Introduction to Network Science

Instructor: Ilya Safro, 228 McAdams Hall
Course: Introduction to Network Science, CP SC 481/681/881
Time: 9:30-10:45 TTh
Place: 211 Daniel

Course Structure

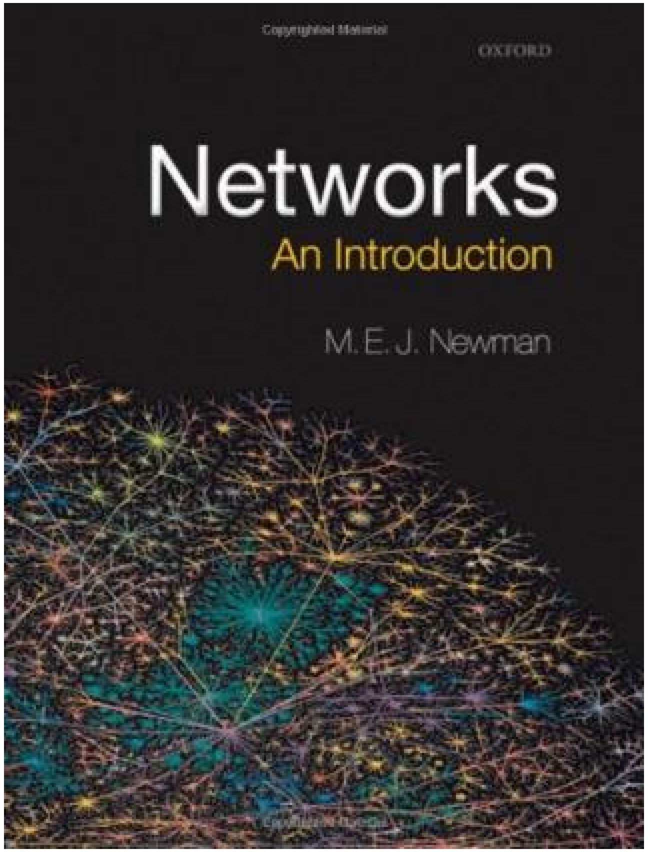
What	How many	Time	Points
Homework		2 weeks	30
Paper reading		1 week	30
Oral presentation	1	3-4 weeks	40
Total			100

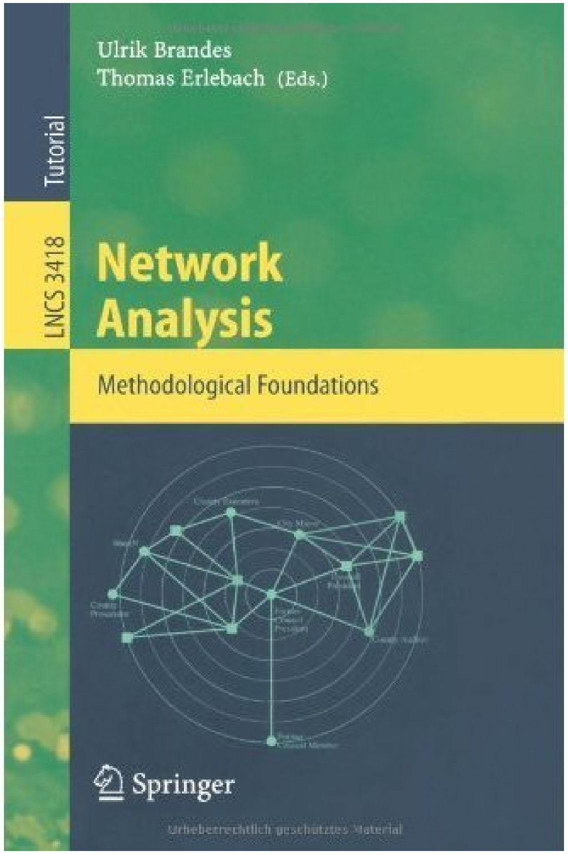
Points	Grade
≥ 90	A
≥ 80	B
≥ 60	C
≥ 0	F

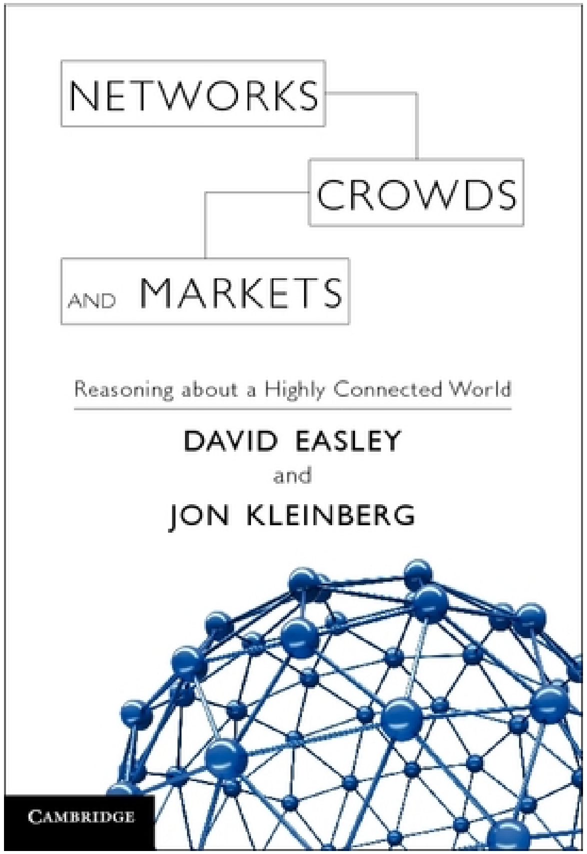
Bonuses

Work in class, extra work in home exercises, etc. - up to 10 points. We do not want to miss the next Turing and Fields laureates, so any submitted conference/journal paper or technical report 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).

