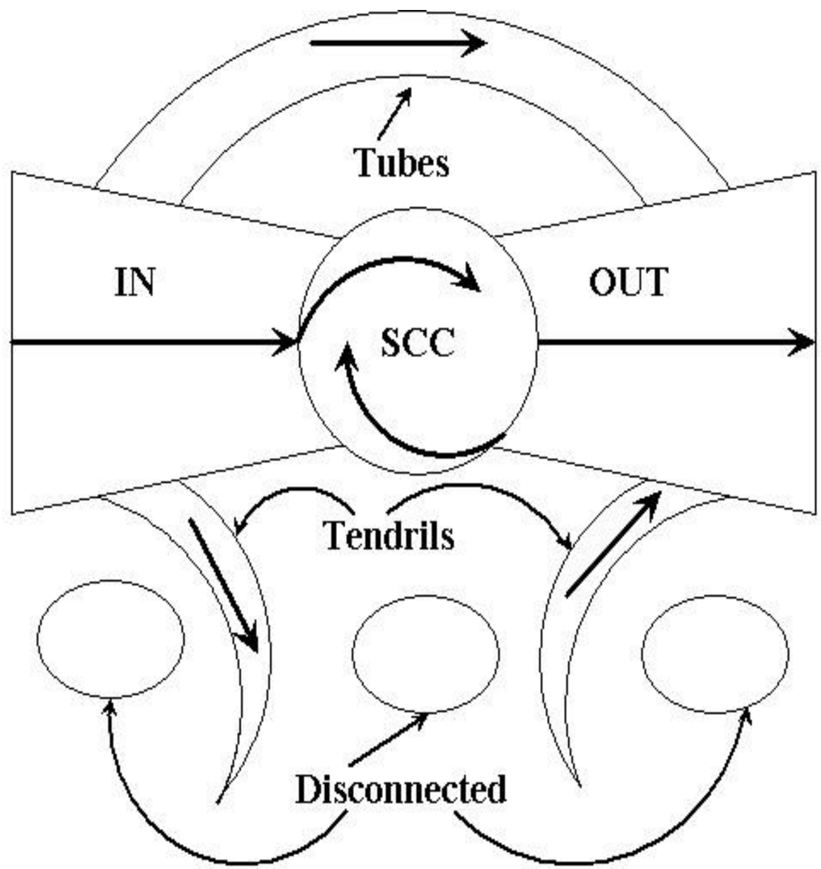
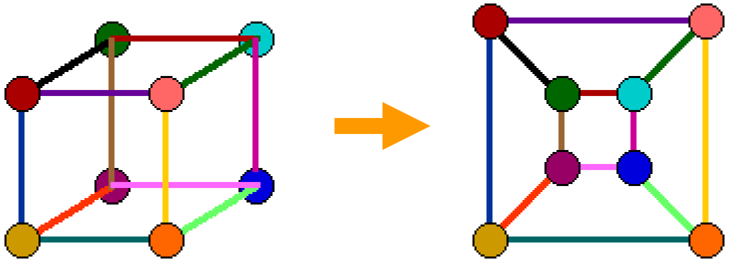


