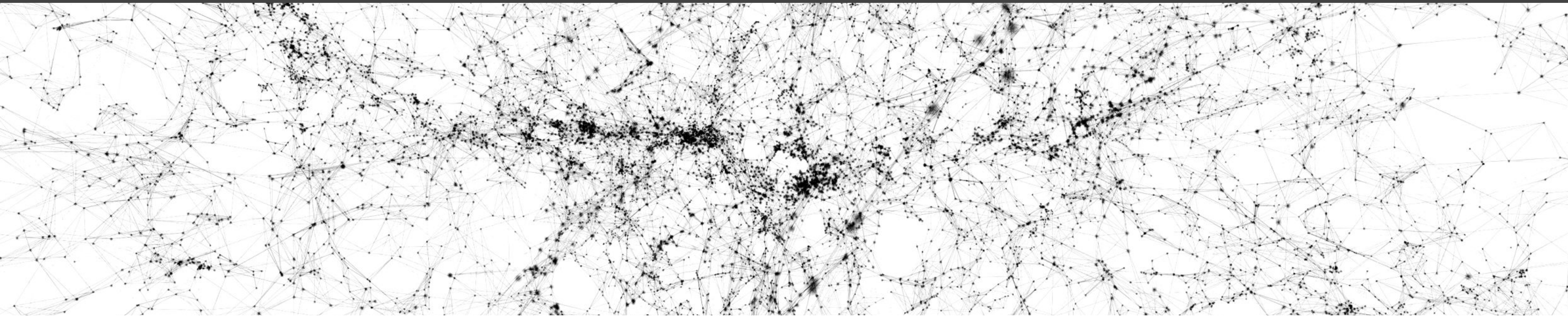
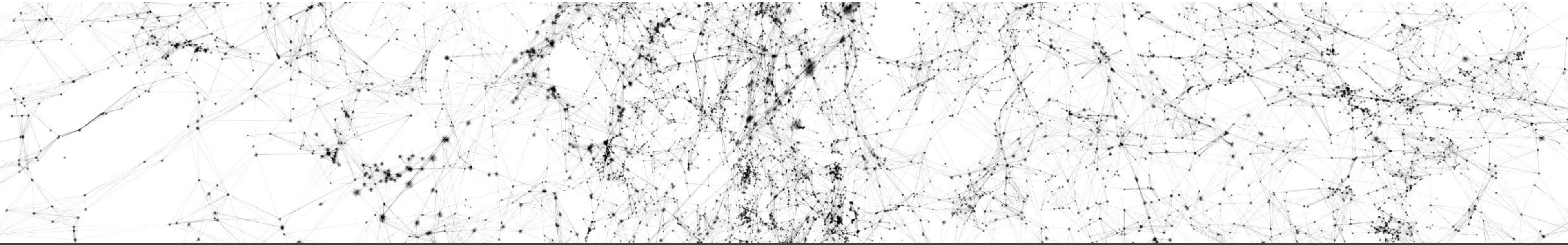


Background

Analysis of complex interconnected data



Announcement:

first assignment out, check the description, and partner up

http://www.reirab.com/Teaching/NS21/Assignment_1.pdf

Proposals are individually

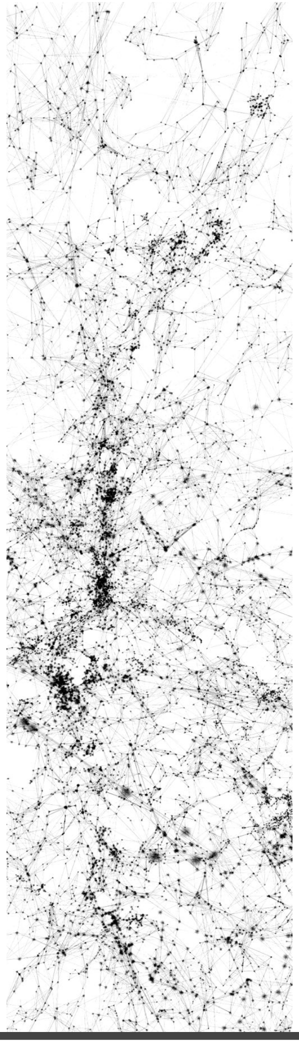
After proposal presentations, you can decide to join another project and continue as a group of two or complete the project individually

Deadlines

- assignment 1 due on Sep. 20th
- assignment 2 due on Oct. 4th
- assignment 3 due on Oct. 18th
- project proposal slides due on Oct. 18th
- project proposal due on Oct. 20th
- Reviews (first round) due on Oct. 27th
- project proposal slides due on Nov. 3rd
- project progress report due on Nov. 5th
- Reviews (second round) due on Nov. 12th
- project final report slides due on Nov. 29th
- project final report due on Dec. 7th
- Reviews (third round) due on Dec. 14th
- project revised report and rebuttal due on Dec. 20th
- note: dates are tentative, subject to change

Outline

- Learning the vocabulary of Network Science
  - Evolution of the field and scale of the data
  - Types of Networks: simple, directed, temporal, bipartite, etc
  - Adjacency matrix, Laplacian matrix

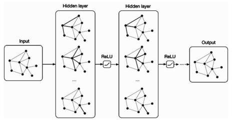


Timeline of notable works in network science

Recent Trend: Deep Learning for Graphs

2015~2021

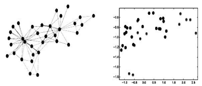
- **Graph Neural Networks**
- Deep Learning for Networks
- High-Order Networks [Benson et al.]

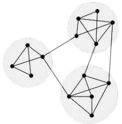


21st Century: More CS

2010~2014

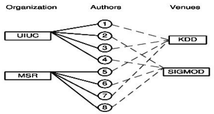
- Info. vs. Social Networks (Twitter) [Kwak et al.]
- **Signed** Networks [Leskovec et al.]
- Semantic Social Networks [Tang et al.]
- Four Deg. Of Separation [Backstrom et al.]
- Structural Diversity [Ugander et al.]
- Computational Social Science [Watts]
- **Network Embedding** [Perozzi et al.]





2005~2009

- Graph Evolution [Leskovec et al.]
- 3 Deg. Of Influence [Christakis & Fowler]
- Social **Influence** Analysis [Tang et al.]
- Six Deg. Of Separation [Leskovec & Horvitz]
- Network **Heterogeneity** [Sun & Han]
- Network **Embedding** [Tang & Liu]
- Computer Social Science [Lazer et al.]