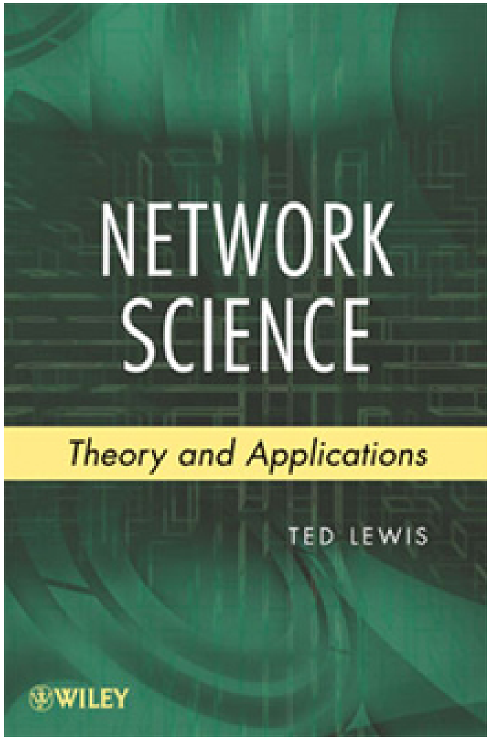
Recommended Books (optional)

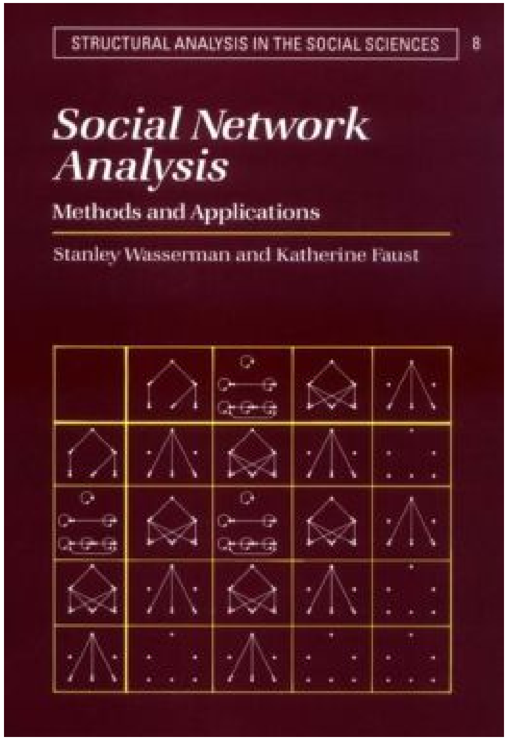






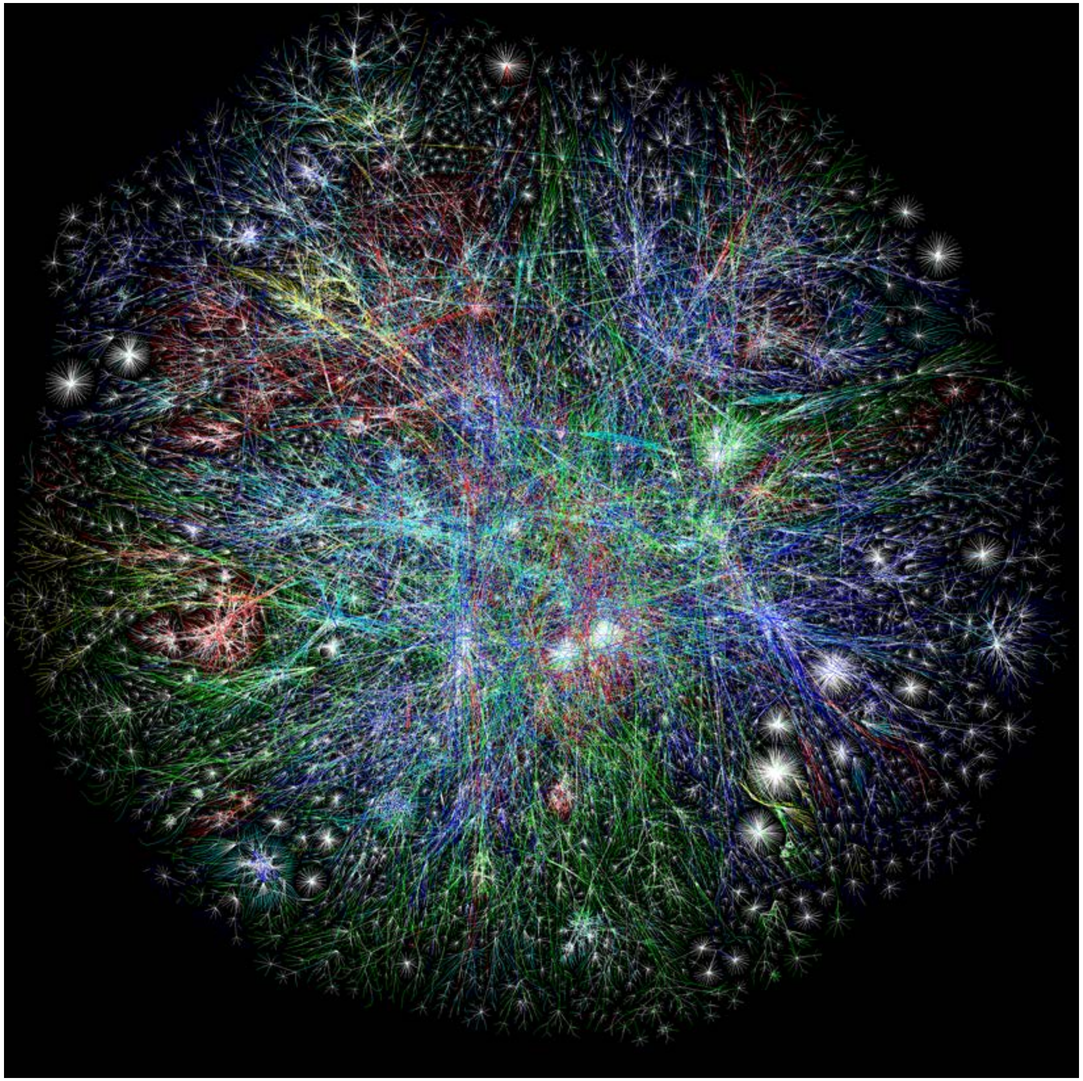
Recommended Books (optional)