image: Mark Levene

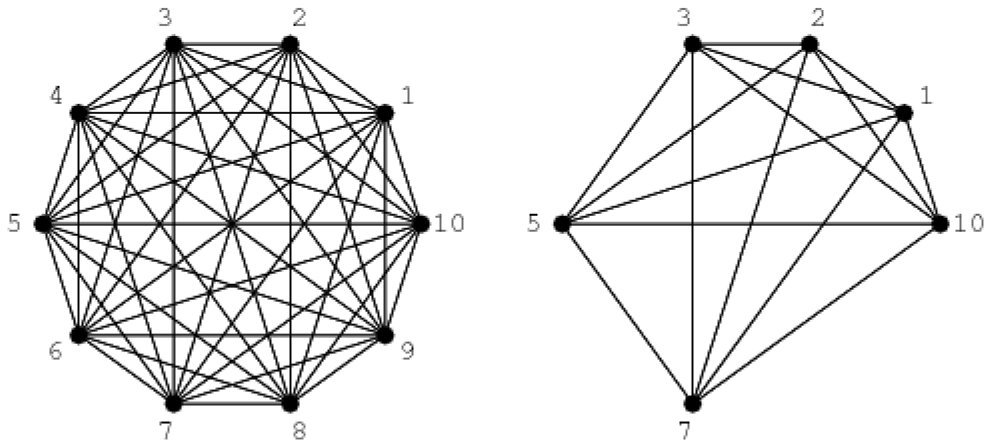
Planar graphs

- A graph is planar if it can be drawn on a plane without any edges crossing



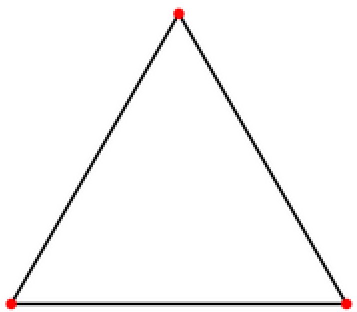
Subgraphs

- **Subgraph**: Given V' $\subset$ V, and E' $\subset$ E, the graph G'=(V',E') is a subgraph of G.
- **Induced subgraph**: Given V' $\subset$ V, let E' $\subset$ E is the set of all edges between the nodes V' in G. The graph G'=(V',E'), is an induced subgraph of G.

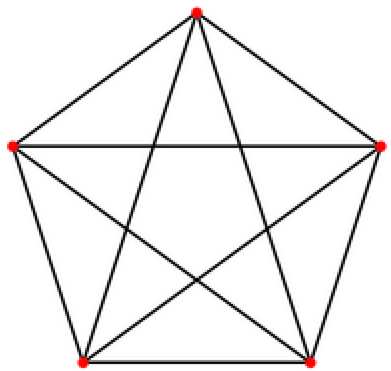


Cliques and complete graphs

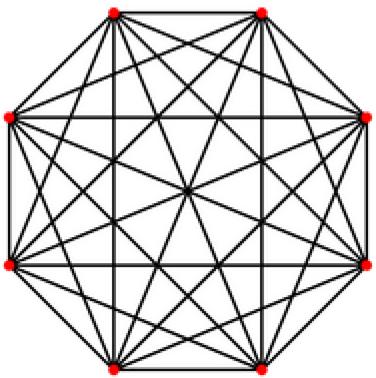
- K_n is the complete graph (clique) with K vertices
  - each vertex is connected to every other vertex
  - there are n*(n-1)/2 undirected edges



K_3



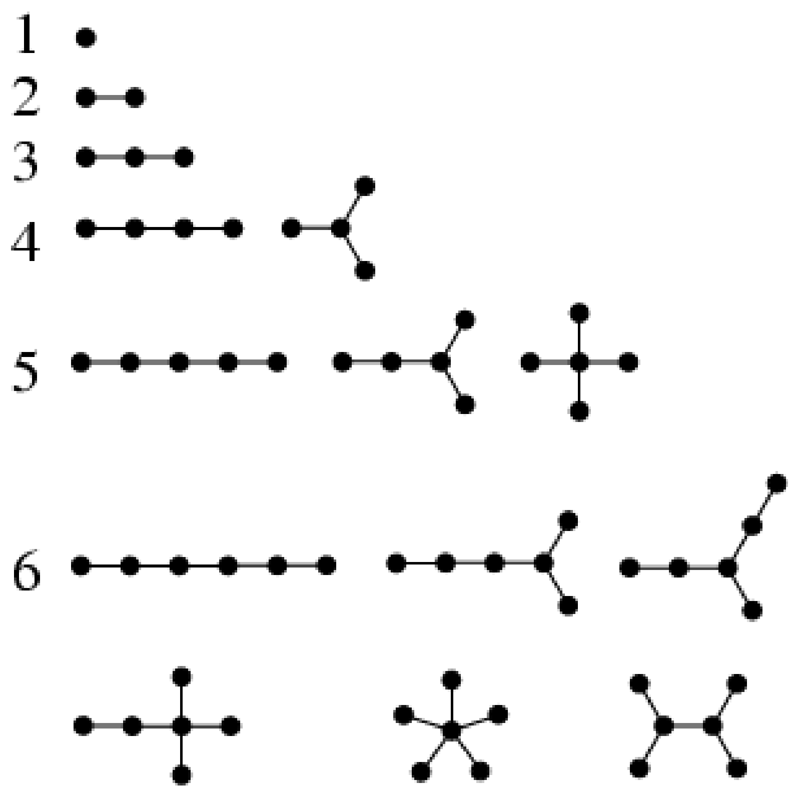
K_5



K_8

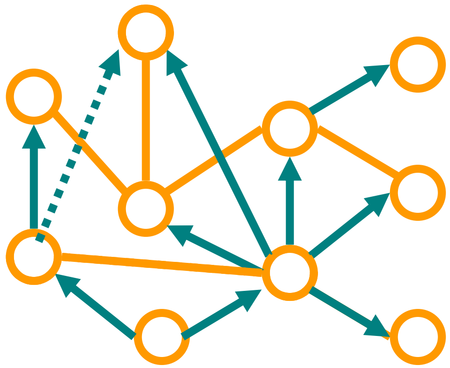
Trees

- Trees are undirected graphs that contain no cycles (loops)