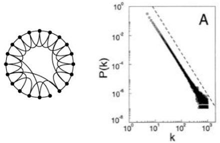
2000~2004

- Influence Max'n [Domingos & Kempe et al.]
- **Community Detection** [Girvan & Newman]
- Network Motifs [Milo et al.]
- Link Prediction [Liben-Nowell & Kleinberg]

Late 20th Century: CS & Physics

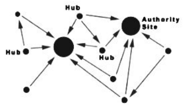
1998~1999

- **Small Worlds** [Watts & Strogatz]
- **Scale Free** [Barabasi & Albert]
- **Power Law** [Faloutsos x3]



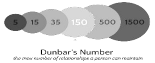
1997

- **HITS** [Kleinberg]
- **PageRank** [Page & Brin]
- Hyperlink Vector Voting [Li]



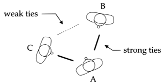
1992

- Structural Hole [Burt]
- **Dunbar's Number** [Dunbar]



1970s

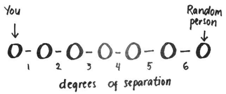
- The Strength Of **Weak Tie** [Granovetter]



20th Century: Sociology

1960s

- **Small Worlds** [Migram]

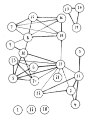


1950s

- **Random Graph** [Erdos, Renyi, Gilbert]
- Degree Sequence [Tuttle, Havel, Hakami]
- **Homophily** [Lazarsfeld & Merton]
- Balance Theory [Heider et al.]

1930s

- **Sociogram** [Moreno]

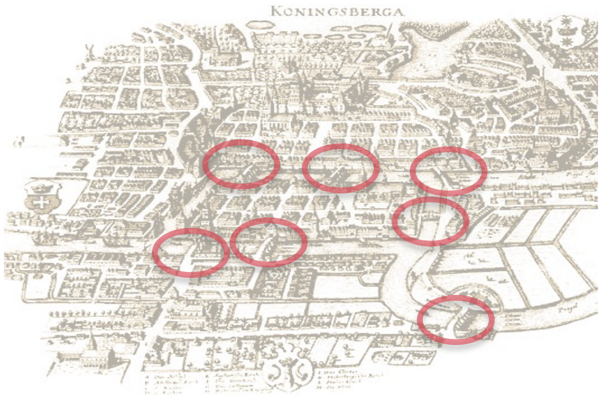


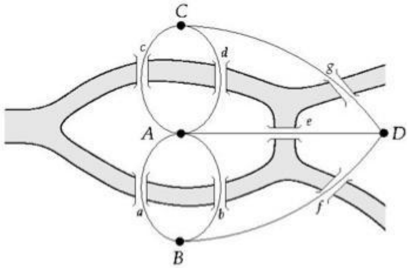
Based on Slides from Jie Tang

Graph Theory & Network Science

Graph theory is older than network science

Can one walk across the seven bridges and never cross the same bridge twice?
[see the video]





1735: **Euler's theorem:**

If a graph has more than two nodes of odd degree, there is no [Eulerian] path. If a graph is connected and has no odd degree nodes, it has at least one path.

Network science borrows many concepts/theories from graph theory. The focus, however, is on **real world** graphs which have specific characteristics, and are different than random graph families commonly studied in math.

for example, regular graphs (same degree for all nodes), are irrelevant here.

Real world graphs are Large Scale

facebook
- 2 billion MAU
- 26.4 billion minutes/day

twitter
- 320 million MAU
- Peak: 143K tweets/s

Instagram
- 700 million MAU
- 95 million pics/day

Snapchat
- 300 million MAU
- 30 minutes/user/day

Alibaba Group 阿里巴巴集团
- >777 million trans. (alipay)
- 200 billion on 11/11

WeChat
- QQ: 860 million MAU
- WeChat: 1.1 billion MAU

Based on Slides from Jie Tang

MAU (Monthly active users)

Example benchmark datasets

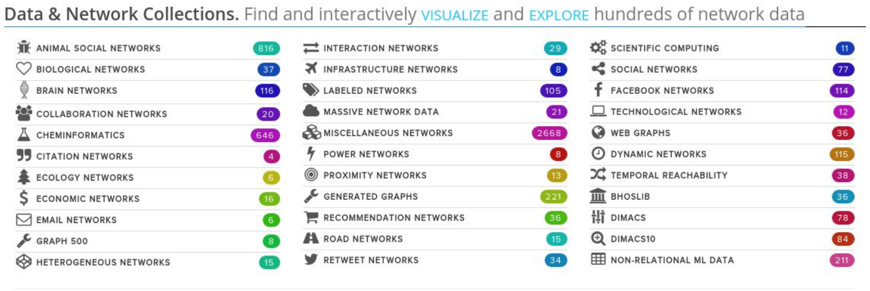
NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

You can download these bundled from Barbasi's website, for the first assignment

Common benchmark repositories

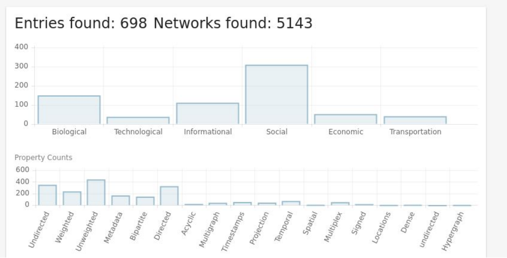
- Stanford Large Network Dataset Collection (SNAP)
  - Social networks : online social networks, edges represent interactions between people
  - Networks with ground-truth communities : ground-truth network communities in social and information networks
  - Communication networks : email communication networks with edges representing communication
  - Citation networks : nodes represent papers, edges represent citations
  - Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
  - Web graphs : nodes represent webpages and edges are hyperlinks
  - Amazon networks : nodes represent products and edges link commonly co-purchased products
  - Internet networks : nodes represent computers and edges communication
  - Road networks : nodes represent intersections and edges roads connecting the intersections

- Network Repository (networkrepository)



Check the visualization demo here: https://networkrepository.com/graphvis.php

- The Colorado Index of Complex Networks (ICON)



- The KONECT Project (KONECT)

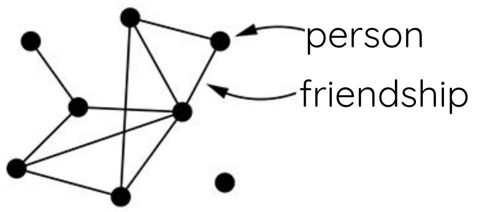
Browse

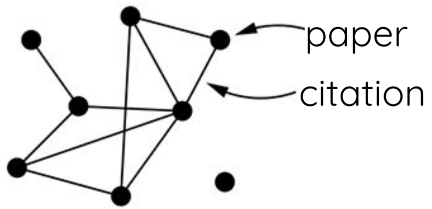
- **Networks**: Karate club • Slashdot Zoo • Twitter followers • more...
- **Statistics**: Clustering coefficient • Diameter • Algebraic connectivity • more...
- **Plots**: Degree distribution • Degree assortativity plot • Hop plot • more...
- **Categories**: Online social networks • Citation networks • Hyperlink networks • more...

