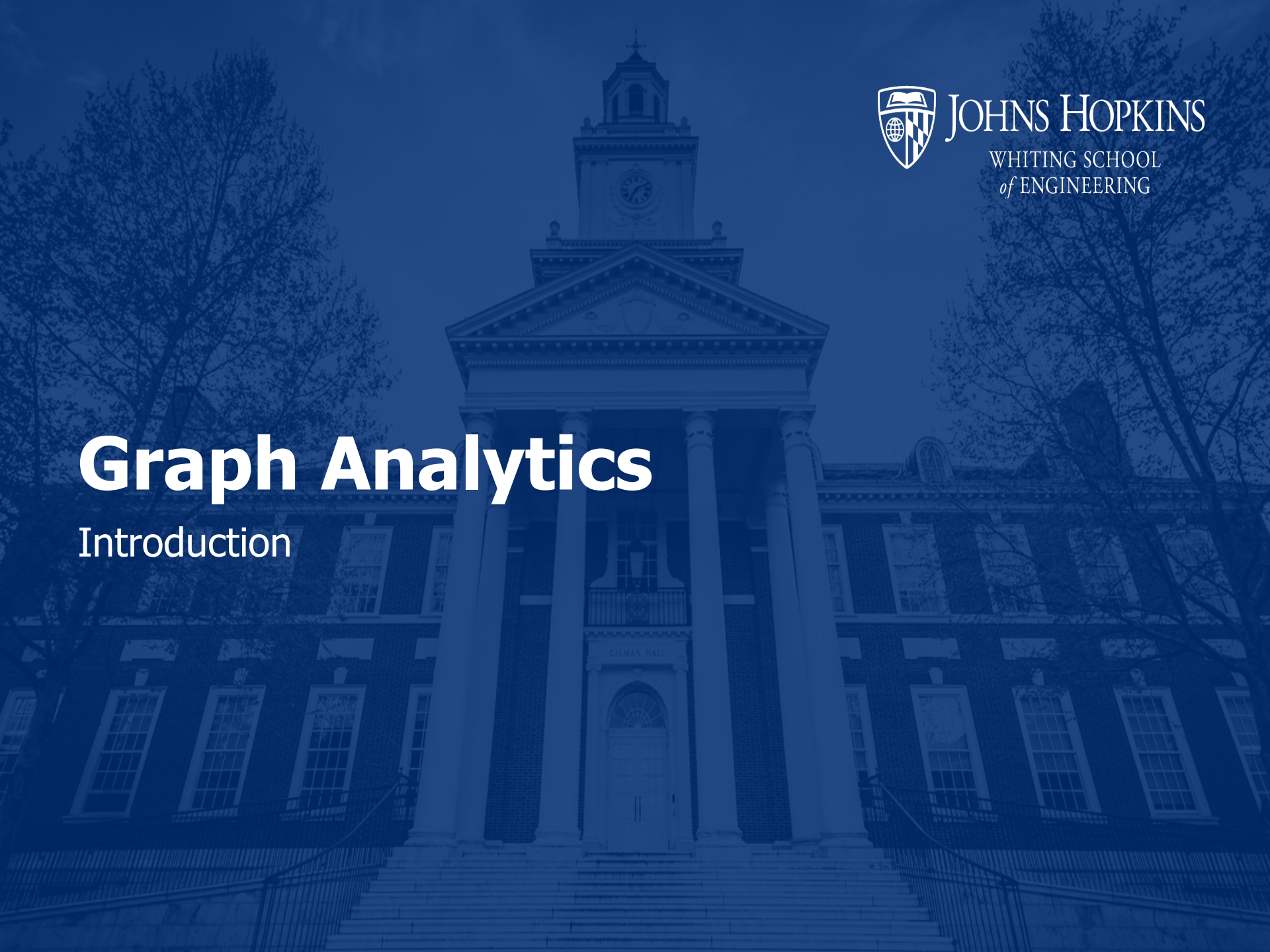


Graph Analytics

Introduction

Topics

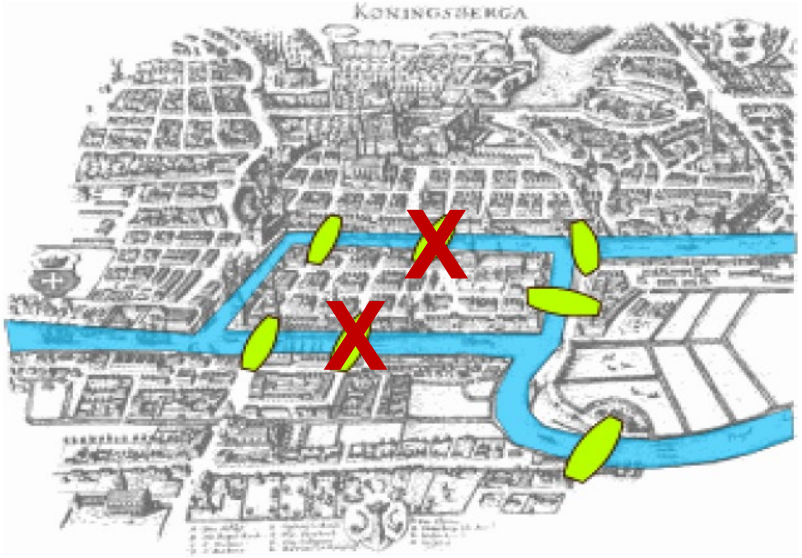
Topics for Today:

- **Introduction to graphs and applications**
- **Detailed plan of the course and what you can expect to get out of it**
- **Orientation to software tools for working with graphs**

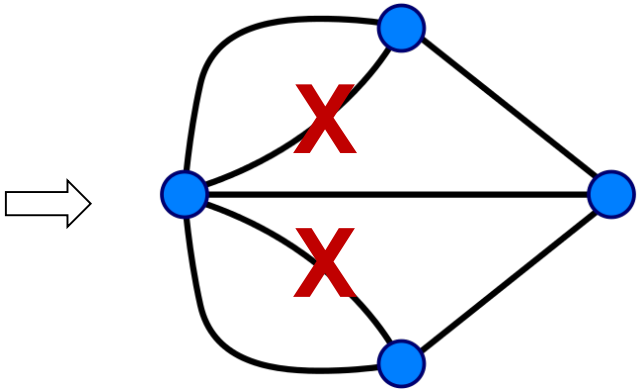
Graphs

Eulerian Path: Can you cross all bridges without crossing any bridge more than once ?