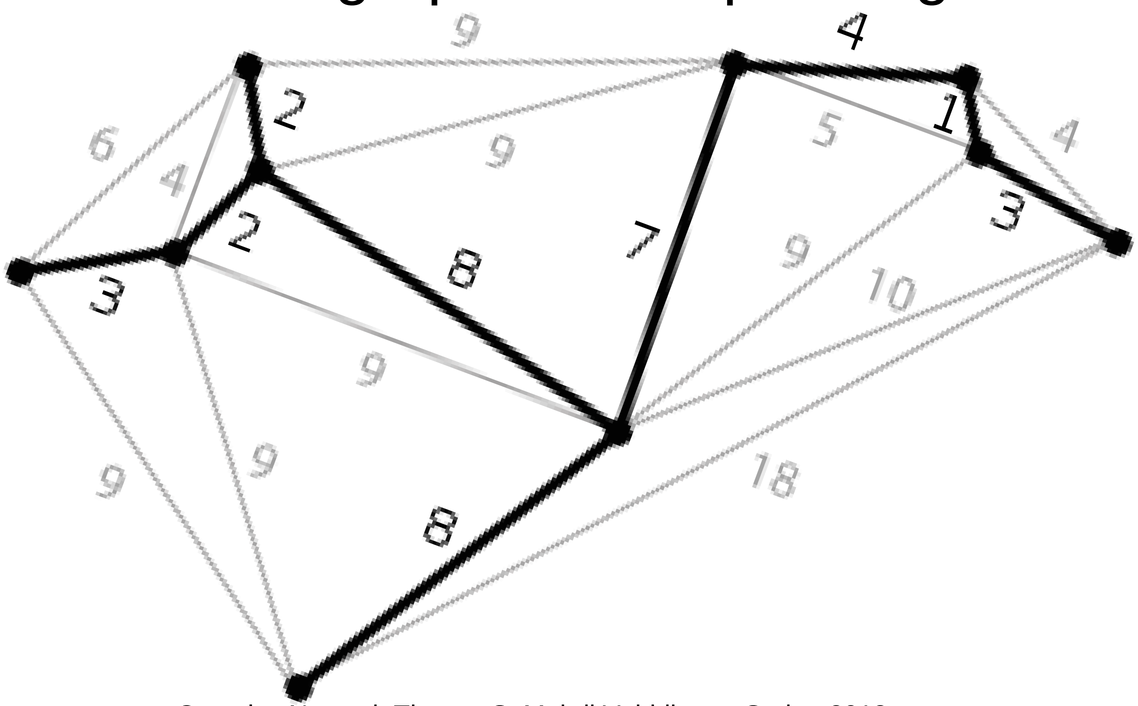
examples of trees

- In nature
  - trees
  - river networks
  - arteries (or veins, but not both)
- Man made
  - sewer system
- Computer science
  - binary search trees
  - decision trees (AI)
- Network analysis
  - minimum spanning trees
    - from one node – how to reach all other nodes most quickly
    - may not be unique, because shortest paths are not always unique
    - depends on weight of edges



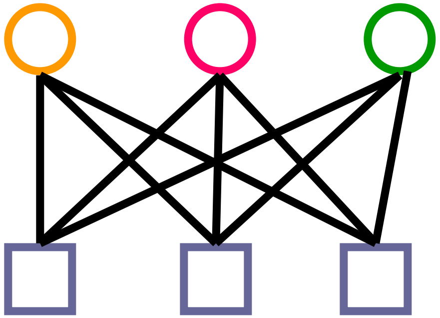
Spanning tree of a graph

- If G(V,E) is a graph and T(V,F) is a subgraph of G and is a tree, then T is a spanning tree of G. That is, T is a tree that includes every vertex of G and has only edges to be found in G. Using a procedure (remove edges from cycles until only a tree remains), we can easily prove that every connected graph has a spanning tree.