Gephi, a notable visualization tool:
https://gephi.org/users/tutorial-visualization/

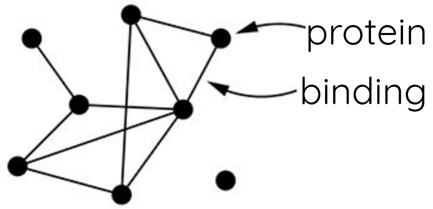
Interconnected Data as Graphs

- Nodes (or Vertices)
  - Proteins, Neurons, People
- Edges (or Links)
  - interactions, friendships





- Two vertices are **adjacent** if they share a common edge
- Two adjacent vertices are **neighbors**
- An edge is **incident** with another edge if they share a vertex
- An edge is incident with two vertices



Adjacency: the default data structure

Adjacency Matrix

	0	1	2	3	4	5	6	7	8	9	10	11
0	0	1	1	0	0	0	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	1	1	0	0	0	0	0	0
4	0	0	0	1	0	1	1	0	0	0	0	0
5	0	0	0	1	1	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	1	1	0	0	0
7	0	0	0	0	0	0	1	0	1	0	0	0
8	0	0	0	0	0	0	1	1	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	1	1	0	1
11	1	0	0	0	0	0	0	0	0	1	1	0

$\mathbf{A} \in \{0,1\}^{N \times N}$

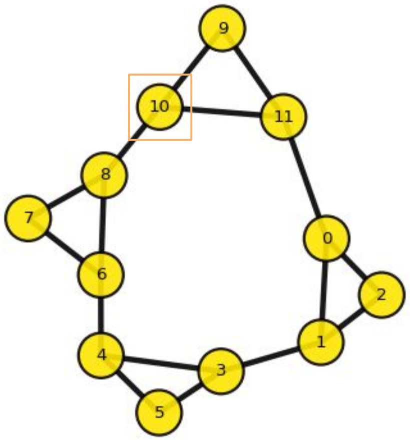
Adjacency List

0 : { 1 , 2 , 11 }
1 : { 0 , 2 , 3 }
2 : { 0 , 1 }
3 : { 1 , 4 , 5 }
4 : { 3 , 5 , 6 }
5 : { 3 , 4 }
6 : { 4 , 7 , 8 }
7 : { 6 , 8 }
8 : { 6 , 7, 10 }
9 : { 10 , 11 }
10 : { 8 , 9 , 11 }
11 : { 0 , 9 , 10 }

Edge List

{ (0, 1), (0, 2), (0, 11),
(1, 0), (1, 2), (1, 3),
(2, 0), (2, 1),
(3, 1), (3, 4), (3, 5),
(4, 3), (4, 5), (4, 6),
(5, 3), (5, 4),
(6, 4), (6, 7), (6, 8),
(7, 8), (7, 6),
(8, 6), (8, 7), (8, 10)
(9, 10), (9,11),
(10, 8), (10, 9), (10, 11),
(11, 0), (11, 9), (11, 10) }

Simple Graph



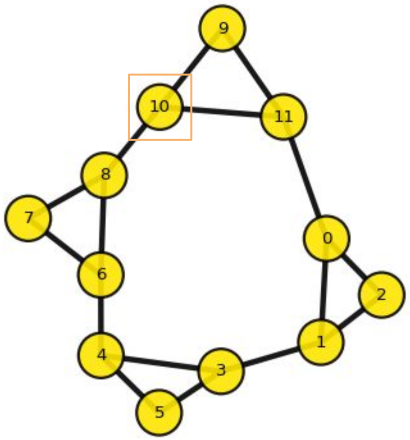
$$G(V,E), V=\{1\ldots n\}, E=\{(i,j)|i,j\in[1\ldots n]) \wedge A_{ij}=1\}$$

Adjacency: sparse representation

Real world graphs are sparse (lots of zeros) and we use sparse matrix representations to only store non-zero elements, in a specific format, often:

- LIL (List of lists): similar to adjacency list
- COO (Coordinate list): similar to edge list
- CSR (Compressed Sparse Row)
  - store only start index of each row
  - fast row access and matrix-vector multiplications

0 : { 1 , 2 , 11 }
1 : { 0 , 2 , 3 }
2 : { 0 , 1 }
3 : { 1 , 4 , 5 }
4 : { 3 , 5 , 6 }
5 : { 3 , 4 }
6 : { 4 , 7 , 8 }
7 : { 6 , 8 }
8 : { 6 , 7, 10 }
9 : { 10 , 11 }
10 : { 8 , 9 , 11 }
11 : { 0 , 9 , 10 }



COL: [1 , 2 , 11 , 0 , 2 , 3 , 0 , 1 , 1 , 4 , 5 , 3 , 5 , 6 , 3 , 4 , 4 , 7 , 8 , 6 , 8 , 6 , 7, 10 , 10 , 11 , 8 , 9 , 11 , 0 , 9 , 10]
ROW: [0 , 3 , 6 , 8 , 11 , 14 , 16 , 19 , 21 , 24 , 26 , 29 , 32]

- CSC (Compressed Sparse Column)

LIL and COO are good for constructing matrices. Once a matrix has been constructed, convert to CSR or CSC format for fast arithmetic and matrix vector operations

Representing Graphs

marginals of A are called degree

$$d_i = \sum_j A_{ij}$$

$$A = \begin{bmatrix} A_{11}, & A_{12}, & \cdots, & A_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1}, & A_{N2}, & \cdots, & A_{NN} \end{bmatrix} \in \mathbb{R}^{N \times N}$$

inlinks — all nodes linking to 1

outlinks

if unweighted then $\in \{0,1\}^{N \times N}$

Laplacian Matrix

$$\underbrace{L}_{\in \mathbb{R}^{N \times N}} = D - A$$

D: diagonal degree matrix

$$L = \begin{bmatrix} d_1, & -A_{12}, & \cdots, & -A_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ -A_{N1}, & -A_{N2}, & \cdots, & d_N \end{bmatrix} \in \mathbb{R}^{N \times N}$$