Eulerian Circuit: An Eulerian path where start and end points are the same.



Seven Bridges of Königsberg
[Euler, 1735]



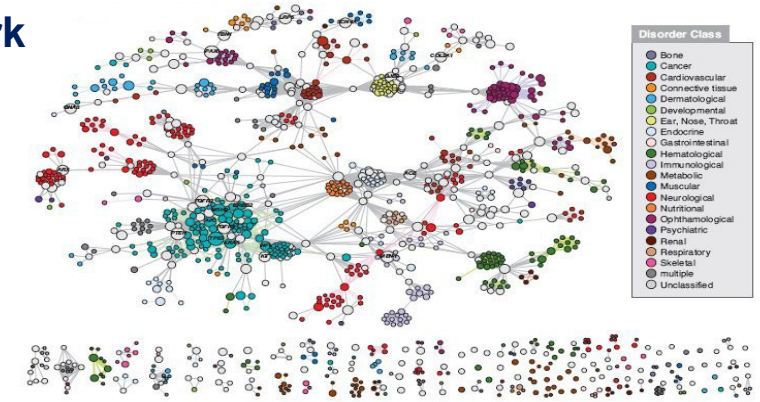
Solution:

- Nodes except start and end points must have an even number of bridges
- No more than 2 nodes (start & end) can have odd degree
- After 2nd World War bombings only 5 bridges are left and only 2 nodes have odd degree. Unfortunately they are both islands !

