Introduction to Network Science, CIS 689-011

Who: Prof. Ilya Safro, Zoom office hours on Thursdays at 11am-12pm
When: T/Th 9:30am
Where: Zoom (fully online)
TA: TBD

Course Structure

What	How many	Time	Points
Homework	≤10*	1 week	20
Paper reading	1-3 per week	1 week	
Oral presentation	1	>2 weeks	20
Final project	1	>4 weeks	50
Quizzes	Every class*	5-7 min	10
Total			100

Points	Grade
≥ 90	A
≥ 80	B
≥ 60	C
≥ 0	F

Bonuses

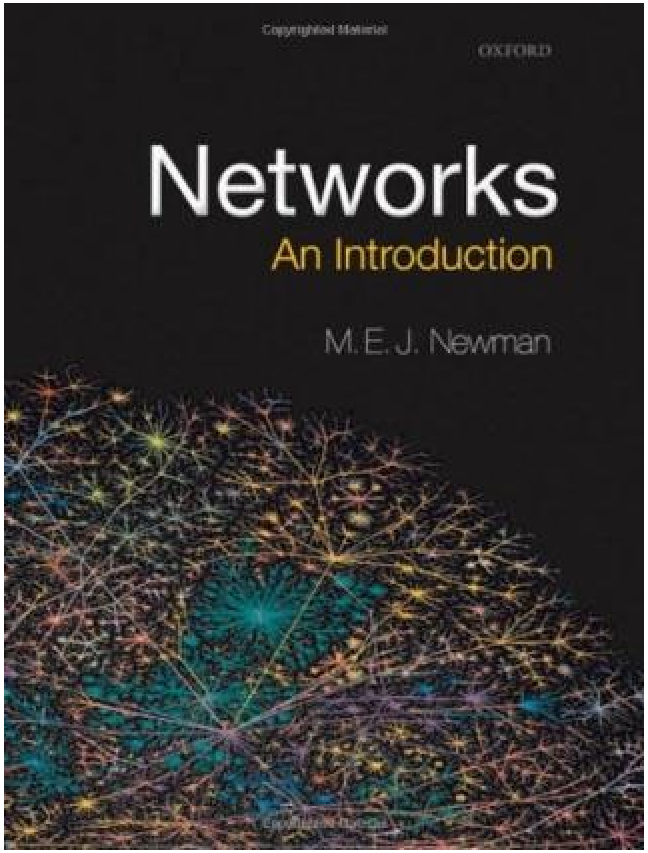
Work in class, extra work in homework, etc. - up to 10 points. We do not want to miss the next Turing, Fields and Nobel laureates, so any submitted conference/journal paper written during and as a result of this course - 100 points, and new interesting ideas - up to 100 points (both are based on instructor's subjective judgment).

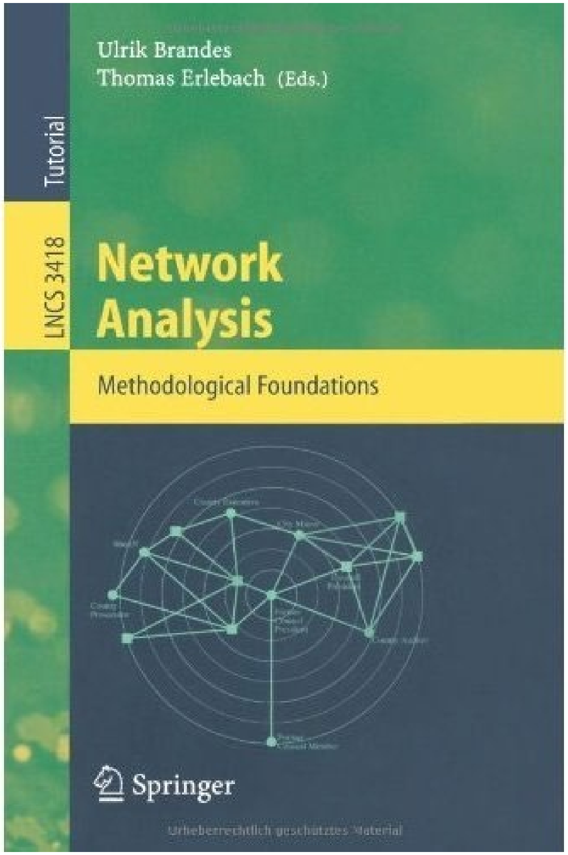
Final project

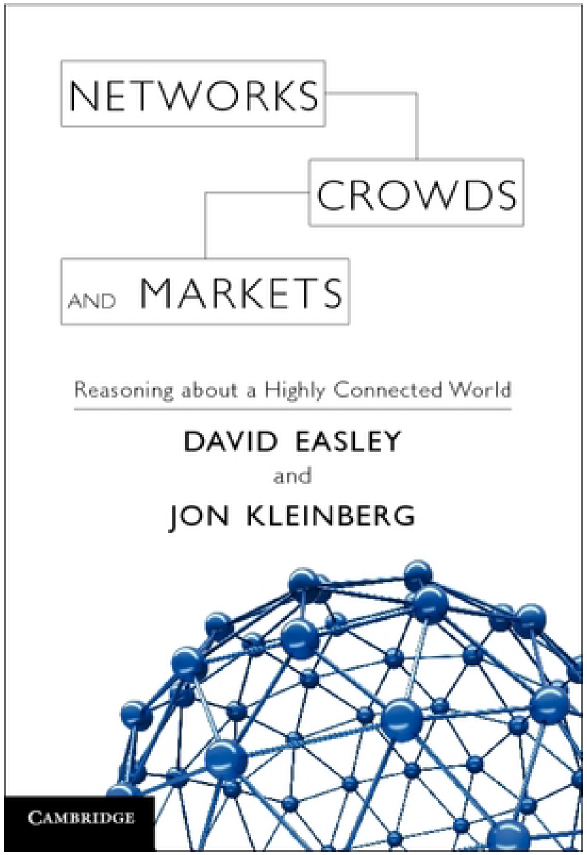
- Can be done in groups of 1-2 students
- Big group → big project
- 100% of your grade: research paper (aka final report) + source code + computational results + presentation to the instructor or TA
- Submit final project proposals by 5/15
- No small networks (unless your work is theoretical)
- All algorithms should be fast (unless you can justify the slowness)

Zero tolerance for plagiarism in homework and final project

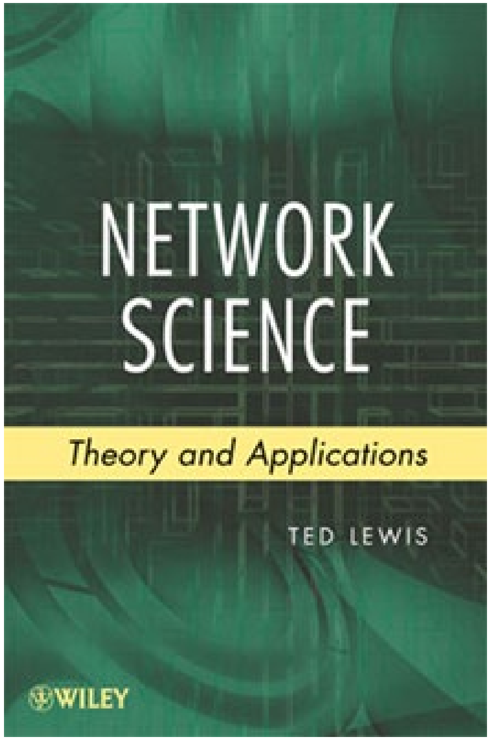
Recommended books







Recommended books

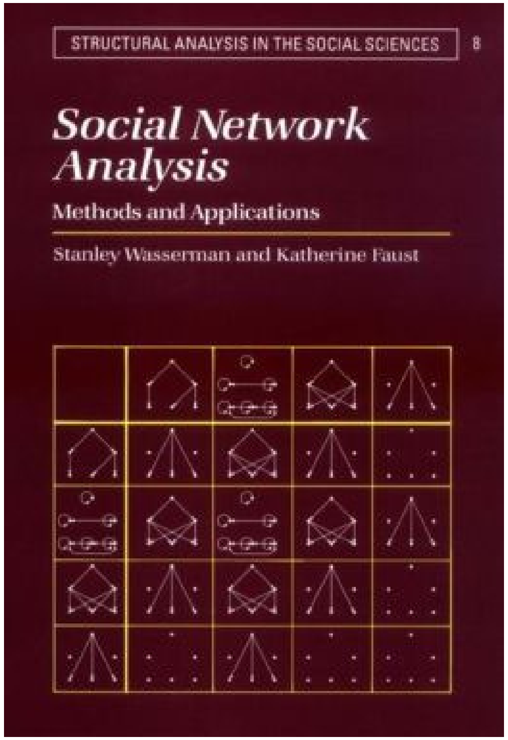


NETWORK SCIENCE

Theory and Applications

TED LEWIS

WILEY

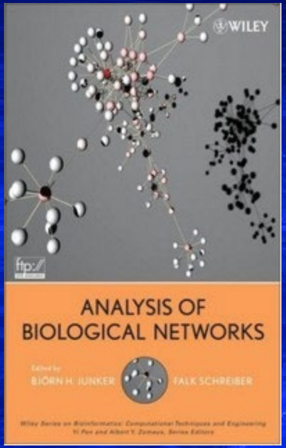


STRUCTURAL ANALYSIS IN THE SOCIAL SCIENCES 8

Social Network Analysis

Methods and Applications

Stanley Wasserman and Katherine Faust



WILEY

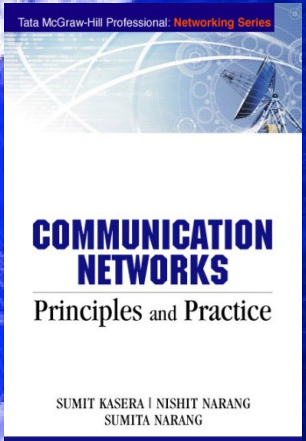
ANALYSIS OF BIOLOGICAL NETWORKS

Edited by

BJÖRN H. JUNKER

FALK SCHREIBER

Wiley Series on Bioinformatics: Computational Techniques and Engineering
Yi Pan and Albert Y. Zomaya, Series Editors