WILEY
ANALYSIS OF
BIOLOGICAL NETWORKS
BJÖRN H. JUNKER
FALK SCHREIBER
Tata McGraw-Hill Professional: Networking Series
COMMUNICATION NETWORKS
Principles and Practice
SUMIT KASERA | NISHIT NARANG
SUMITA NARANG
Advanced Electric
Power Network Analysis
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
The SAGE Handbook of
Social Network Analysis
Edited by
John Scott and
Peter J. Carrington

facebook



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

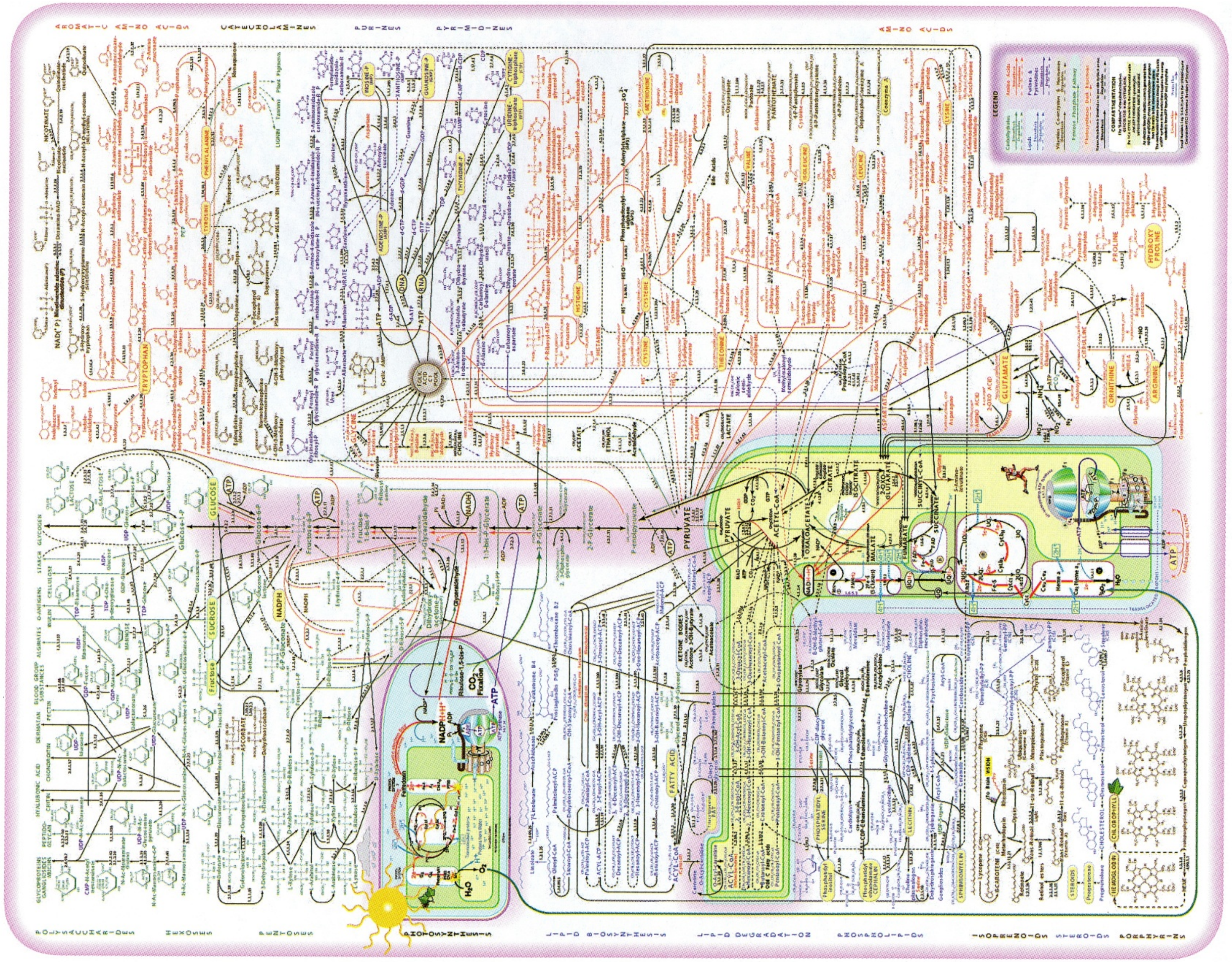
VoIP: will be merged in future?