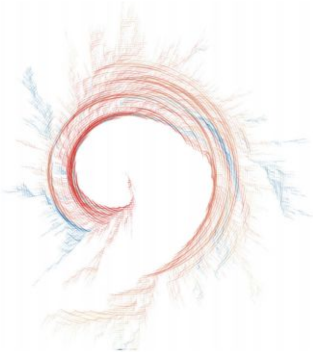
sums to zero (rows)

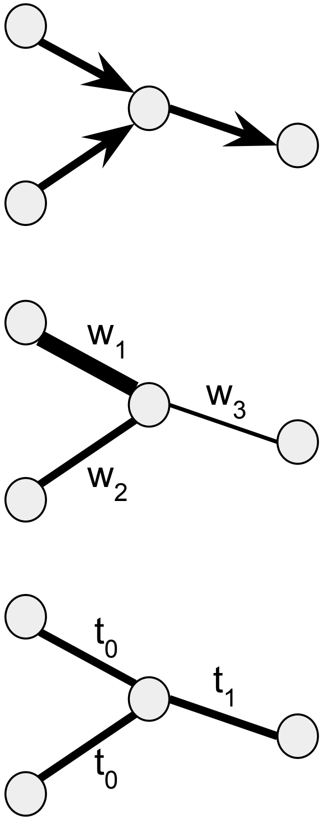
sums to zero (columns)

Eigenvalues of Graph laplacian tells us about the connectivity of the grap

Beyond Simple Graphs

- Directions
  - E.g. who follows who at Twitter
- Weights
  - E.g. friendship strength, or travel cost
- Time
  - E.g. your friendships changes



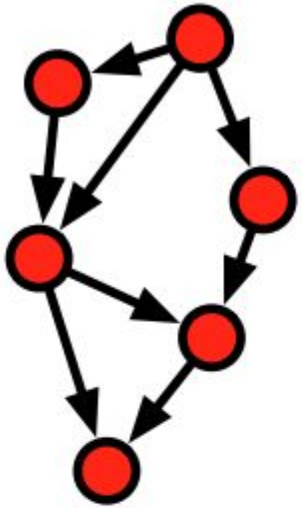


Directed Networks Examples

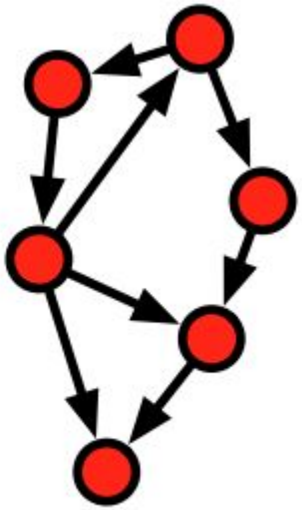
citation networks

foodwebs*

epidemiological



directed acyclic graph



directed graph

WWW

friendship?

flows of goods, information

economic exchange

dominance

neuronal

transcription

time travelers

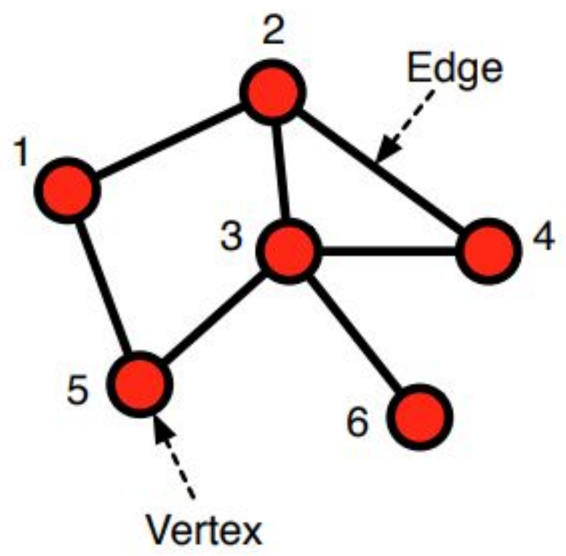
[From Clauset's slides]

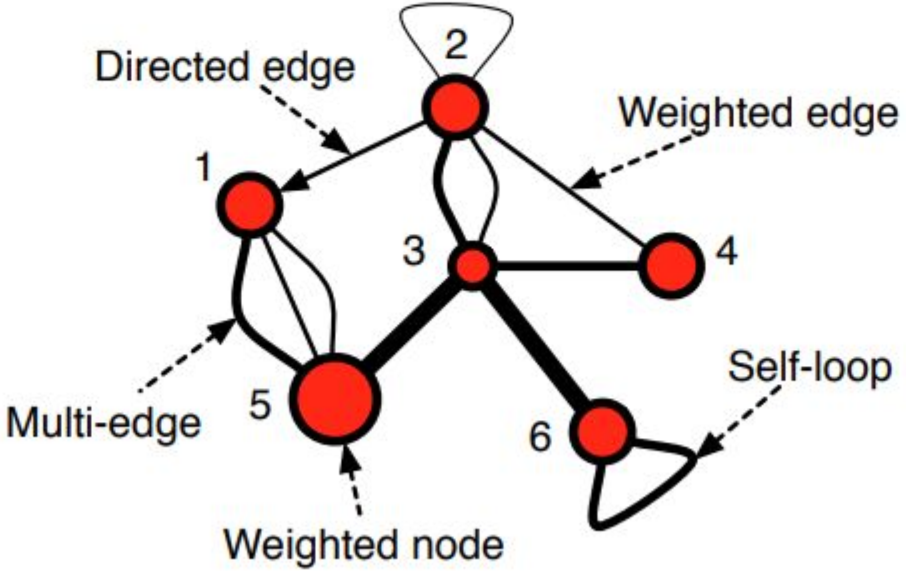
Adjacency Matrix

- Symmetric if graph is undirected
  - $A_{ij} = A_{ji}$
- Directed, not symmetric
  - $A_{ij} \neq A_{ji}$
- Weighted, not binary
  - $[0,1] \Rightarrow R^{+}$
- Temporal
  - **Matrix ⇒ Tensor**

	0	1	2	3	4	5	6	7	8	9	10	11
0	0	1	1	0	0	0	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	1	1	0	0	0	0	0	0
4	0	0	0	1	0	1	1	0	0	0	0	0
5	0	0	0	1	1	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	1	1	0	0	0
7	0	0	0	0	0	0	1	0	1	0	0	0
8	0	0	0	0	0	0	1	1	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	1	1	0	1
11	1	0	0	0	0	0	0	0	0	1	1	0