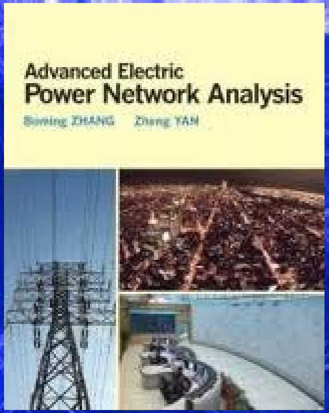


Tata McGraw-Hill Professional: Networking Series

COMMUNICATION NETWORKS

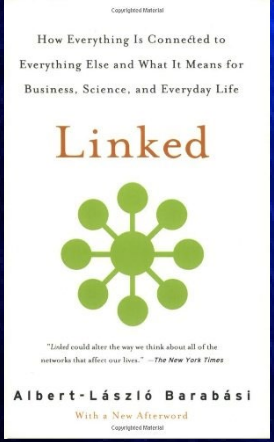
Principles and Practice

SUMIT KASERA | NISHIT NARANG
SUMITA NARANG



Advanced Electric
Power Network Analysis

Boming ZHANG Zheng YAN



Copyrighted Material

How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life

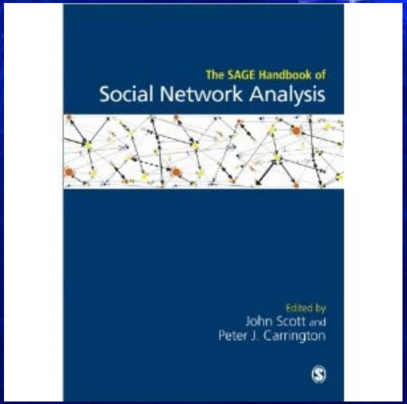
Linked

"*Linked* could alter the way we think about all of the networks that affect our lives." —*The New York Times*

Albert-László Barabási

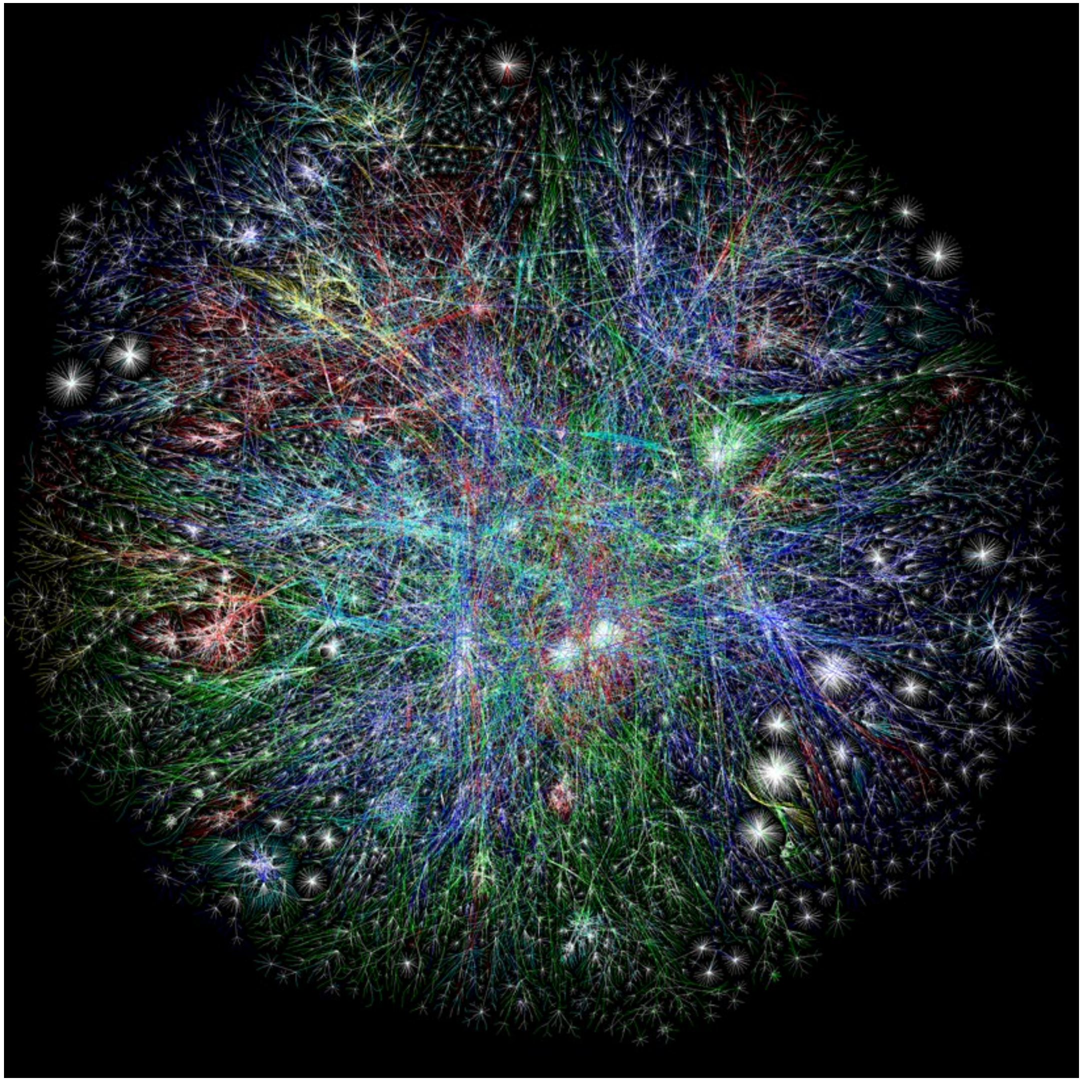
With a New Afterword

Copyrighted Material



The SAGE Handbook of
Social Network Analysis

Edited by
John Scott and
Peter J. Carrington



Internet (from opta.org), |V|=5M, |E|=50M,
Colors are geographic regions
Pages/Hyperlinks is a different network!

- **Packet switched data network**
- IP – communications protocol
- Packets are small
- Packets can disappear
- TCP – transmission control, error checking

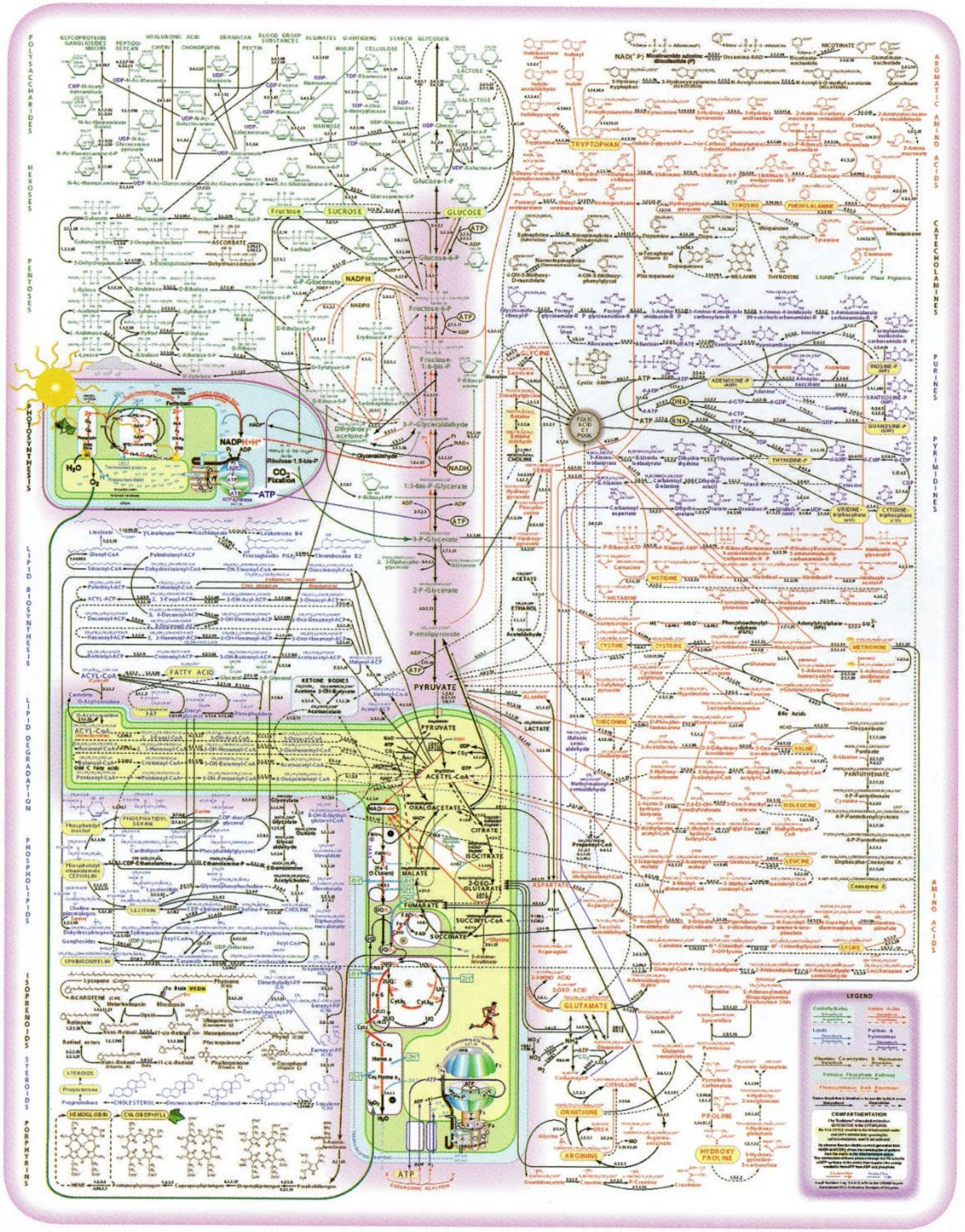
VoIP: will be merged in future?

Alternative to PSN is a **circuit switched network**. Example: telephone systems.