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

- Separate circuit for each call
- No packets

Research Problems

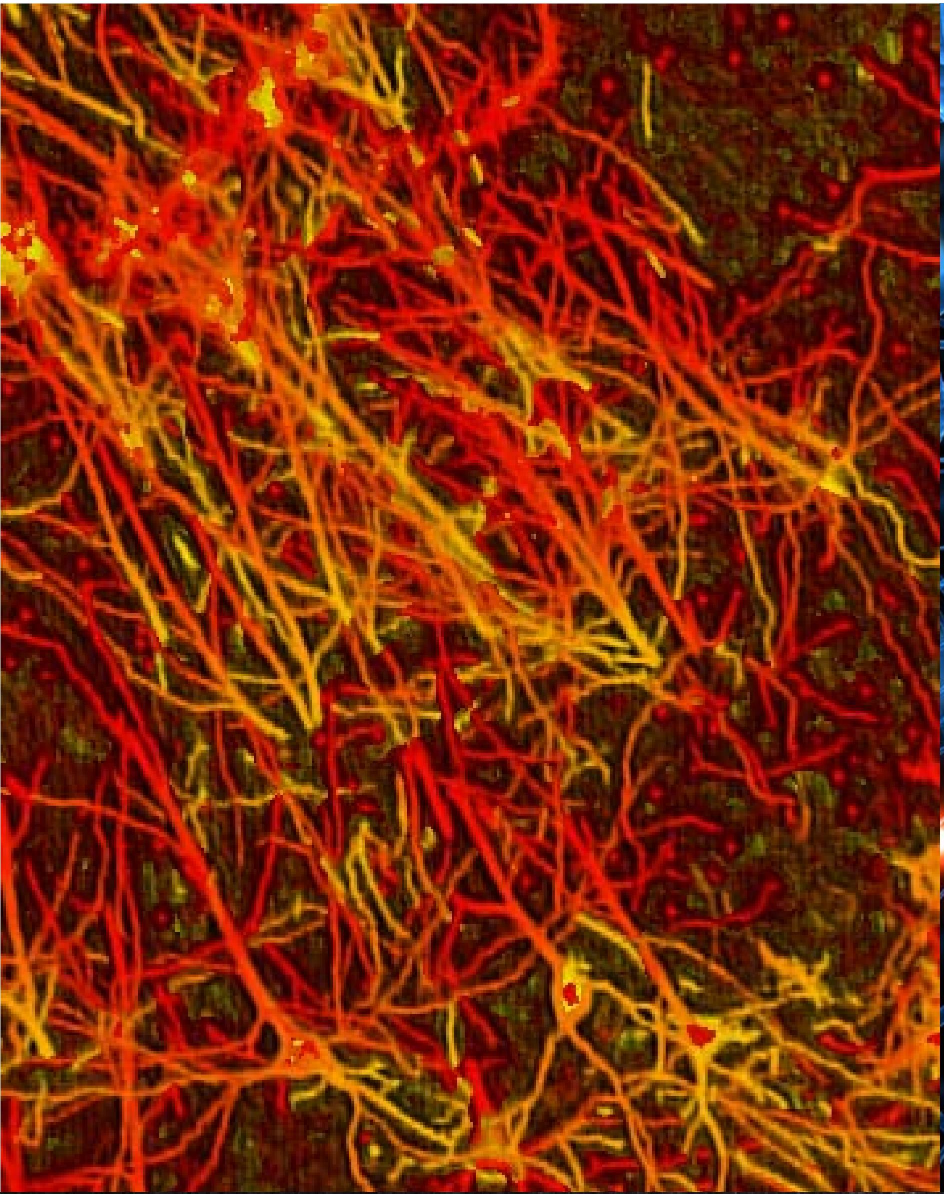
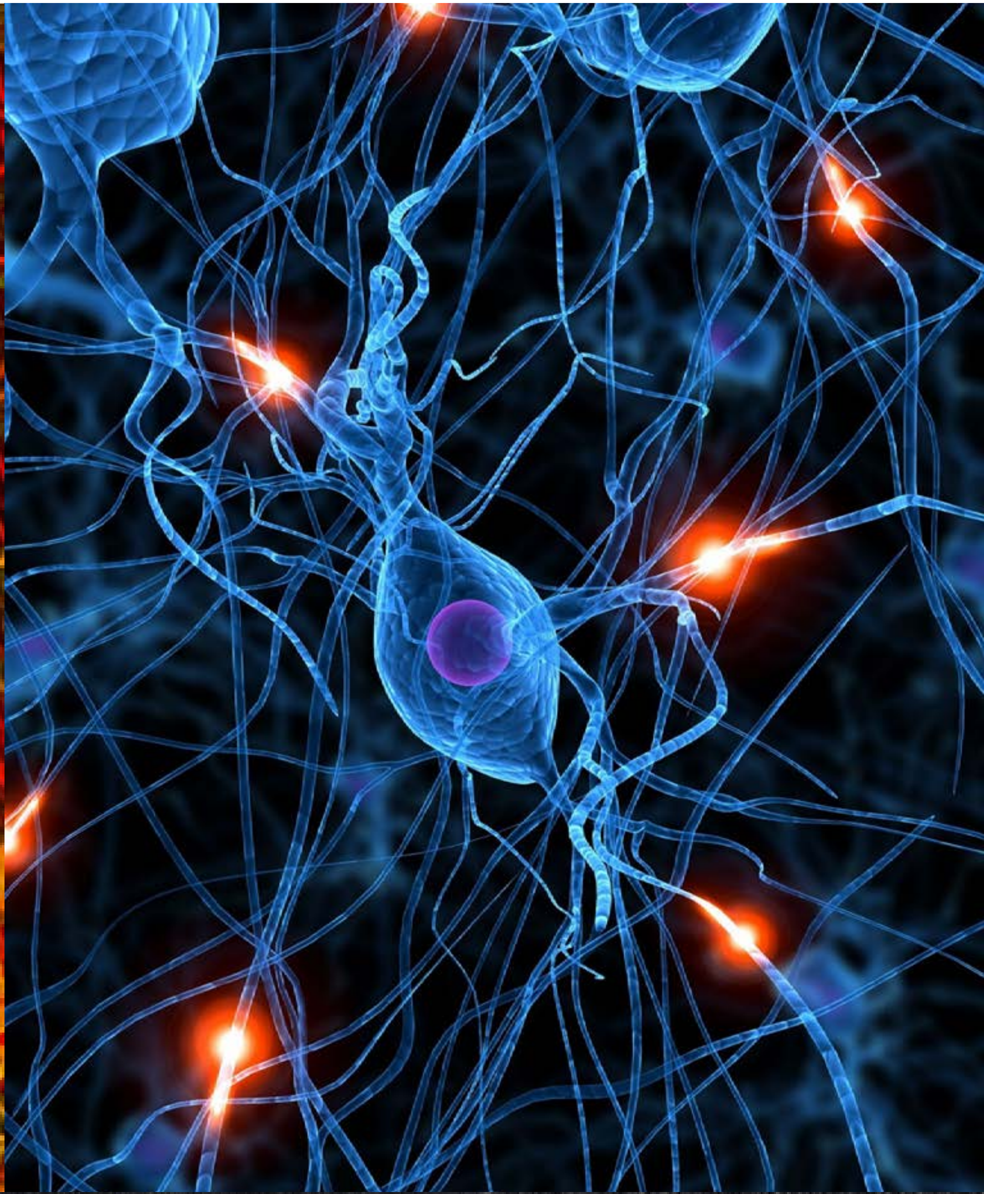
- Network congestion
- Robustness (wrt disasters, etc.)
- Dynamics