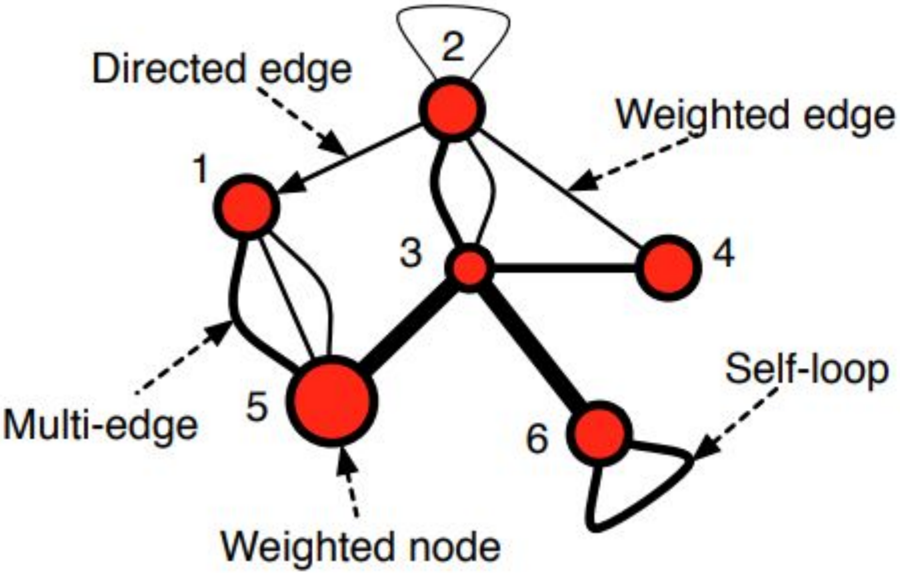
Simple and Not Simple





From Clauset's slides

Example



adjacency matrix

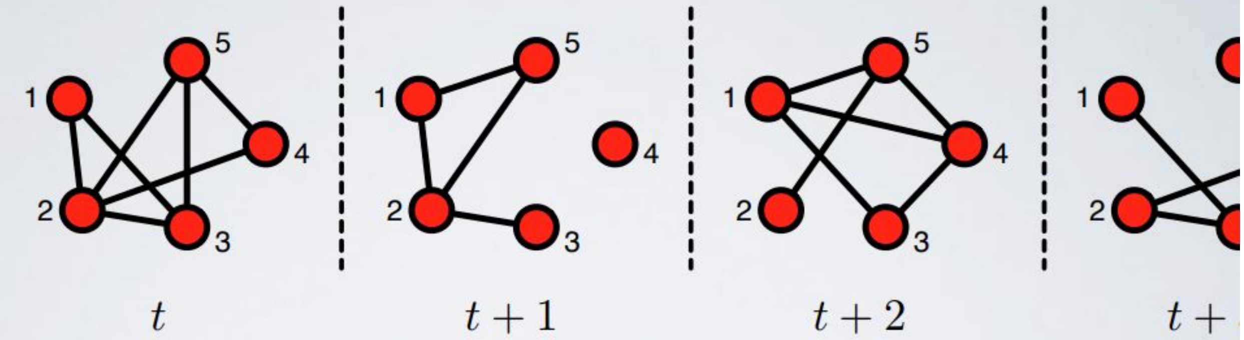
A	1	2	3	4	5	6
1	0	0	0	0	$\{1,1,2\}$	0
2	1	$\frac{1}{2}$	$\{2,1\}$	1	0	0
3	0	$\{2,1\}$	0	2	4	4
4	0	1	2	0	0	0
5	$\{1,1,2\}$	0	4	0	0	0
6	0	0	4	0	0	2

adjacency list

A	
1	$\rightarrow \{(5,1),(5,1),(5,2)\}$
2	$\rightarrow \{(1,1),(2,\frac{1}{2}),(3,2),(3,1),(4,1)\}$
3	$\rightarrow \{(2,2),(2,1),(4,2),(5,4),(6,4)\}$
4	$\rightarrow \{(2,1),(3,2)\}$
5	$\rightarrow \{(1,1),(1,1),(1,2),(3,4)\}$
6	$\rightarrow \{(3,4),(6,2)\}$

From Clauset's slides

Temporal Networks, snapshots or continuous



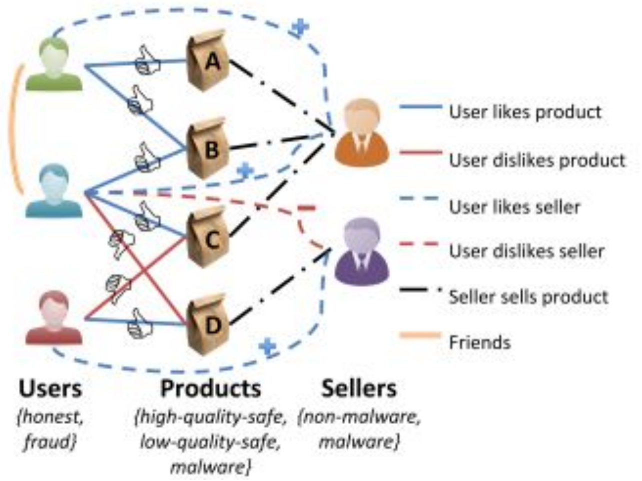
any network over time

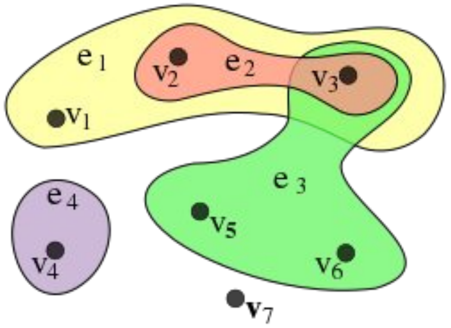
discrete time (snapshots), edges (i, j, t)

continuous time, edges $(i, j, t_s, \Delta t)$

Not Simple Graphs

- Multigraph: Multiple edges
  - E.g. followership & friendship
- Heterogeneous Graphs: Different Types
  - E.g. people, places, interest
- Relation between more than two nodes
  - Hypergraphs, E.g. family
- Relationships in different layers
  - Multiplex or multilayer network