- Separate circuit for each call
- No packets

Research problems?

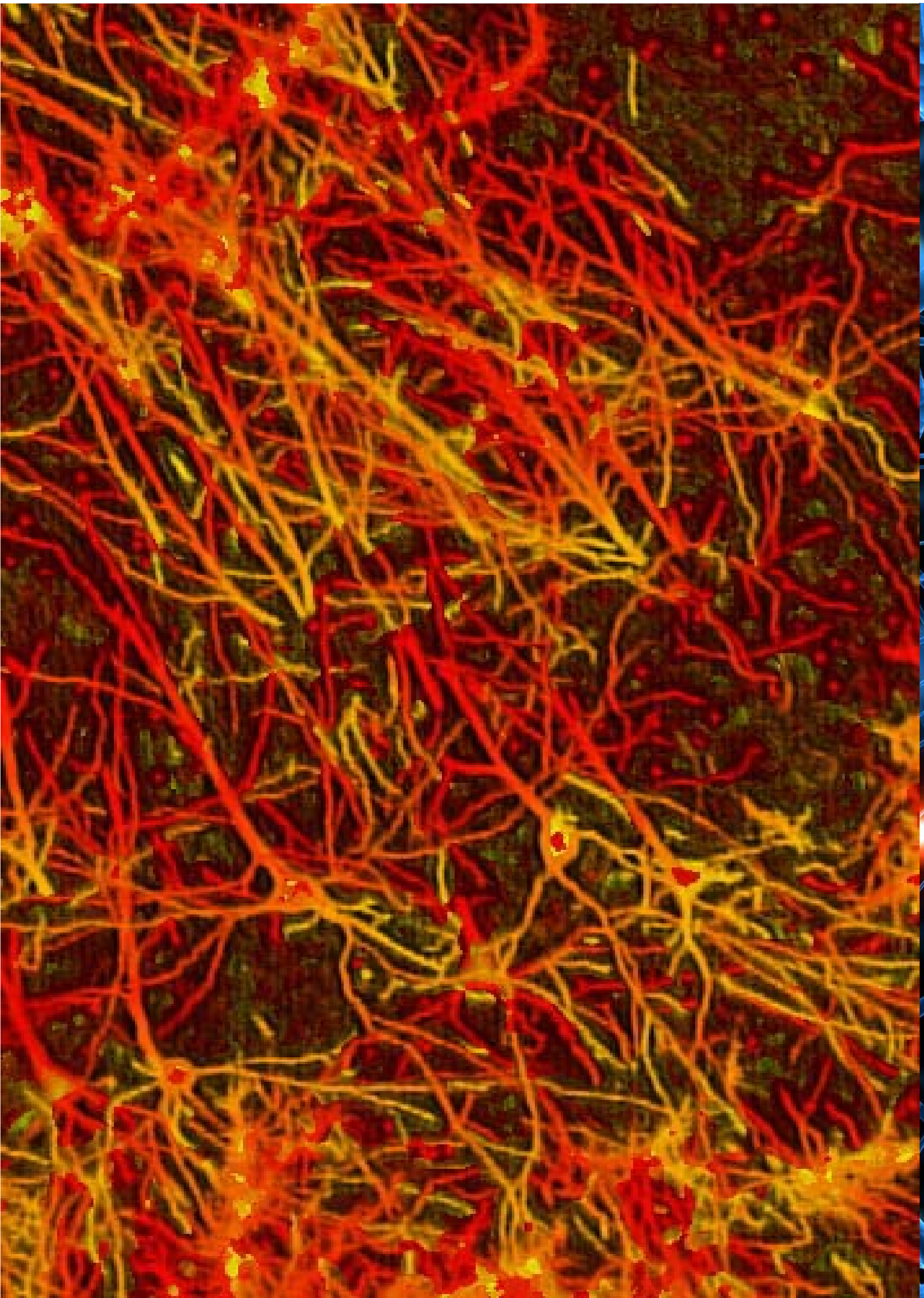
- Network congestion
- Robustness (wrt disasters, etc.)
- Dynamics

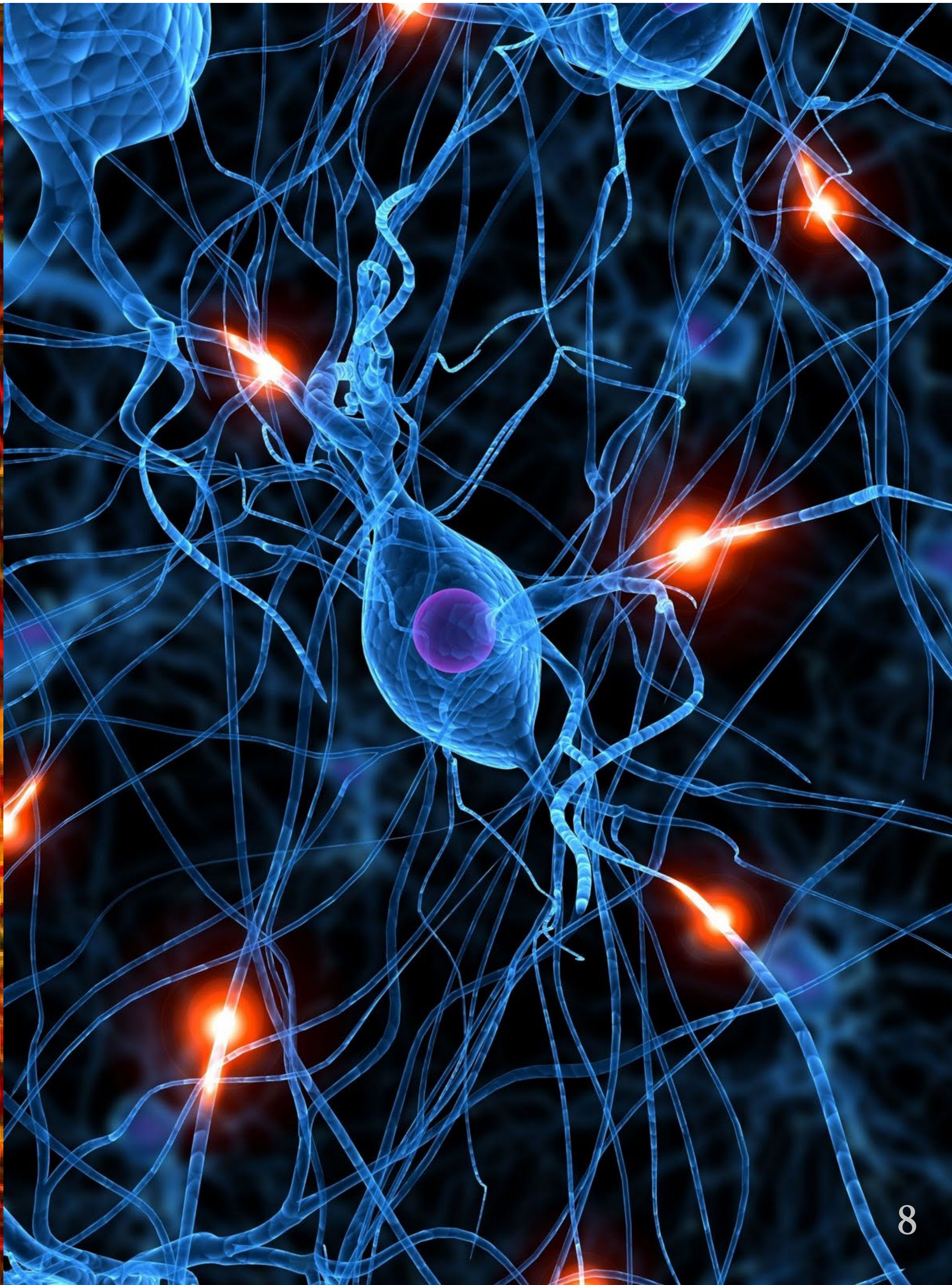


Network formed by the major metabolic pathways (Newman “Networks: An Introduction”)