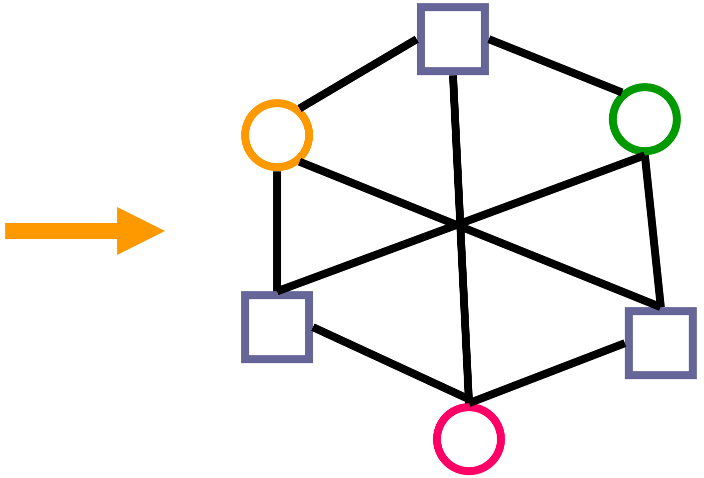


Bi-cliques (cliques in bipartite graphs)

- $K_{m,n}$ is the complete bipartite graph with **m** and **n** vertices of the two different types
- $K_{3,3}$ maps to the utility graph
  - Is there a way to connect three utilities, e.g. gas, water, electricity to three houses without having any of the pipes cross?



$K_{3,3}$



Utility graph

Eigenvalues and Eigenvectors

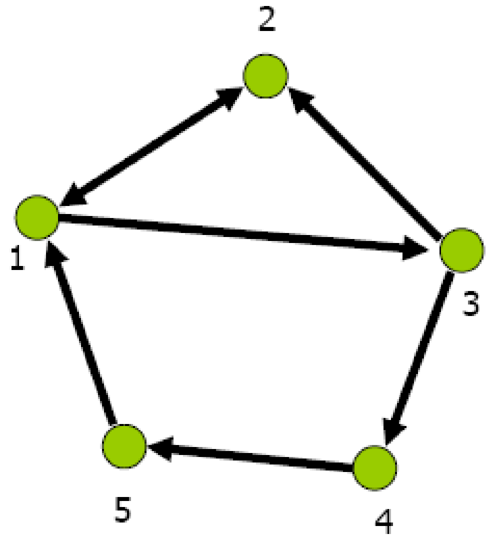
- The value λ is an eigenvalue of matrix A if there exists a non-zero vector x, such that $Ax=\lambda x$. Vector x is an eigenvector of matrix A
  - The largest eigenvalue is called the principal eigenvalue
  - The corresponding eigenvector is the principal eigenvector
  - Corresponds to the direction of maximum change
  - $Ax=\lambda x \rightarrow Ax - \lambda x = 0 \rightarrow (A-\lambda I)x=0$
  - Eig function in MATALB

Random Walks

- Start from a node, and follow links uniformly at random.
- Stationary distribution: The fraction of times that you visit node i, as the number of steps of the random walk approaches infinity
  - if the graph is strongly connected, the stationary distribution converges to a unique vector.
  - stationary distribution: principal left eigenvector of the normalized adjacency matrix
  - x = xP
  - for undirected graphs, the degree distribution

Transition matrix P

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$



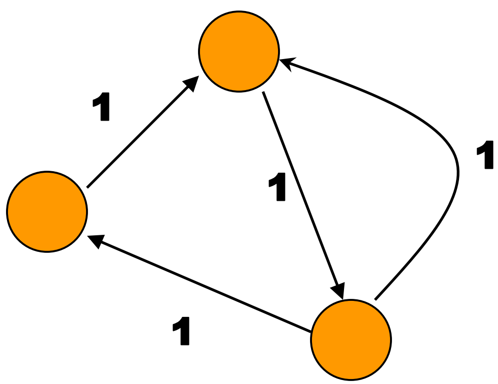
Random walks (Example)