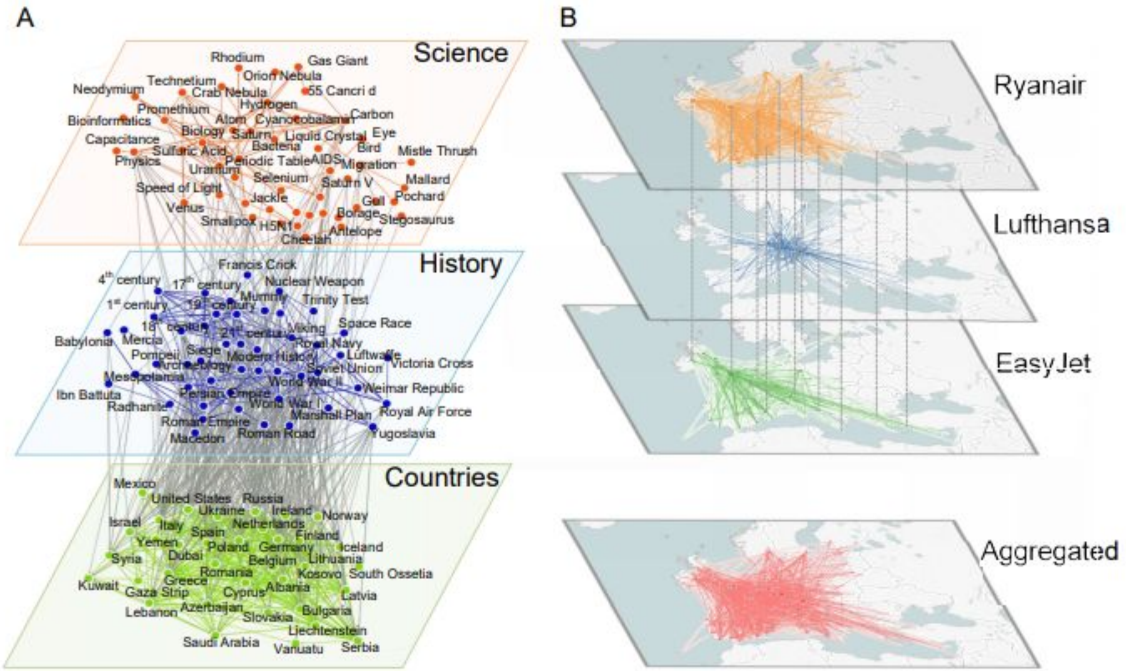




Multilayer Networks

different sets of nodes

E.g. wiki pages layered by subject



Multiplex: same set of nodes

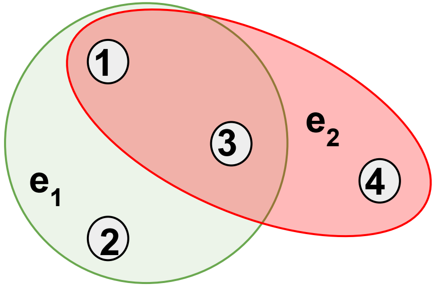
different types of connections

E.g. flights layered by airlines

https://arxiv.org/pdf/1708.07763.pdf

Incidence Matrix

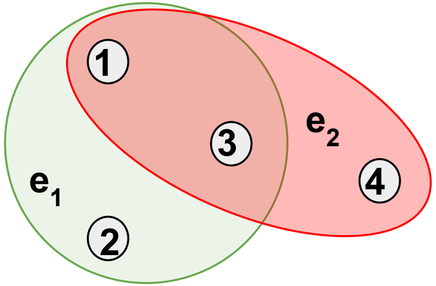
- Adjacency Matrix:
  - A_{ij}= 1 if node i is connected to node j & 0 otherwise
- Incidence Matrix:
  - B_{ik}= 1 if node i is incident to edge k & 0 otherwise
- If a simple graph G has n nodes and m edges what are the dimensions of A & B ?
- How many non-zero elements are in A & B?
- If simple graph, we have 2 ones in each column
  - What is the row marginal of B?
  - $BB^T = A + D$
- Can be used for hypergraphs

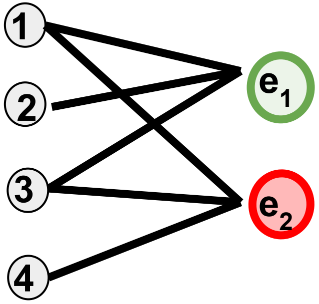


B	e_1	e_2
1	1	1
2	1	0
3	1	1
4	0	1

Incidence Matrix

- Can be used for hypergraphs
  - hyper-edges with more than one node
- Can be used for **bipartite** graphs
  - Two sets of nodes
  - Edges only between them



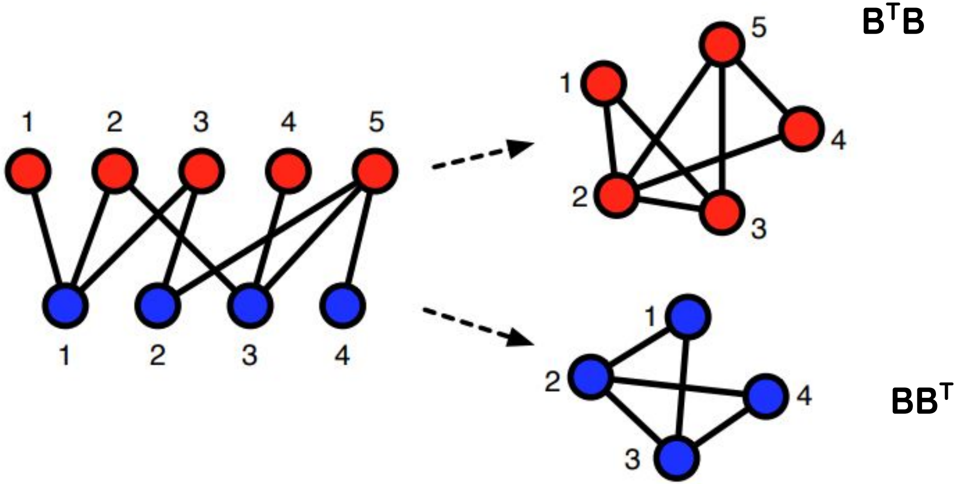


B	e_1	e_2
1	1	1
2	1	0
3	1	1
4	0	1

$V = A \cup B$ where $A \cap B = \emptyset$, and $\forall (i,j) \in E((i \in A) \wedge (j \in B)) \vee ((i \in B) \wedge (j \in A))$