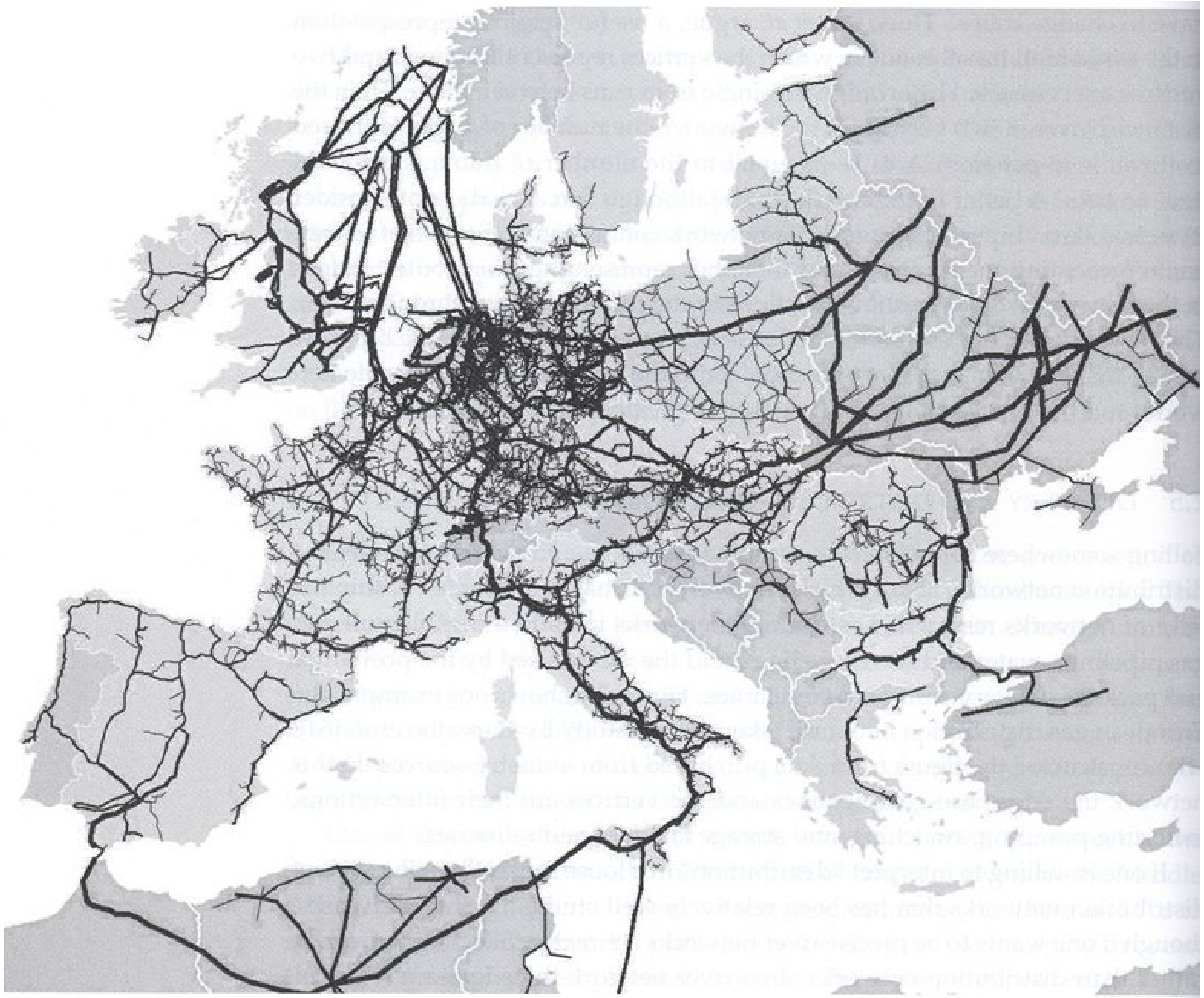
Brain Networks

from networksandservers.blogspot.com

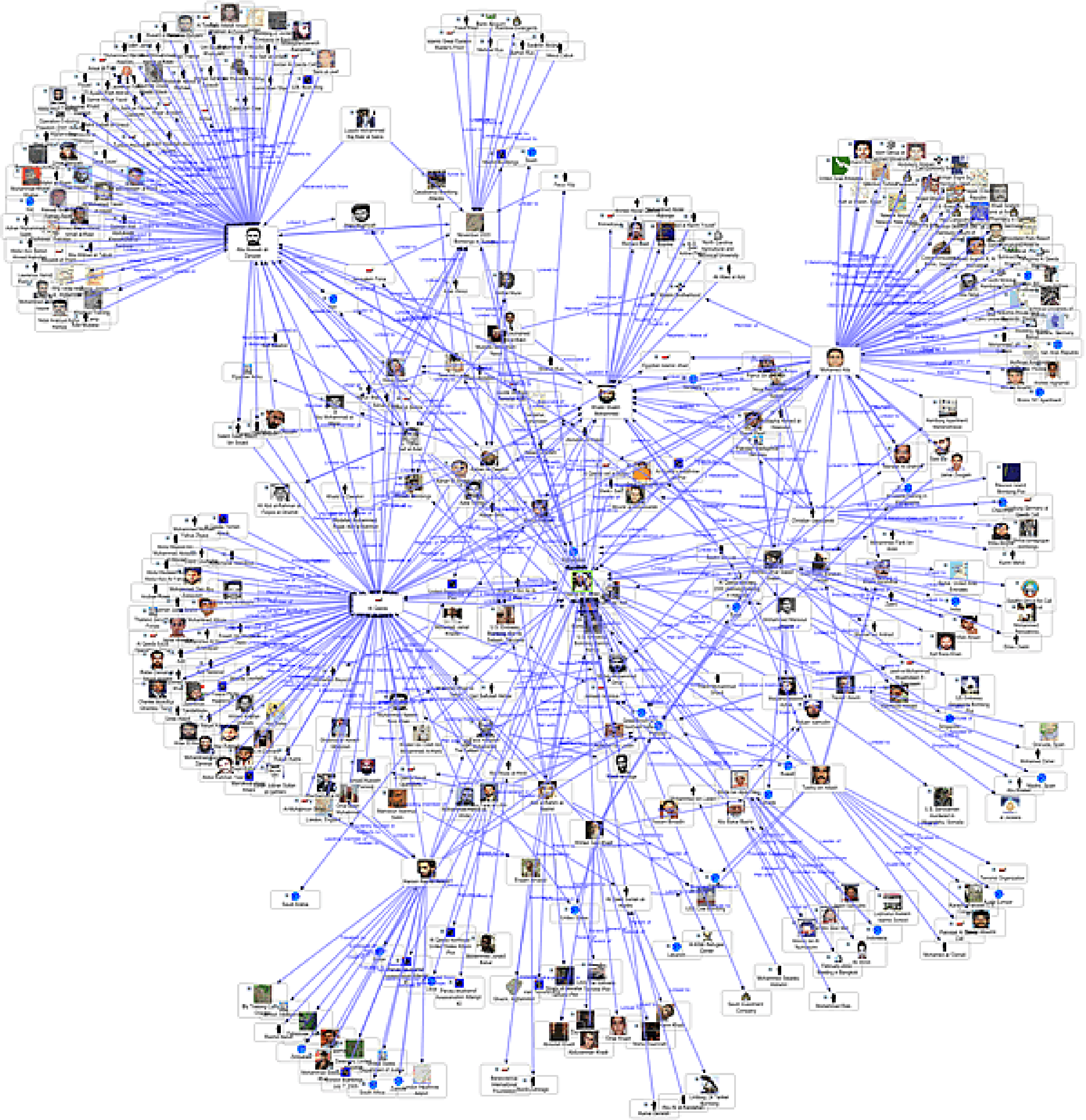




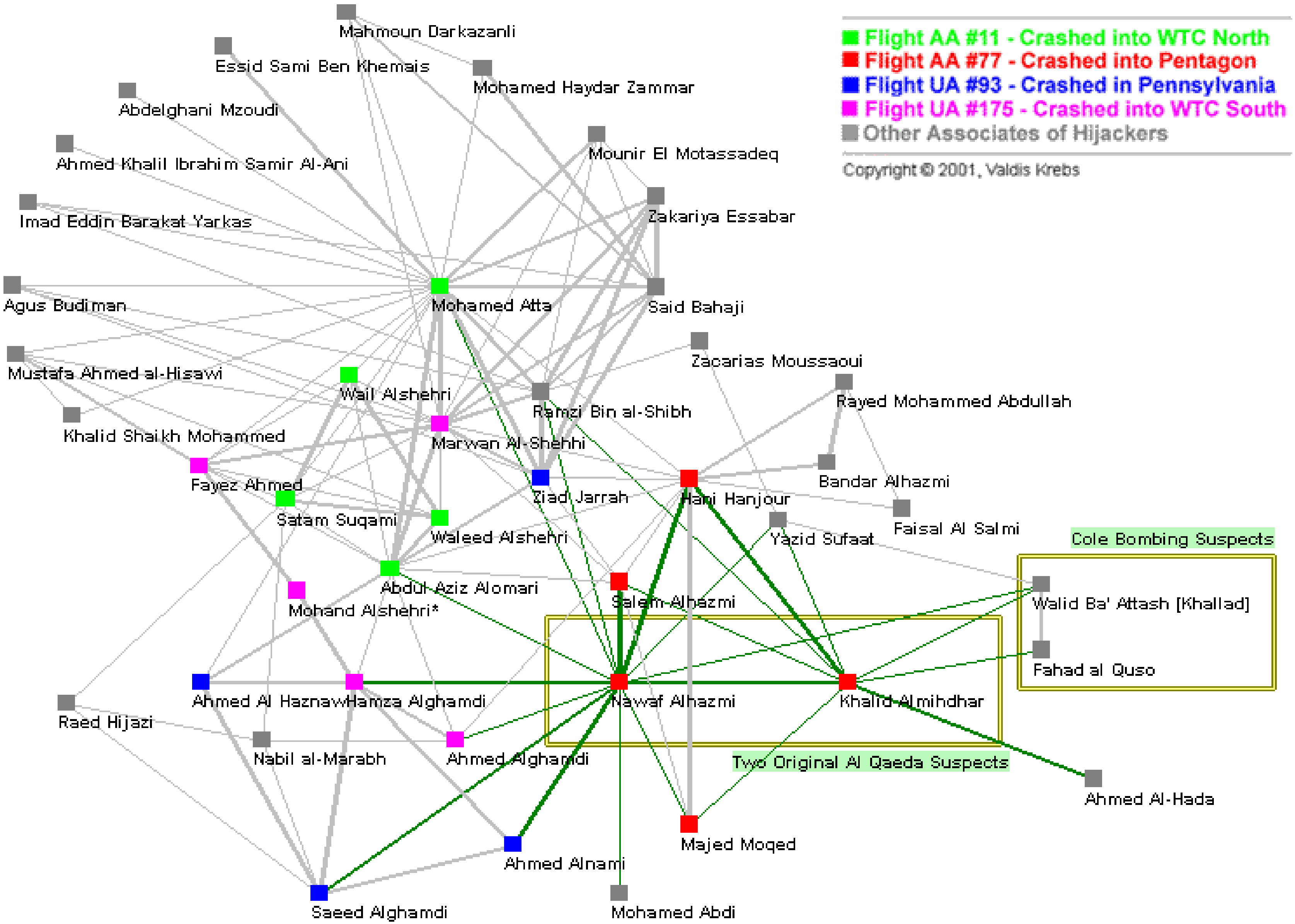
Delivery and Distribution Networks



Natural gas major pipelines in Europe (Newman “Networks: An Introduction”)