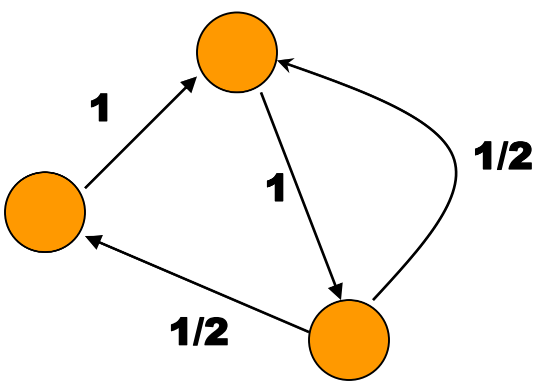
0	1	0
0	0	1
1	1	0

Adjacency matrix A

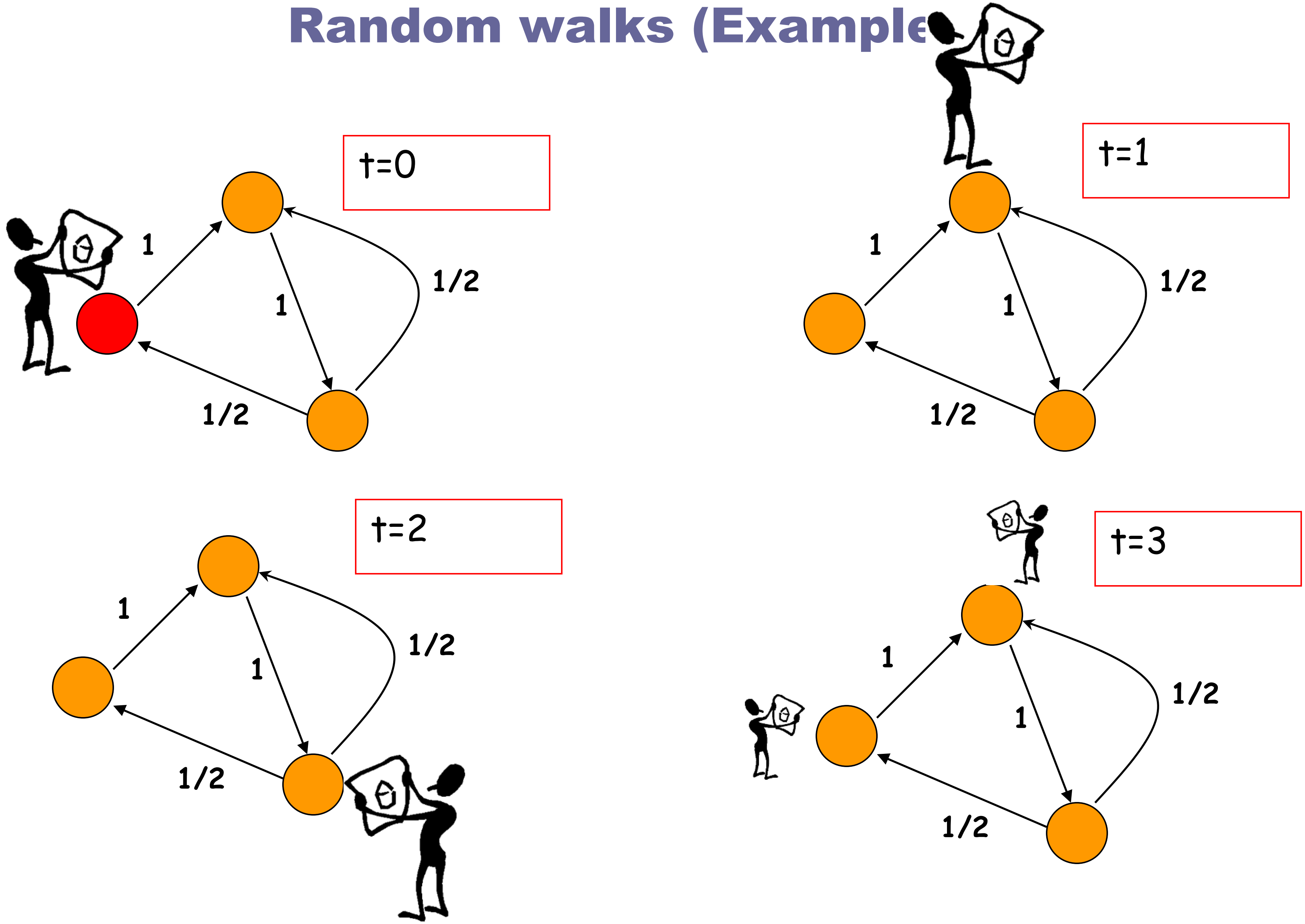
0	1	0
0	0	1
1/2	1/2	0

Transition matrix P





Random walks (Example



Probability Distributions

- $x_t(i)$ = probability that the surfer is at node *i* at time *t*
- $x_{t+1}(i) = \sum_j$(Probability of being at node j)*Pr(j->i) $=\sum_j x_t(j)*P(j,i)$
- $x_{t+1} = x_t P = x_{t-1}*P*P= x_{t-2}*P*P*P = \ldots =x_0 P^t$