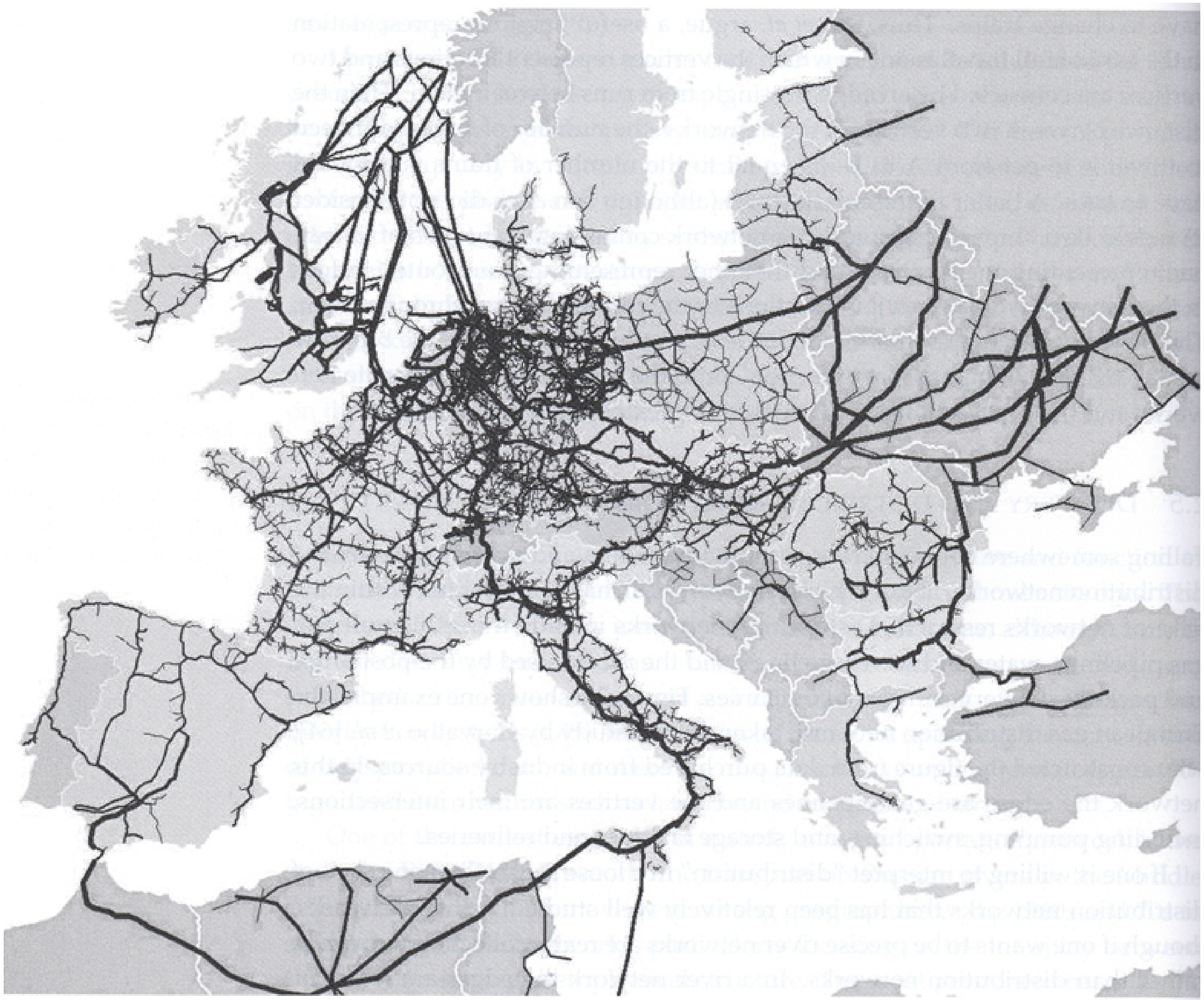
Network formed by the major metabolic pathways (Newman “Networks: An Introduction”)