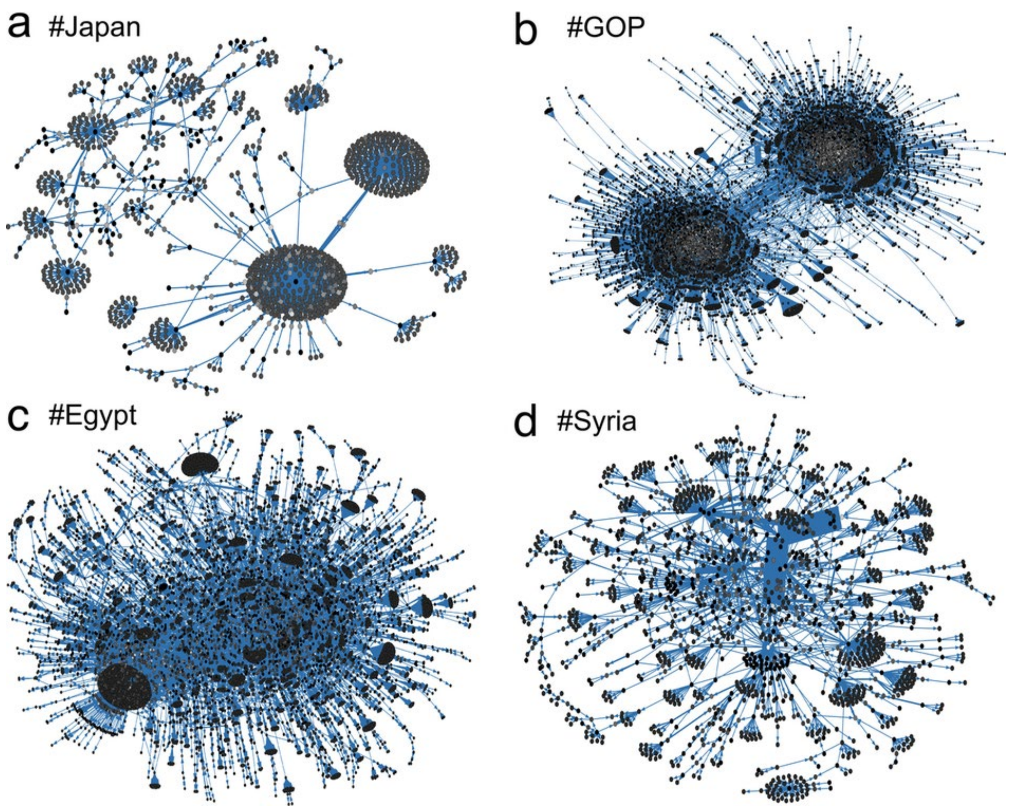
Network of 9/11 contacts (Krebs)



Copyright © 2001, Valdis Krebs

Another network of 9/11 contacts (Krebs)

- Nodes = Twitter users
- Directed edges = retweeted posts that carry the meme.



Networks of retweets (Weng, Flammini, Vespignani, Menczer)

The brightness of a node indicates the activity (number of retweets) of a user, and the weight of an edge reflects the number of retweets between two users.

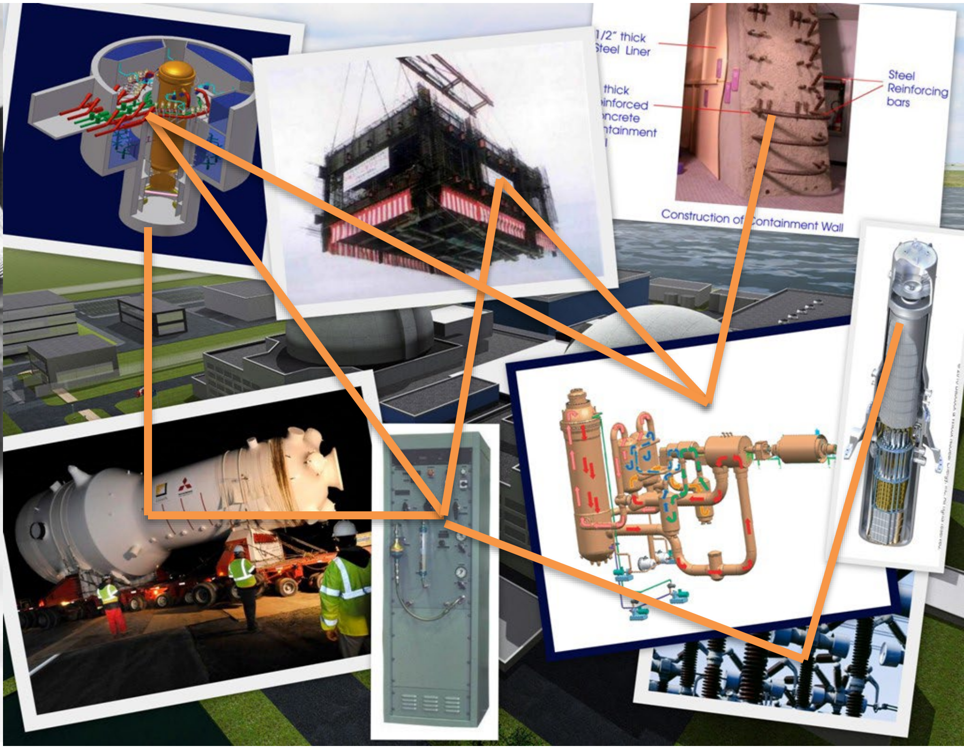
(a) The **#Japan** meme shows how news about the March 2011 earthquake propagated.

(b) The **#GOP** tag stands for the US Republican Party and as many political memes, displays a strong polarization between people with opposing views.

(c) and (d) Memes related to the “Arab Spring” and in particular the 2011 uprisings in (c) **#Egypt** and (d) **#Syria** display characteristic hub users and strong connections, respectively.

Actors, Reactors and More

IMDB, Network of movie actors



Transportation and logistics network of nuclear power plant

Definition

Network is an object composed of elements and interactions between them.

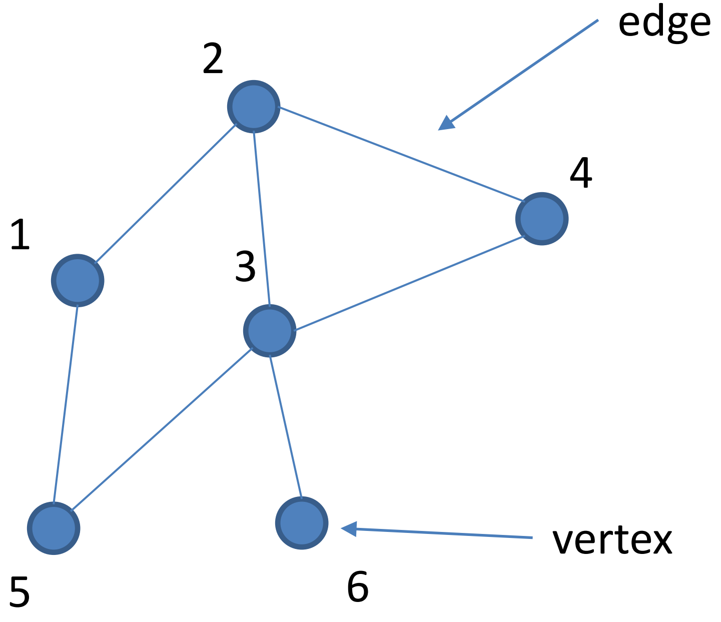
Examples

- Internet (servers/routers/computers and fiber/optic/wireless connections)
- Epidemiology (people/places and contacts)

Graphs

A graph *G*=(*V*, *E*) is an abstract mathematical object formed by

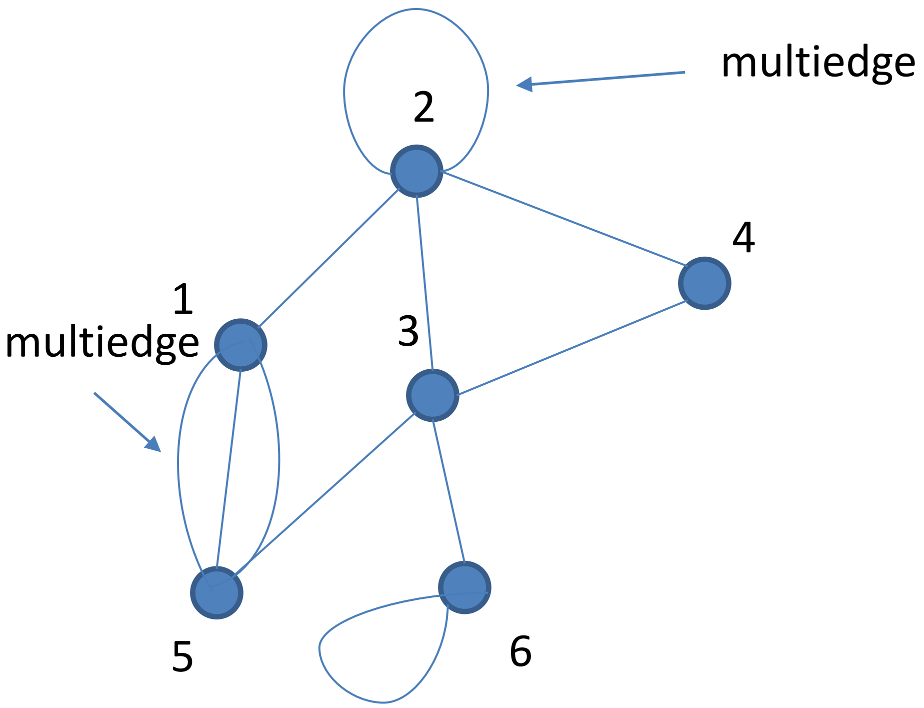
- set of vertices or nodes *V*
- set of edges *E* that connect pairs of vertices
- |V| = n, |E| = m
- l_E , l_V are labeling functions for edges and nodes, respectively