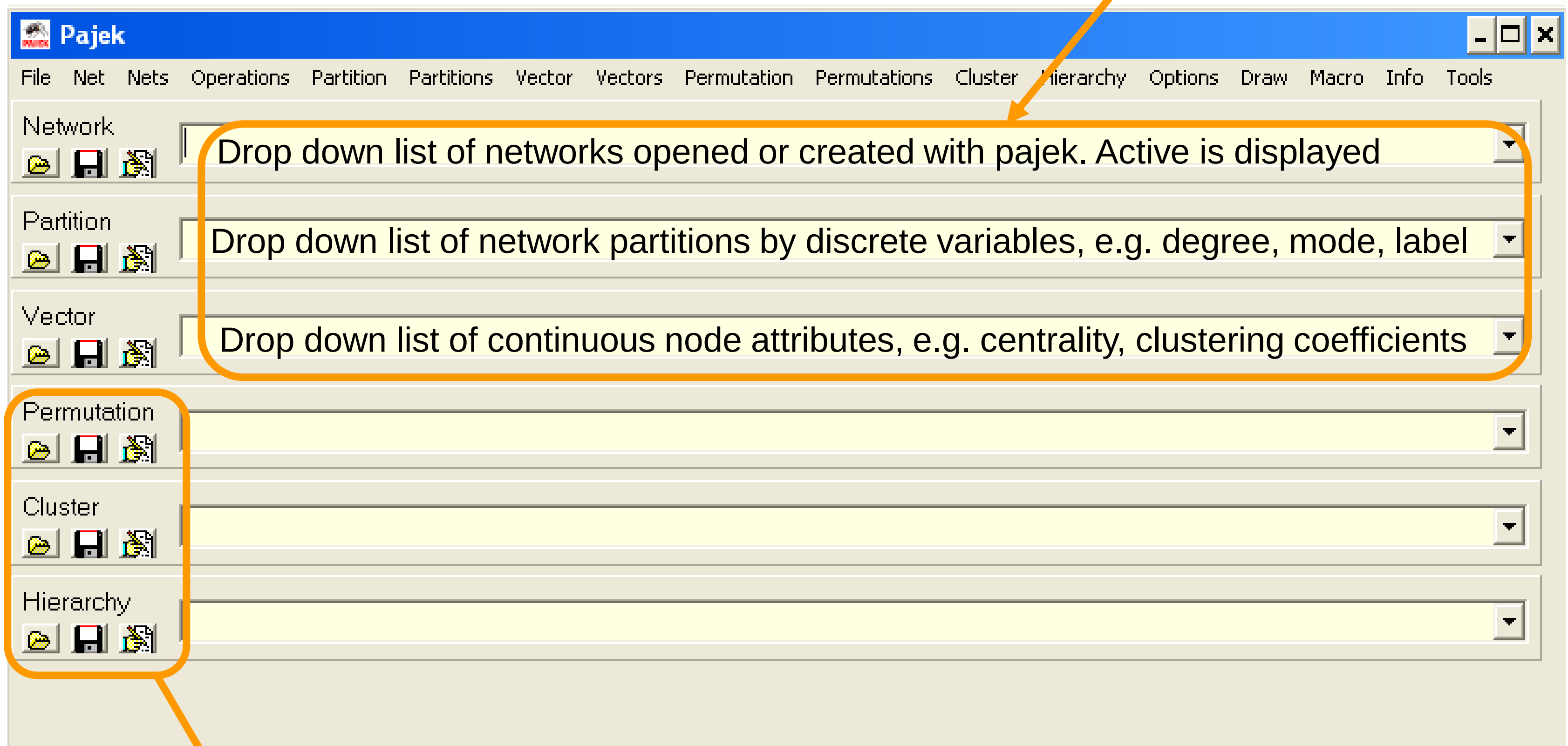
- What happens when the surfer keeps walking for a long time?
- Stationary Distribution
  - When the surfer keeps walking for a long time
  - When the distribution does not change anymore
    - i.e. $x_{T+1} = x_T$
  - For “well-behaved” graphs this does not depend on the start distribution!!

Using Pajek for exploratory social network analysis

- Pajek – (pronounced in Slovenian as Pah-yek) means 'spider'
- website: vlado.fmf.uni-lj.si/pub/networks/**pajek**/
  - download application (free)
  - tutorials
  - lectures
  - data sets
- Windows only (works on Linux via Wine)
- can be installed via NAL in the student lab (DIAD)
- helpful book: 'Exploratory Social Network Analysis with Pajek' by Wouter de Nooy, Andrej Mrvar and Vladimir Batagelj
  - first 2 chapters are required reading and on cTools
- Pajek
  - Opening a network
  - Visualization
  - Essential measurements