Brain Networks

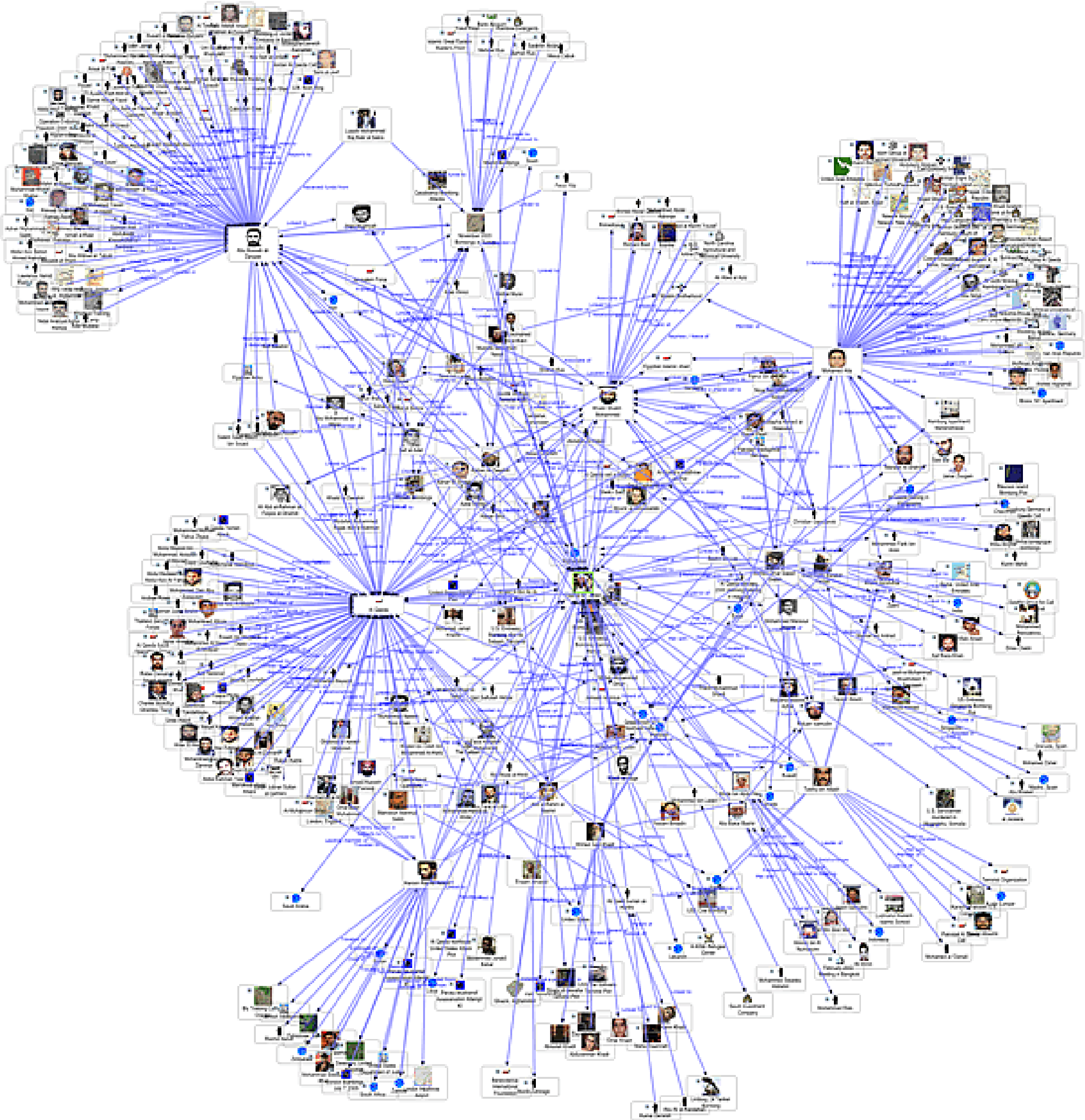
from networksandservers.blogspot.com

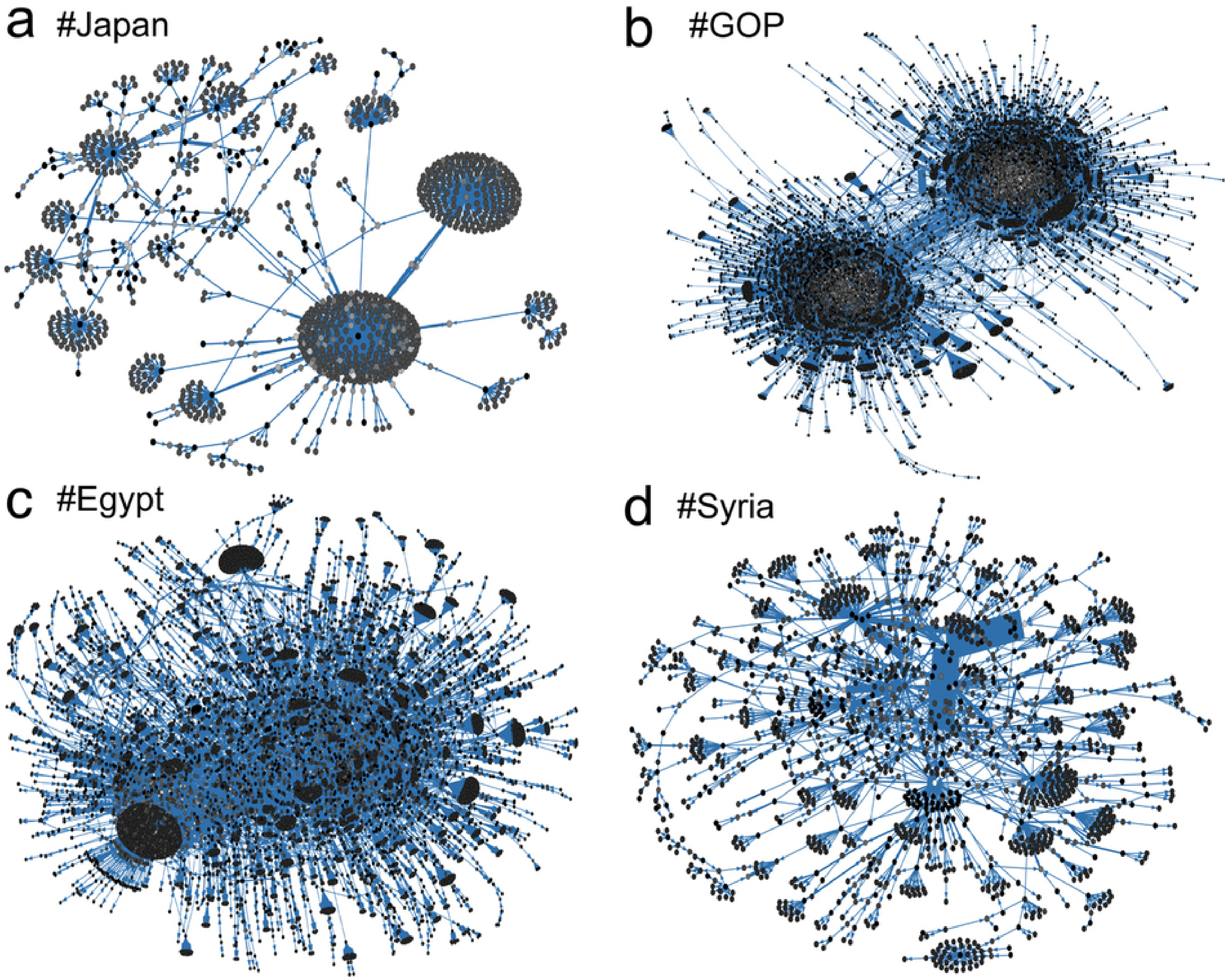
Delivery and Distribution Networks



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



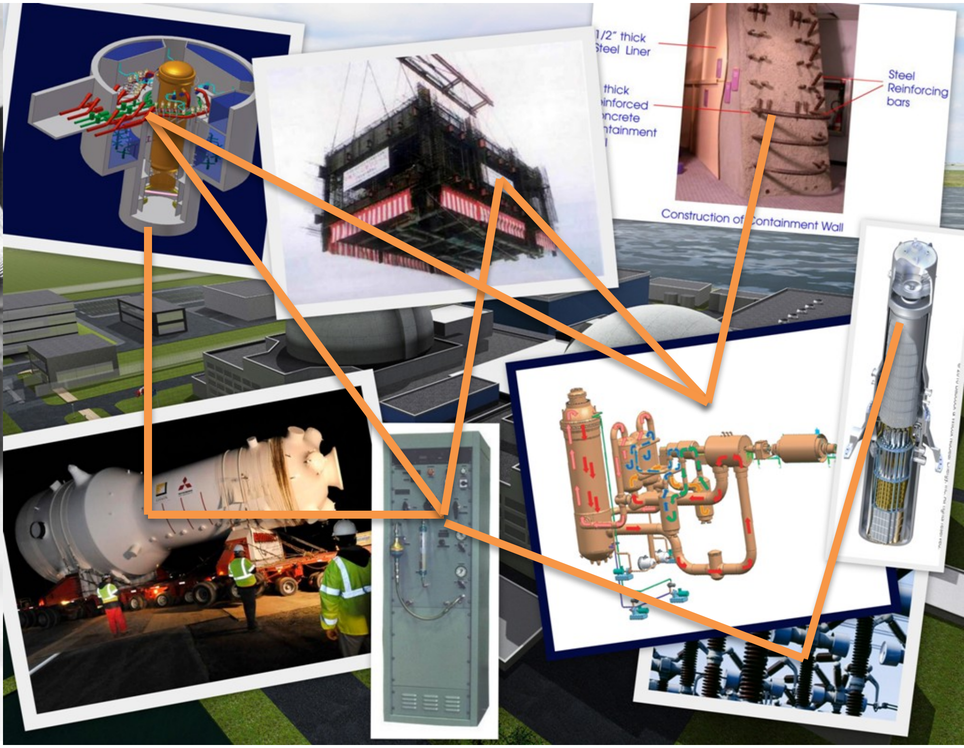
Network of 9/11 contacts (Krebs)



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

Actors, Reactors and More

IMDB, Network of movie actors



Transportation and logistics network of nuclear power plant

Definition

Network is an object composed of (not necessarily equal) elements and interactions or connections between these elements. Examples: Internet (servers-routers-computers/fiber-optic-wireless connections), Epidemiology (people-places/contacts).

Definition

A labeled (attributed) graph $G = (V, E, l_E, l_V)$ is an abstract object formed by a set V of vertices (nodes), a set E of edges (links) that join (connect) pairs of vertices, and functions l_E and l_V that map V and E, respectively, onto the set of numeric and non-numeric labels. Example: Internet link $ij \to$ (500 common users, 1Gb capacity, optic, blue).

- vertex set = $V(G)$; edge set $E(G)$; n = cardinality of V; $m = |E|$;
- graphs can be undirected and directed ($ij \in E$ means $i \to j$)
- directed graph has an underlying undirected graph

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 $i \in V$ $d(i) = |\{ij \in E\}|$
- 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 $\Gamma(i)$, $\Gamma^+(i)$, $\Gamma^-(i)$
- Δ(G), $\overline{d}$ (G) , and δ(G) - maximum, average and minimum degrees of G, respectively

Subgraphs

• A graph $G' = (V', E')$ is a subgraph of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$
• The edge-induced subgraph with edge set $E' \subseteq E$, denoted by $G[E']$, is the subgraph $G' = (V', E')$ of G, where $V' = \{i \in V | ij \in E'\}$
• A vertex-induced subgraph if E' contains all edges $ij \in E$ that join vertices in V'

Connected Components

• Undirected G is connected if there is a path (alternating sequence $i_1 e_1 i_2 e_2 \cdots e_{k-1} i_k$) from every vertex to every other vertex. For a given undirected graph $G = (V, E)$, a connected component of G is an induced subgraph $G' = (V', E')$ that is connected and maximal.
• Checking whether a graph is connected and finding all its connected components can be done in time $\mathcal{O}(n + m)$ 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 graph G is an induced subgraph that is strongly connected and maximal. Checking - same time.