edge

vertex

Simple graph with no multiedges and self-edges

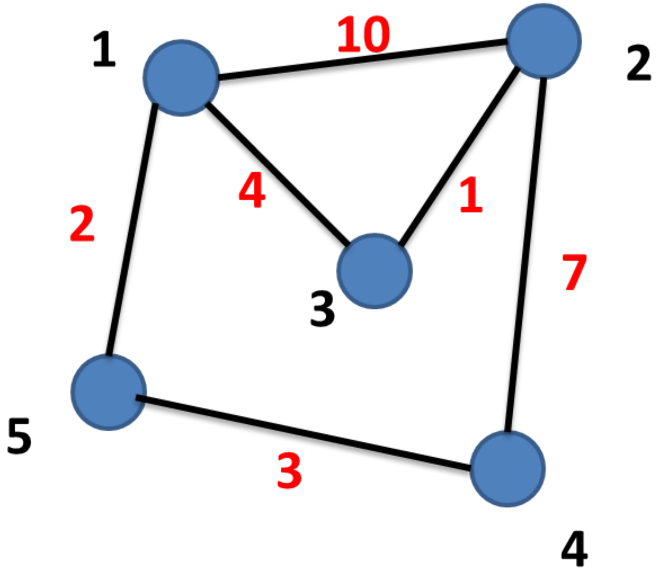


multiedge

multiedge

Graph with multiedges and self-edges

Weighted graphs and matrix representation



Adjacency Matrix *A*

	1	2	3	4	5
1		10	4		2
2	10		1	7	
3	4	1			
4		7			3
5	2			3	

Laplacian *L*

	1	2	3	4	5
1	16	-10	-4		-2
2	-10	18	-1	-7	
3	-4	-1	5		
4		-7		10	-3
5	-2			-3	5

real, symmetric matrices

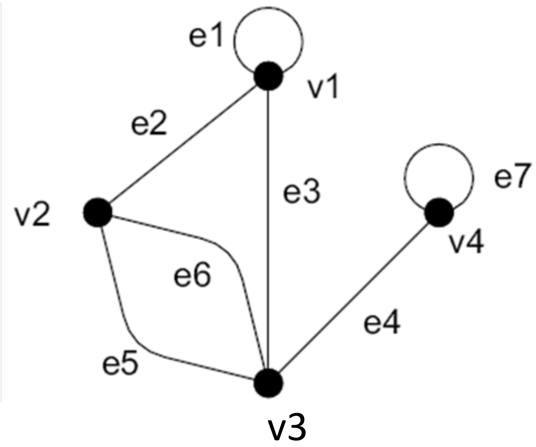
eigenvalues are real

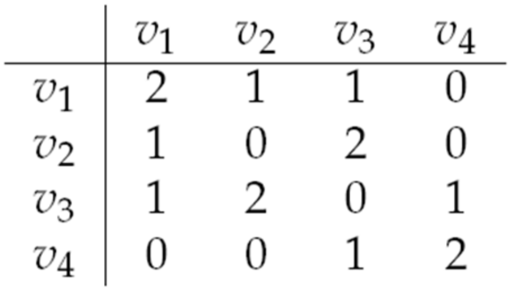
$$L_{ij} = -\omega_{ij}$$

$$L_{ii} = \sum_{ij\in E} \omega_{ij}$$

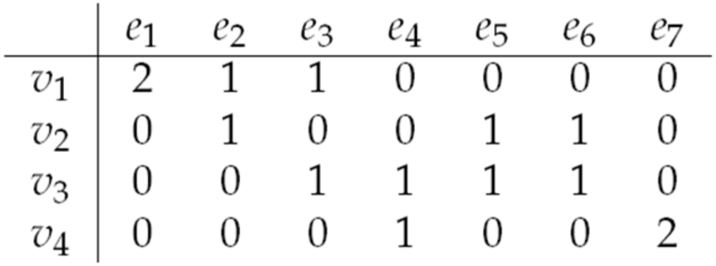
normalized Laplacian $\mathcal{L} = D^{-\frac{1}{2}} \cdot L \cdot D^{-\frac{1}{2}}$

Data structure





	v_1	v_2	v_3	v_4
v_1	2	1	1	0
v_2	1	0	2	0
v_3	1	2	0	1
v_4	0	0	1	2



	e_1	e_2	e_3	e_4	e_5	e_6	e_7
v_1	2	1	1	0	0	0	0
v_2	0	1	0	0	1	1	0
v_3	0	0	1	1	1	1	0
v_4	0	0	0	1	0	0	2

Matrix formats

Large space but fast access. Space can be improved with compressed row representation. Often used in static problems, i.e., given a graph, solve a problem on it.

edge weight

Edge list format: (1,1,1), (1, 2, 1), (2, 3, 2), (3, 1, 1), (4, 3, 1), (4, 4, 1)

Often used in streaming frameworks and sublinear algorithms

Directed networks

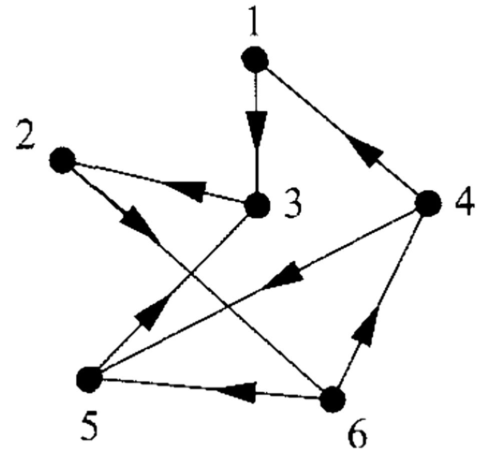


Figure 6.2: A directed network. A small directed network with arrows indicating the directions of the edges.

A *directed network* or *directed graph*, also called a *digraph* for short, is a network in which each edge has a direction, pointing *from* one vertex *to* another. Such edges are themselves called *directed edges*, and can be represented by lines with arrows on them.

A directed network is represented using non-symmetric matrix.

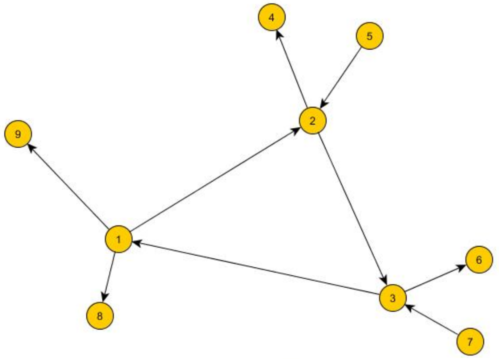
$$A_{ij} = \begin{cases} 1 & \text{if there is an edge } \textit{from } j \textit{ to } i, \\ 0 & \text{otherwise.} \end{cases}$$

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

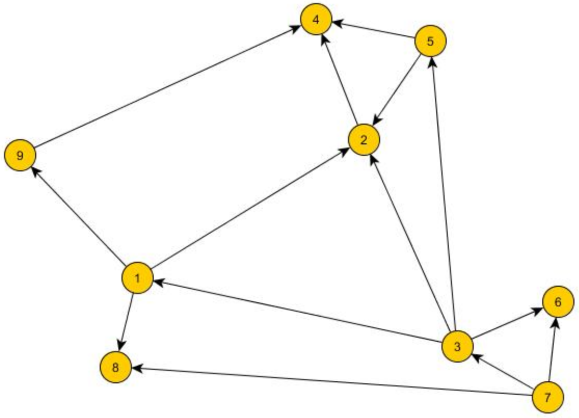
Directed networks can have multiedges and self-edges. Self-edges are represented by setting the corresponding diagonal element of the adjacency matrix to 1, not 2.

Acyclic directed networks

A *cycle* in a directed network is a closed loop of edges with the arrows on each of the edges pointing the same way around the loop.



Some directed networks however have no cycles and these are called *acyclic* networks.



We can always find a “root” node in acyclic network.

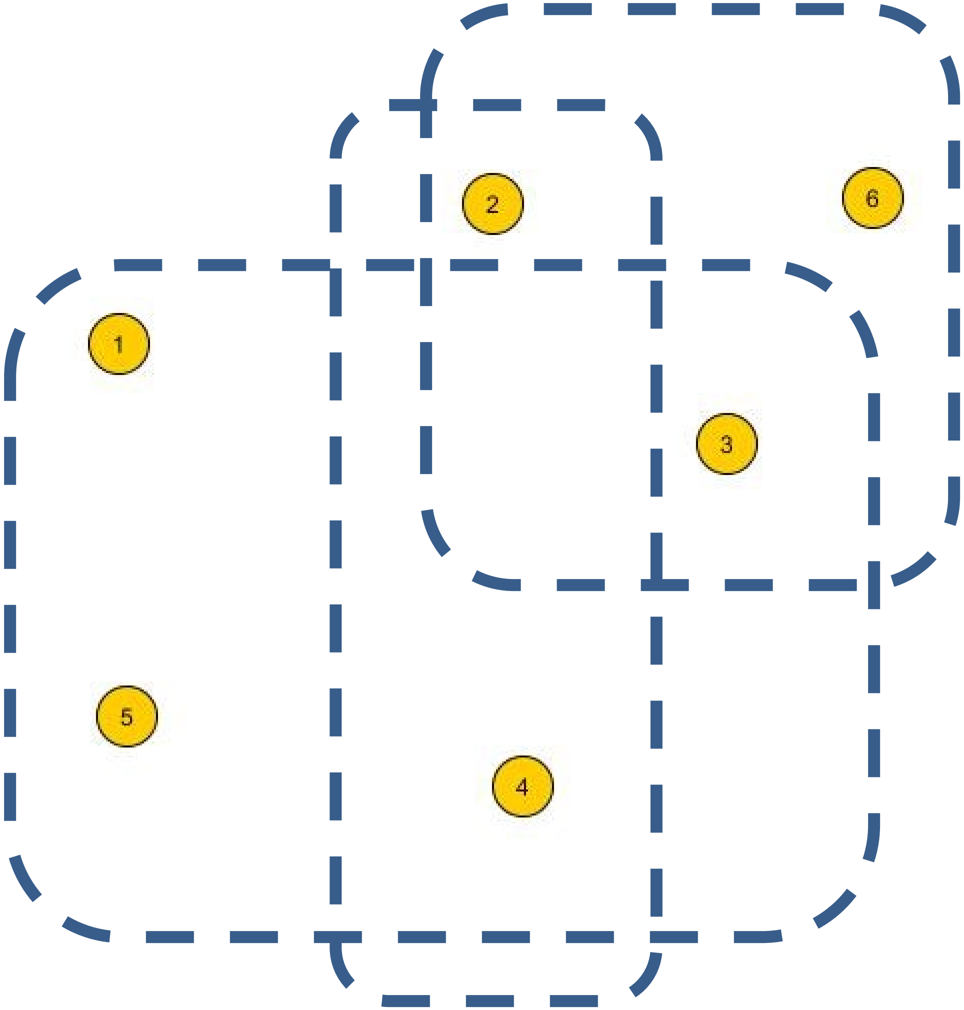
Algorithm to determine if a network is acyclic

1. Find a node with no outgoing edges
2. In no such vertex exists the network is cyclic, otherwise remove it and all its edges
3. If all vertices are removed the network is acyclic, otherwise go to step 1

There exist a labeling of an undirected network such that its adjacency matrix is strictly upper triangular. See CLR book.