Networks

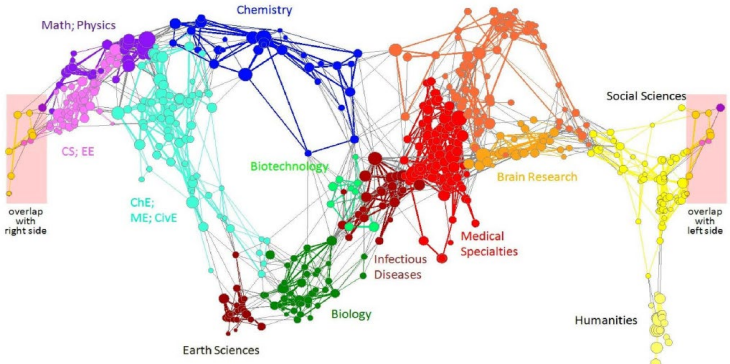
Social Networks

Disease Gene Network



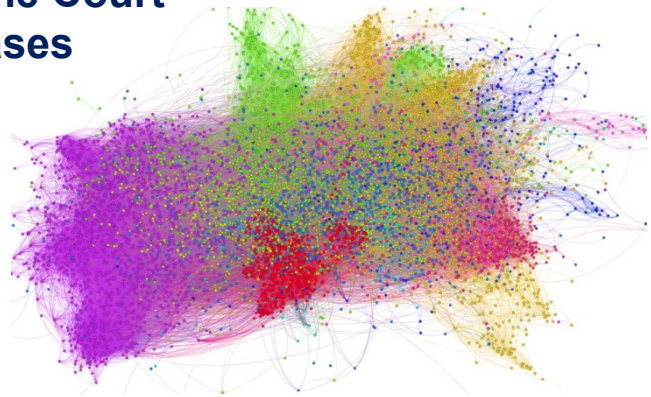
https://www.nature.com/scitable/topicpage/genome-wide-association-studies-and-human-disease-788#

Scientific Literature



Citation networks and Maps of Science
Borner et al, 2012

Supreme Court Cases



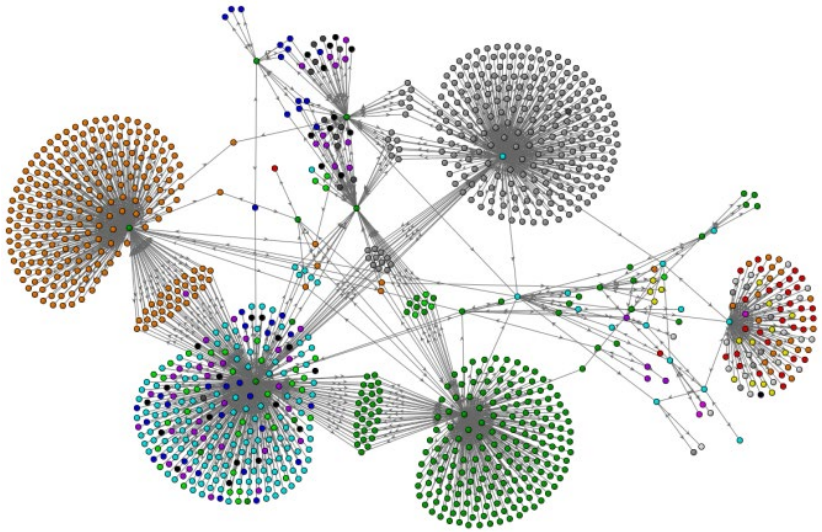
http://visualfa.org/citation-network-maps/network-maps-a-guided-tour/

Cyber & Social Networks

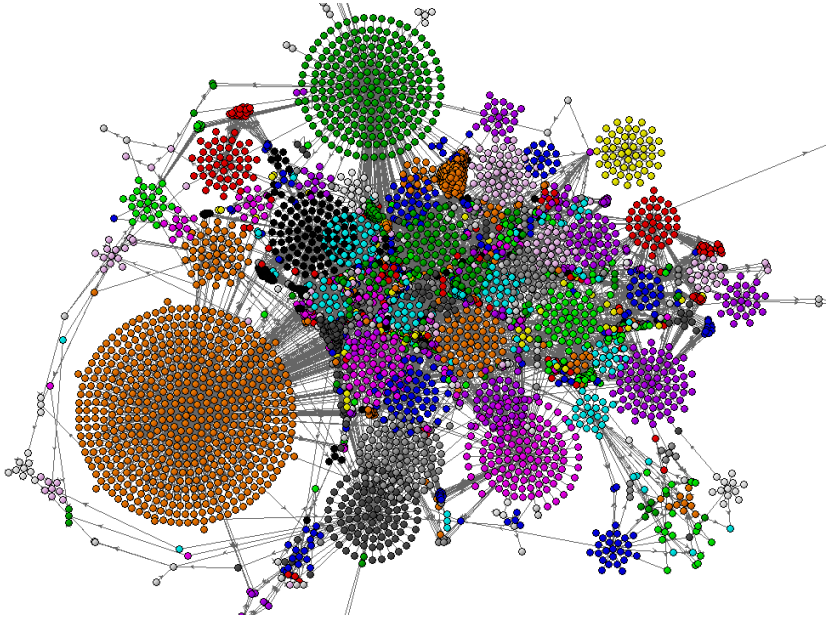
Problems:

- Leaders, power brokers ?
- Nodes critical for spreading of information?
- Nodes critical for defending against spread of virus.
- Communities ?