Bipartite Networks

$$B = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



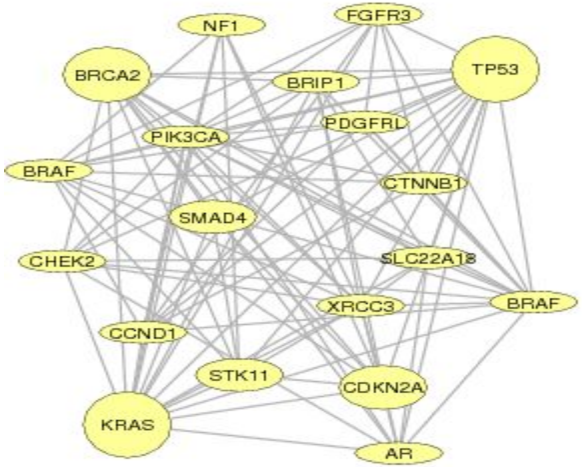
No within edges
& Two possible
one-mode projections
Make the graph to show connections between only one type of node

authors & papers	people & locations (checkins)
actors & movies/scenes	metabolites & reactions
musicians & albums	genes & substrings
people & online groups	words & documents
people & corporate boards	plants & pollinators

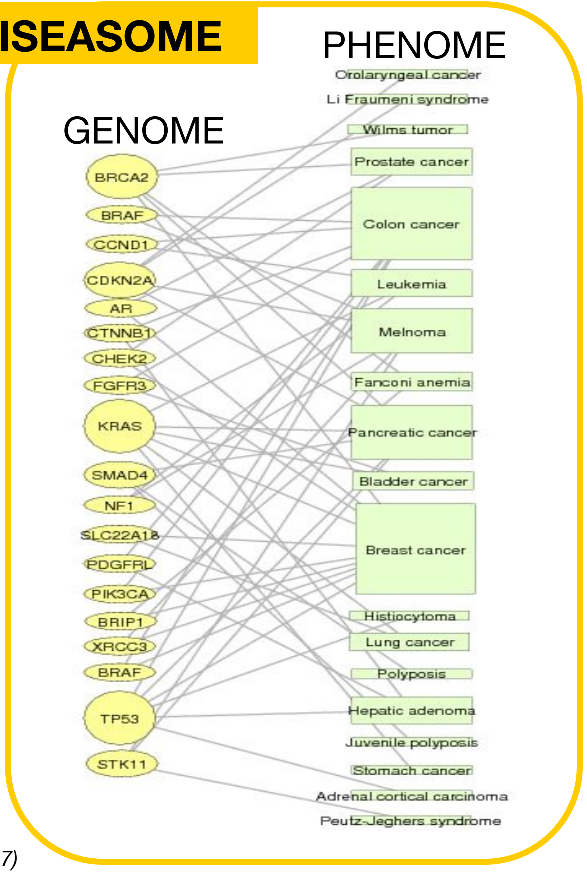
What are the one-mode projection of actors & movies graph?

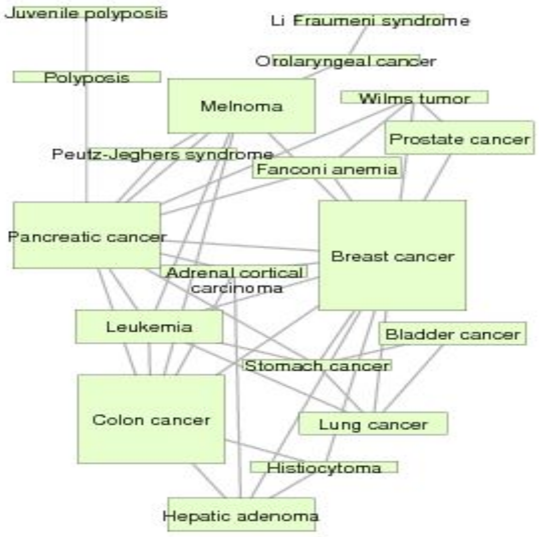
[From Clauset's slides]

Bipartite Networks example



Gene network



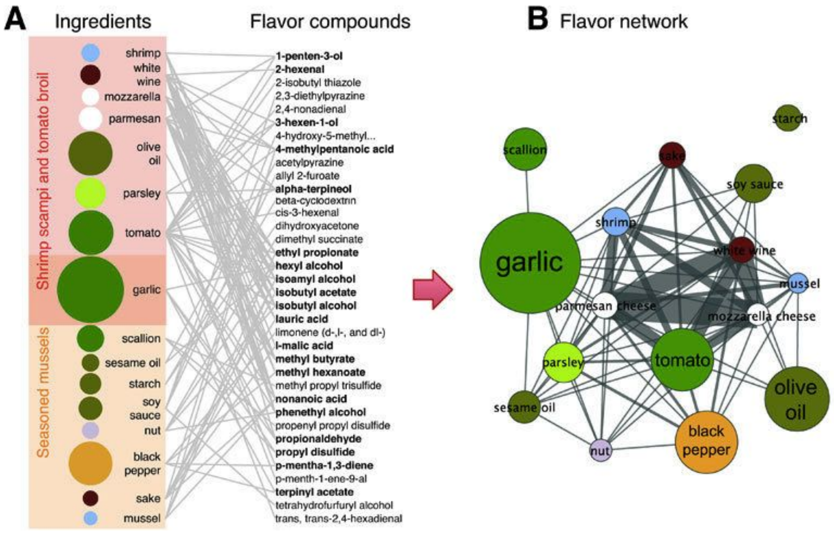


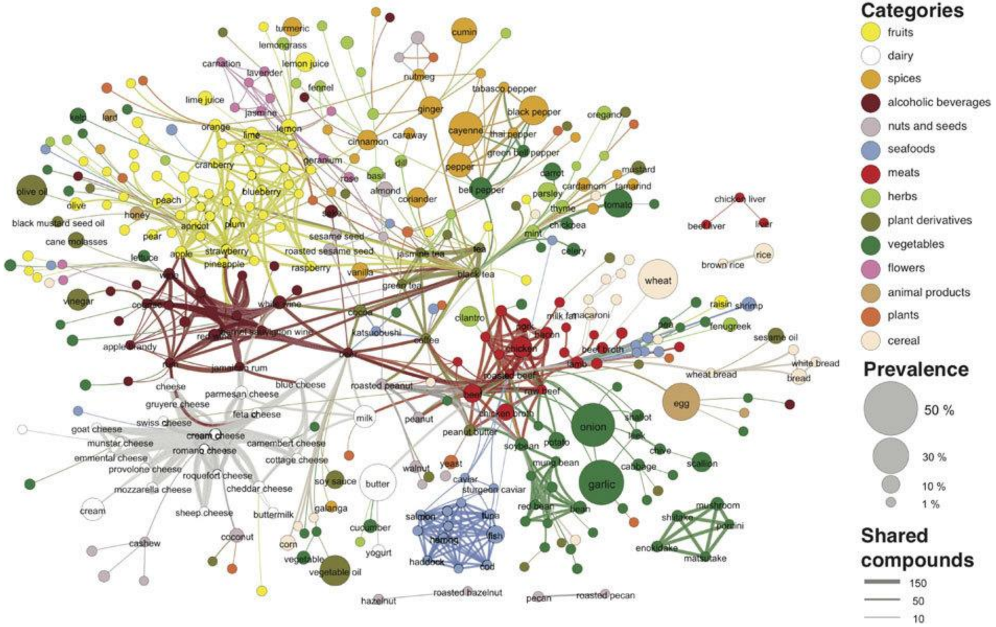
Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