Pajek interface

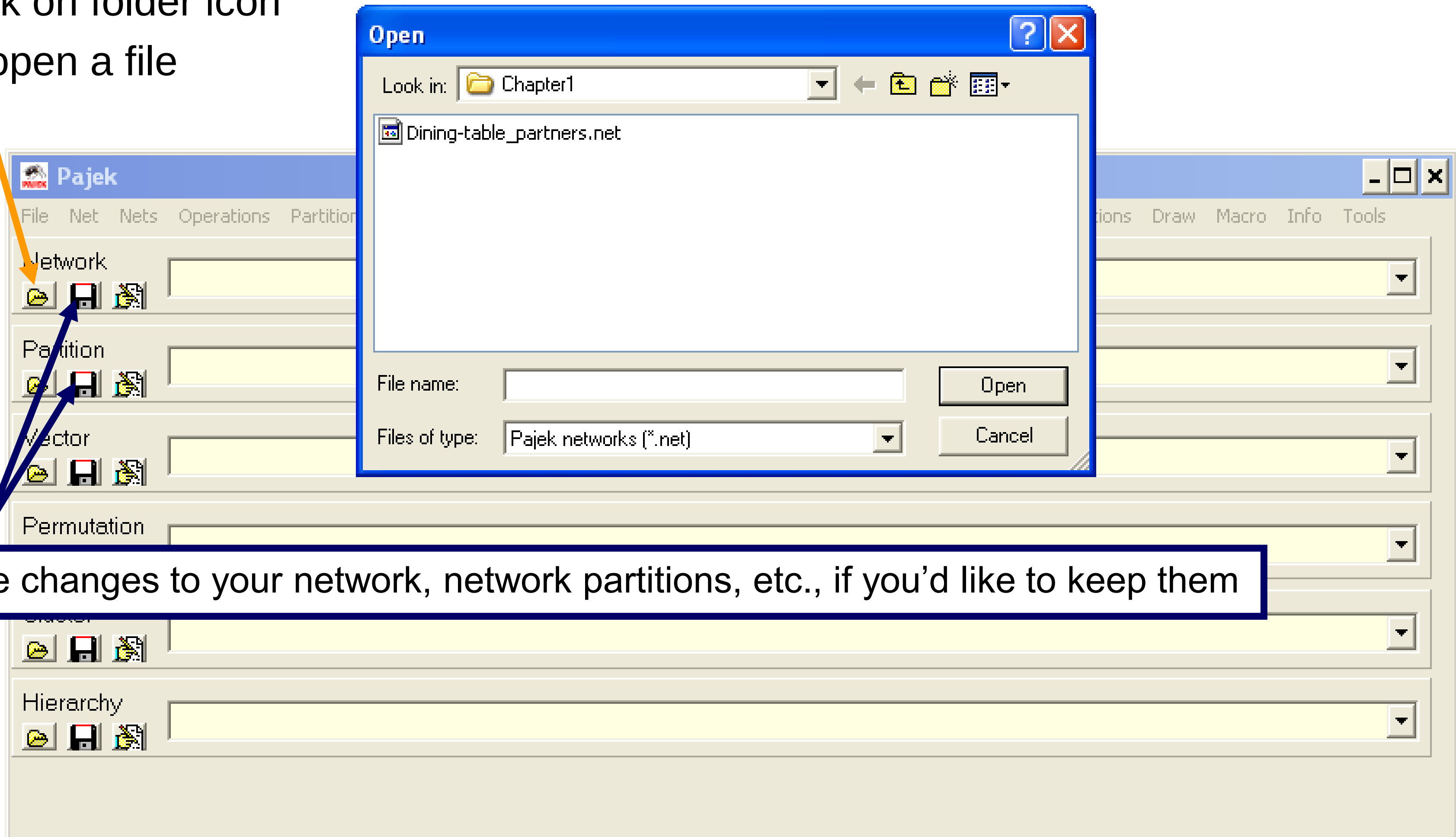
things we'll use right away



things we'll use later for clustering

opening a network file

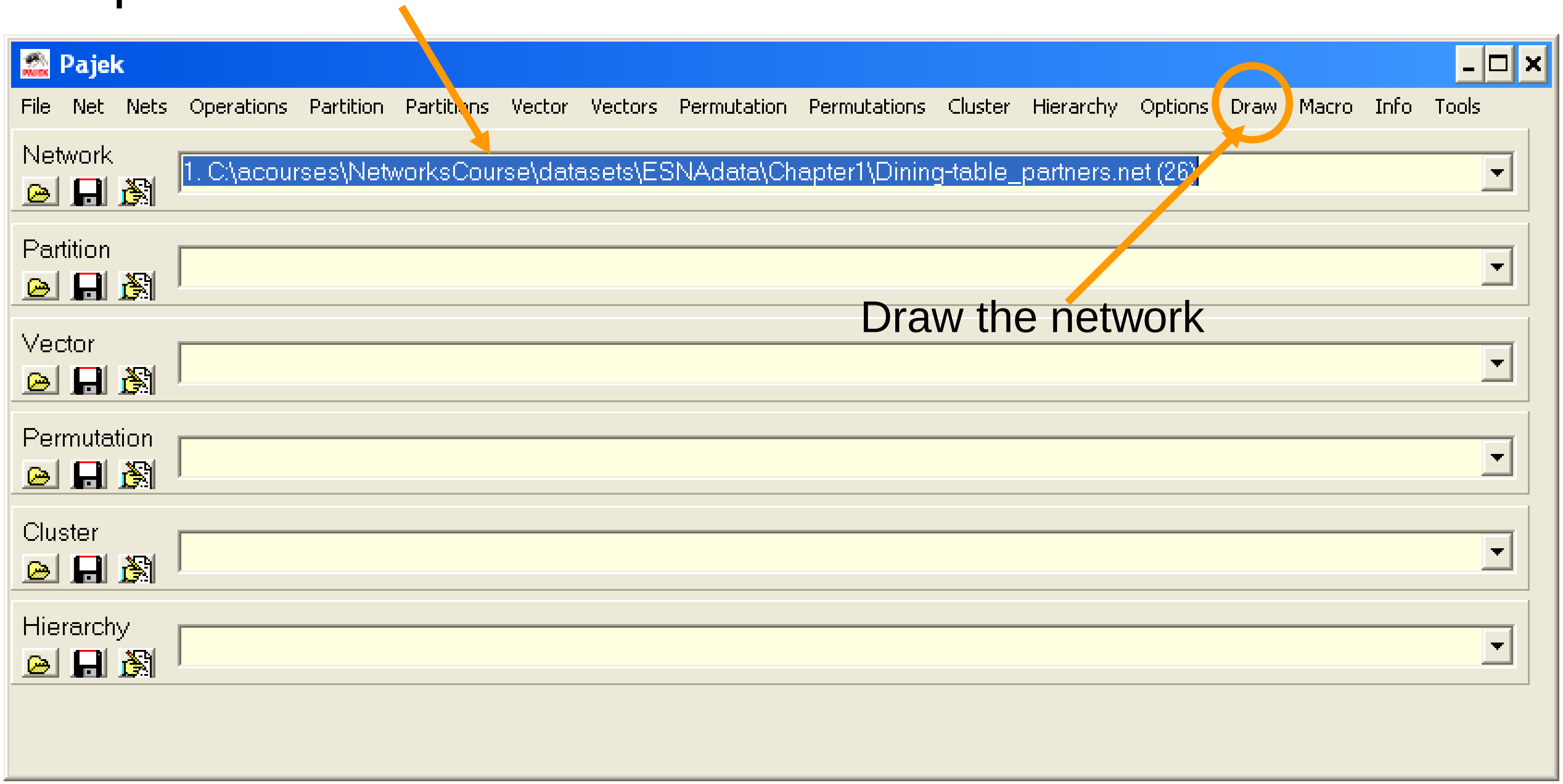
click on folder icon
to open a file



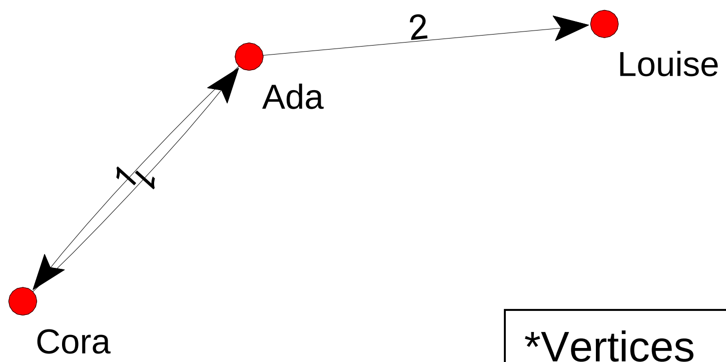
Save changes to your network, network partitions, etc., if you'd like to keep them

Working with network files in Pajek

- The active network, partition, etc is shown on top of the drop down list



Pajek data format



number of vertices

vertex x,y,z coordinates (optional)

```
*Vertices     26
    1 "Ada"          0.1646   0.2144   0.5000
    2 "Cora"         0.0481   0.3869   0.5000
    3 "Louise"       0.3472   0.1913   0.5000
     ..

*Arcs
    1 3 2 c Black
     ..
*Edges
    1 2 1 c Black
     ..
```

directed edges

from Ada(1) to Louise(3) as choice “2” and color Black

undirected edges

between Ada(1) to Cora(2) as choice “1” and color Black

Readings

- Easley, David, and Jon Kleinberg. **Networks, crowds, and markets: Reasoning about a highly connected world**. Cambridge University Press, 2010. (Ch.1-2)

- Newman, Mark. **Networks: an introduction**. Oxford University Press, 2010. (Ch. 6)

- L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. **Characterization of complex networks: A survey of measurements**. Advances in Physics, 56(1):167 – 242, 2007.