Moreover, the adjacency matrix of a directed network has all 0 eigenvalues **iff** it is acyclic.

Hypergraphs



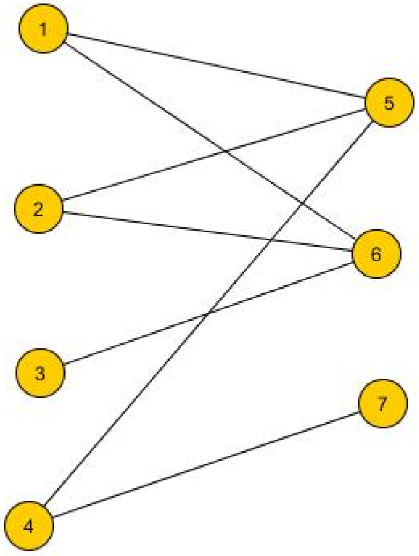
Network	Vertex	Group
Film actors	Actor	Cast of a film
Coauthorship	Author	Authors of an article
Boards of directors	Director	Board of a company
Social events	People	Participants at social event
Recommender system	People	Those who like a book, film, etc.
Keyword index	Keywords	Pages where words appear
Rail connections	Stations	Train routes
Metabolic reactions	Metabolites	Participants in a reaction

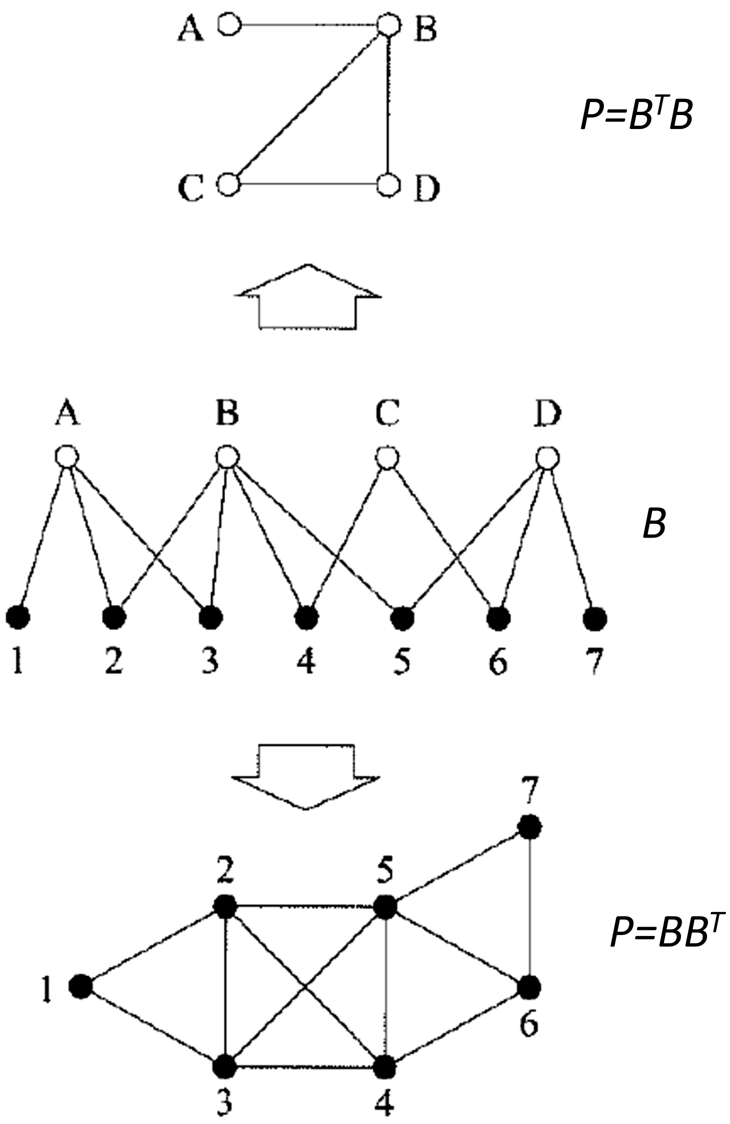
Bipartite networks

The membership of vertices in groups represented by hyperedges in a hypergraph can equally and often more conveniently be represented as a *bipartite network.*

In sociology it is also called a *two-mode network*. In such a network there are two kinds of vertices, one representing the original vertices and the other representing the groups to which they belong.

There are many more applications to bipartite graphs than just in sociology.

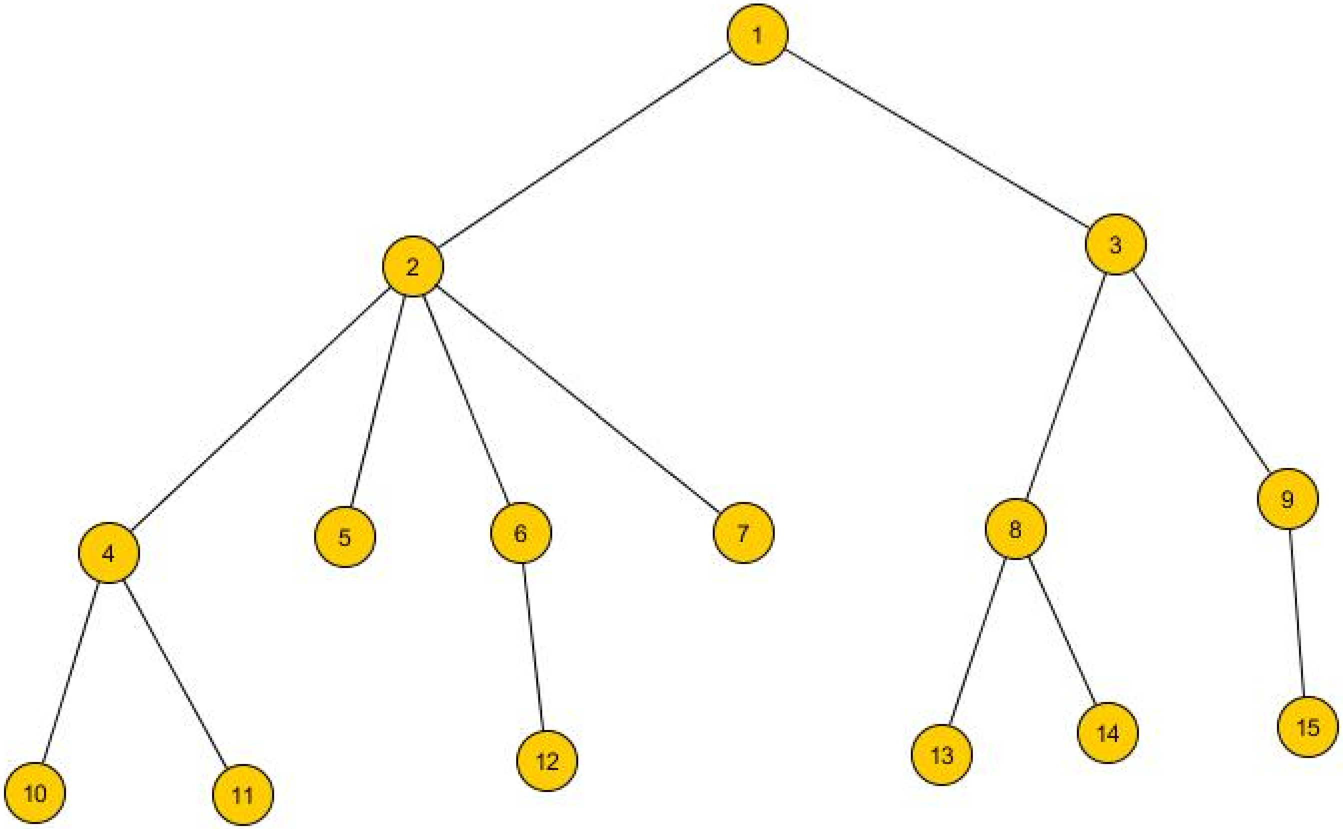




Examples of 1-mode projections of 2-mode network .

Trees and forests

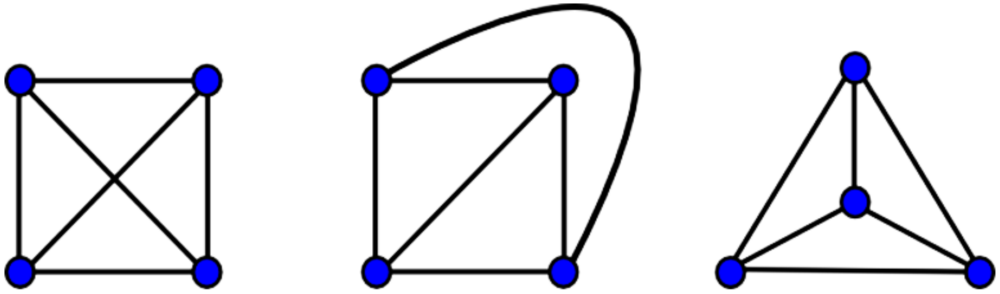
A *tree* is a connected (every vertex in the network is reachable from every other via some path through the network), undirected network that contains no closed loops. A network can also consist of two or more parts, disconnected from one another, and if an individual part has no loops it is also called a tree. If all the parts of the network are trees, the complete network is called a *forest*.