Shortest Paths

• For a path $p = (e_1 e_2 \cdots e_k)$ in G with edge weights ω, we define the weight of the path $\omega(p) = \sum_{e_i \in p} \omega_{e_i}$.
• Shortest path from i to j (wrt ω) is the smallest possible among all paths from i to j.
• No cycle of negative weight: $\mathcal{O}(m + n \log n)$ (version of Dijkstra alg.); otherwise detect a cycle $\mathcal{O}(mn)$ (Bellman-Ford)

Minimum Cuts

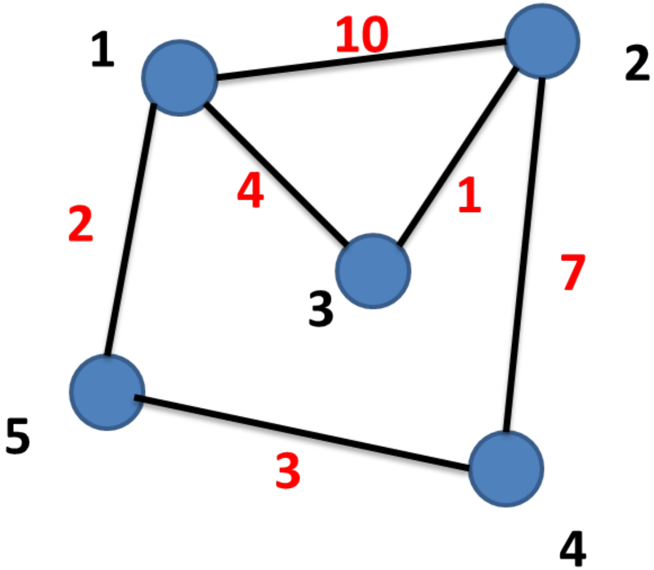
• A cut is a partition $(S, \overline{S})$ of V, s.t. $S, \overline{S} \neq \emptyset$. The capacity of a cut $(S, \overline{S})$ is defined as

$$\sum_{ij \in E,\ i \in S,\ j \in \overline{S}} \omega_{ij}.$$

A minimum cut is a cut whose capacity is minimum among all cuts.
• Finding min cut is easy; Finding min cut with certain constraints on $|S|$ is one of the most important NP-hard problems.

k-Connectivity

- An undirected graph G = (V,E) is called k-vertex-connected if $|V| > k$ and G-X is connected for every $X \subset V$ with $|X| < k$.
- The vertex-connectivity of G is the largest k s.t. G is k-vertex-connected.
- Edge connectivity is defined similarly.
- Finding minimum balanced vertex separator is extremely important for showing robustness of the network. The problem is hard. Enumerating the separators is even more important and difficult.



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

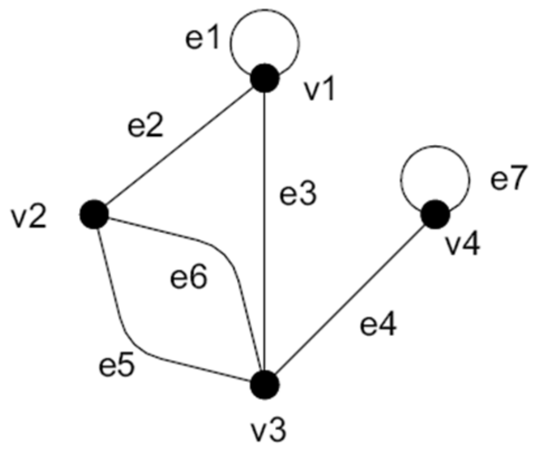
eigenvalues are real

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

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

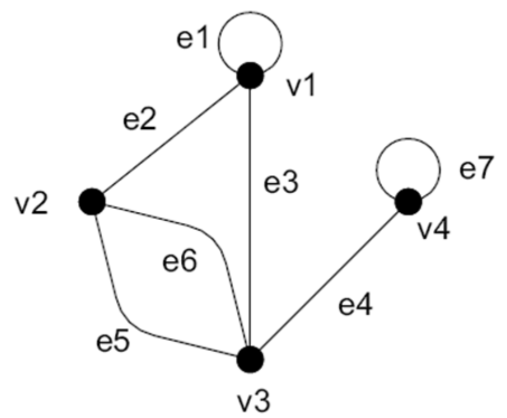
normalized Laplacian $\mathcal{L} = D^{-\frac{1}{2}} \cdot L \cdot D^{-\frac{1}{2}}$

Data Structures



	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

Large space, fast access. Space be improved with compressed row format.



	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

Homework 1

Do not submit #1

1. Additional topics for a brief review. Can be found in many textbooks (including Cormen, Leiserson, Rivest "Intro to Algs")

- Network flow; Ford-Fulkerson; relationship between maximum flow and minimum cut
- Linear programming (LP) is solvable in polynomial time; Integer LP is NP-hard
- Basic properties of graph Laplacian spectrum
- Classical NP-complete problems: k-partitioning, traveling salesman problem, minimum linear arrangement and bandwidth, balanced vertex separator
- Random walks
- Graph isomorphism
- Weighted matching problem

Homework 1

Submit report

2. Download and play with graph libraries and packages

- Graphs and Algorithms Libraries: LEDA (Algorithmic Solutions, C++), LEMON (COIN-OR, C++), Boost Graph Library (C++), iGraph (C++), NetworkX (python) - choose one
- Graph Drawing: sfdp (Linux distributions or Graphviz), Gephi, Tulip, Tom Sawyer Software. Draw graph 2K-50K nodes.

3. Find and browse datasets with real-world networks

- University of Florida Sparse Matrix Collection
  `http://www.cise.ufl.edu/research/sparse/matrices`
  Stanford Network Collection
  `http://snap.stanford.edu`
- Upload graphs using your library (what is the maximum size?)
- Can you add attributes to nodes/edges?
- Can you implement/run basic algorithms such as BFS?
- Try to compute up to 5 smallest eigenvalues of L with one cc (LAPACK++, GSL, SLEPc, PetSc).