[From Barbasi's slides]

Bipartite Networks example





Ingredient-Flavor Network

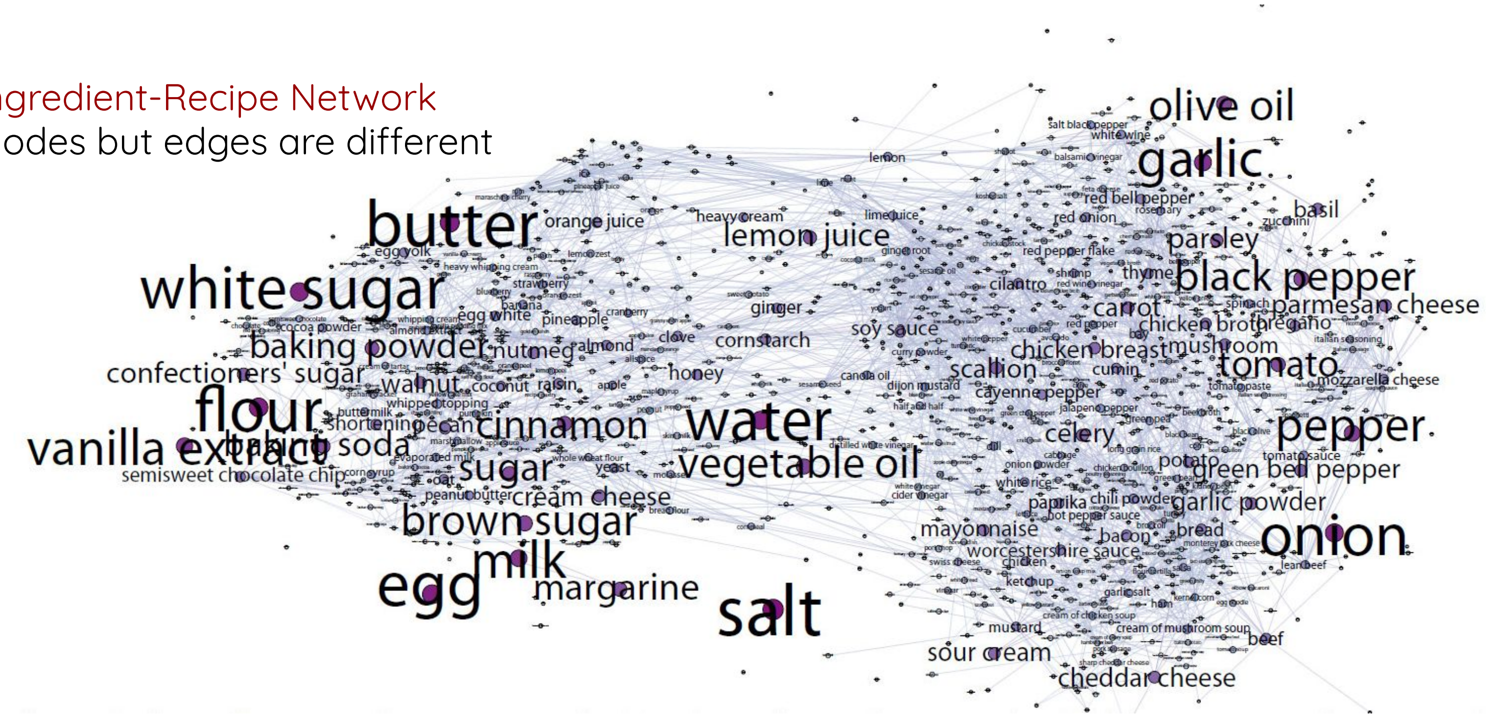
From Barbasi's slides

Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási *Flavor network and the principles of food pairing* , *Scientific Reports 196, (2011).*

Bipartite Networks example

From Ingredient-Recipe Network

Same nodes but edges are different

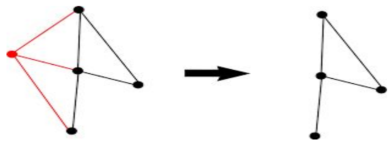


https://arxiv.org/pdf/1111.3919.pdf

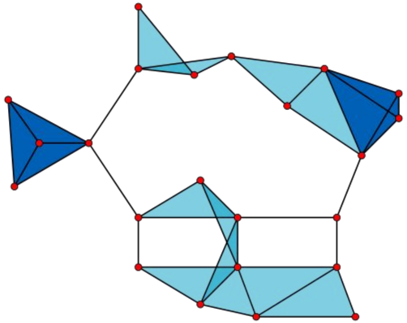
https://studentwork.prattsi.org/infovis/labs/visualizing-ingredient-networks/ browse for visualizarions and project ideas

Subgraphs, cliques and k-cores

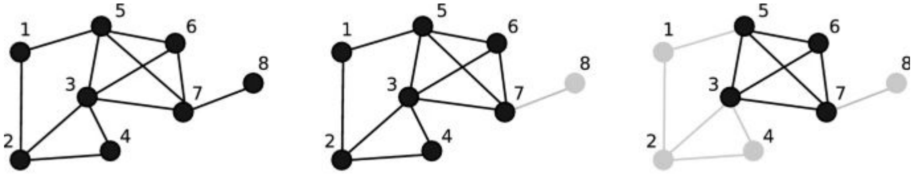
- Induced subgraph:
  - Edges between a subset of nodes in the Graph



- Clique: a.k.a. complete subgraphs
  - A subgraph where every two nodes are adjacent
    - How many 4-vertex cliques?



- K-core:
  - Maximal subgraph where degree of each node is at least k



(a) 1-core (b) 2-core (c) 3-core

Graphlets & Motifs

- Graphlets
  - small, connected, and non-isomorphic induced subgraphs
- Motifs
  - Statistically over- or under-represented graphlets