Planar networks

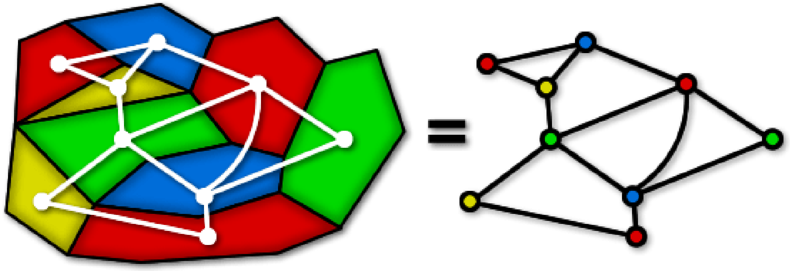
A *planar network* is a network that can be drawn on a plane without having any edges cross. Note that it is in most cases possible to find a way to draw a planar network so that some edges do cross. The definition of planarity only specifies that at least one arrangement of the vertices exists that results in no crossing.

Planarity testing is in P (many linear algorithms).



A planar graph (left), a plane drawing (center), and a straight line drawing (right), all of the same graph

Four-color theorem: it is possible to color any set of regions on a two-dimensional map with at most four colors such that no two adjacent regions have the same color (no matter how many regions there are or of what size or shape).



Weighted Graphs

- Usually l_E and l_V will be restricted to map onto numeric labels only such as $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$. In such cases if $dim(\text{Range}(l_E)) = 1$ then we will call this graph weighted.

- Unweighted graph $G = (V, E)$ is equivalent to a weighted graph with unit edge weights $\omega_{ij} = 1$ for all $ij \in E$.

- Examples: cost, distance, capacity, strength of interaction, and similarity.

Degrees

- G is undirected, degree of $i \in V$ $d(i) = |\{ij \in E\}|$ (also denoted by k_i)

- G is directed, out-degree of $i \in V$ $d^+(i) = |\{ij \in E\}|$

- G is directed, in-degree of $i \in V$ $d^-(i) = |\{ji \in E\}|$

- sets of neighbors are defined similarly $\Gamma(i)$, $\Gamma^+(i)$, $\Gamma^-(i)$

- $\Delta(G)$, $\overline{d}(G)$, and $\delta(G)$ - maximum, average and minimum degrees of G, respectively

Degrees

A is the adjacency matrix of unweighted undirected graph

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

- $m = \frac{1}{2} \sum_{j=1}^{n} k_j = \frac{1}{2} \sum_{ij} A_{ij}$

- The mean degree is defined as

$$c = \frac{1}{n} \sum_{j=1}^{n} k_j = \frac{2m}{n}$$

- The maximum possible number of edges in a graph (a clique)

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

Directed networks → Undirected networks

- Ignore the edge directions entirely, an approach that can work in some cases. However, it throws out a lot of potentially useful information about the network's structure.

- Or use $A+A^T$ instead of A

- In some applications, such methods as "cocitation" network are useful.

The *cocitation* of two vertices *i* and *j* in a directed network is the number of vertices that have outgoing edges pointing to both *i* and *j*. In the language of citation networks, for instance, the cocitation of two papers is the number of other papers that cite both.

If we use the adjacency matrix, we can see that

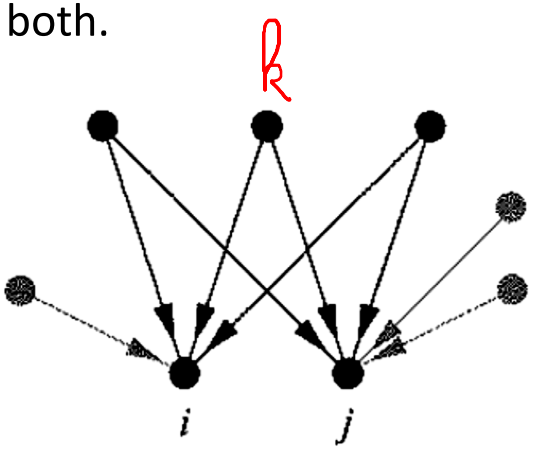
$$A_{ik}A_{jk} = 1$$

if *i* and *j* are both cited by *k* and 0 otherwise.

Summing over all *k*, we get matrix *C*

$$C_{ij} = \sum_{k=1}^{n} A_{ik}A_{jk} = \sum_{k=1}^{n} A_{ik}A^T_{kj}, \text{ i.e.,}$$

$$C = AA^T = (AA^T)^T = C^T$$

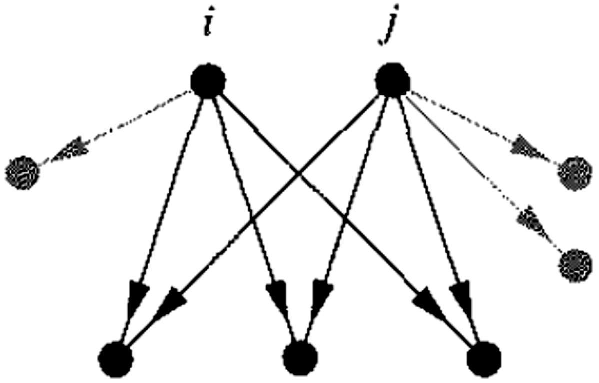


Vertices *i* and *j* are cited by three common papers, so their cocitation is 3.

Directed networks → Undirected networks

- In some applications, such methods as "bibliographic coupling" networks are useful.

The *bibliographic coupling* of two vertices in a directed network is the number of other vertices to which both point.



Vertices i and j cite three of the same papers and so have a bibliographic coupling of 3.

If we use the adjacency matrix, we can see that

$$A_{ki}A_{kj} = 1$$

if *i* and *j* both cite *k* and 0 otherwise.

Summing over all *k*, we get matrix *B*

$$B_{ij} = \sum_{k=1}^{n} A_{ki}A_{kj} = \sum_{k=1}^{n} A_{ik}^{T}A_{kj}, \text{ i.e.,}$$

$$B = A^T A = (A^T A)^T = B^T$$

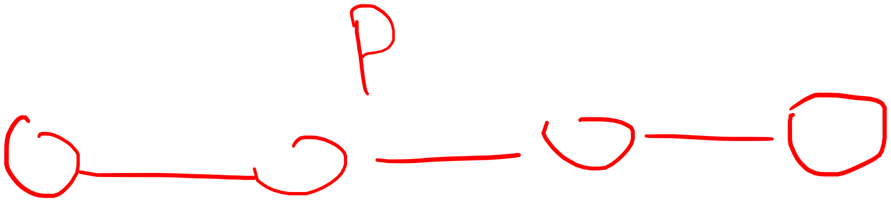
Paths

- A *path* in a network is any sequence of vertices such that every consecutive pair of vertices in the sequence is connected by an edge in the network.

- Paths can be defined for both directed and undirected networks. In a directed network, each edge traversed by a path must be traversed in the correct direction for that edge. In an undirected network edges can be traversed in either direction.

- In general a path can intersect itself, visiting again a vertex it has visited before, or even running along an edge or set of edges more than once. Paths that do not intersect themselves are called *self-avoiding paths.*

- The *length* of a path in an unweighted graph is the number of edges traversed along the path (not the number of vertices).



$|P|=3$

Computing the number of paths

For either a directed or an undirected simple graph the element A_{ij} is 1 if there is an edge from j to i, and 0 otherwise.

- $A_{ik}A_{kj} = 1$ if there is a path of length 2 from j to i via k, and 0 otherwise.

- The total number of paths of length 2 from j to i, via any other vertex, is

$$N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik}A_{kj} = [A^2]_{ij}$$

- The total number of paths of length 3 from j to i, via any other 2 vertices, is

$$N_{ij}^{(3)} = \sum_{k,l=1}^{n} A_{ik}A_{kl}A_{lj} = [A^3]_{ij}$$

- Generalizing to the paths of length r, $N_{ij}^{(r)} = [A^r]_{ij}$

Shortest paths

- For a path $p = (e_1 e_2 e_3 \dots e_k)$ in G with edge weights ω, we define the weight of p

$$\omega(p) = \sum_{e_i \in p} \omega_{e_i}$$

- Shortest path from $i \in V$ to $j \in V$ (wrt to ω) is a path p with smallest possible $\omega(p)$ among all $i - j$ paths.
- No negative weight cycle: $O(m + n \log n)$ (version of Dijkstra alg.); otherwise detect a cycle $O(mn)$ (Bellman-Ford)

Subgraphs

- A graph $G' = (V', E')$ is a subgraph of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$.
- An edge-induced subgraph G', denoted by $G[E']$, is created by chosing $E' \subseteq E$, and $V' = \{i \in V \mid ij \in E'\}$
- A vertex-induced subgraph G', denoted by $G[V']$, is created by chosing $V' \subseteq V$, and $E' = \{ij \in E \mid i, j \in V'\}$

Connected components

- Undirected graph G is connected if there is a path from every vertex to every other vertex.
- A connected component of G is an induced subgraph G' that is connected and *maximal*.
- Checking whether G is connected and finding all its connected components can be done in time $O(n+m)$ using DFS or BFS.
- A directed graph is strongly connected if there is a directed path from every vertex to every other vertex.
- A strongly connected component of a directed G is an induced subgraph that is strongly connected and maximal. Checking - same time.

Minimum cuts

- A cut is a partition $(S, \overline{S})$ of V, s.t. $S, \overline{S} \neq \emptyset$. The capacity of a cut $(S, \overline{S})$ is defined as
$$\sum_{ij \in E, i \in S, j \in \overline{S}} \omega_{ij}.$$
- A minimum cut is a cut whose capacity is minimum among all cuts.
- Finding minimum cut is easy (polynomial time).
- Finding minimum cut with certain constraints on $|S|$ is hard. For example, when
$$|S| = |V|/2$$
the problem is called graph bisectioning. This is one of the most important NP-hard problems.

k-connectivity

- An undirected $G = (V, E)$ is called k-vertex-connected if $|V| > k$, and $G[\overline{X}]$ is connected for every $X \subset V$ with $|X| < k$.

- The vertex connectivity of G is the largest k such that G is k-vertex-connected

- Edge connectivity is defined similarly.

- Finding minimum balanced vertex separator is extremely important for showing robustness of network. It is NP-hard.

- Enumerating the separators is even more important.

Discuss homework 1