Cyber Network:



Social Network:

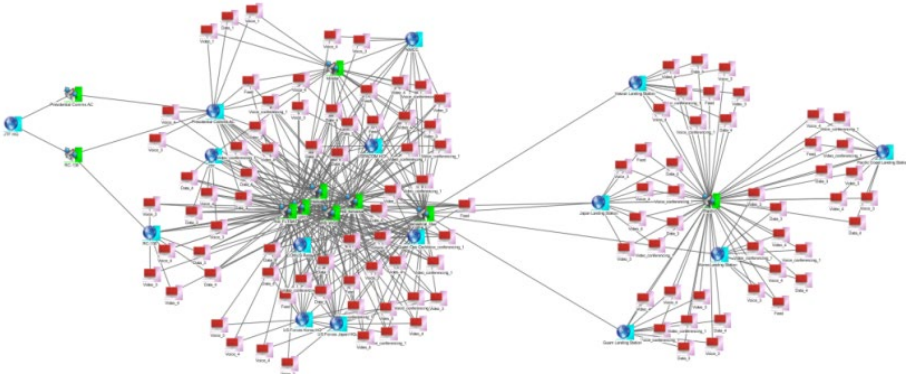


Communication & Power Networks

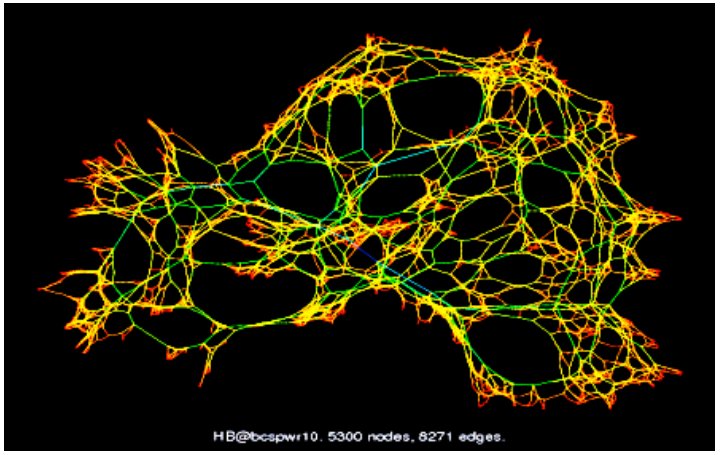
Problems:

- Nodes critical for maintaining connectivity
- Resiliency of a network

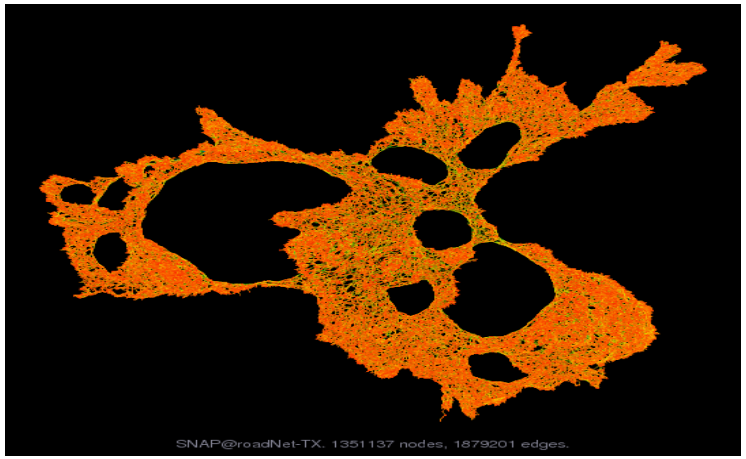
Communication Network:



US Power Network:



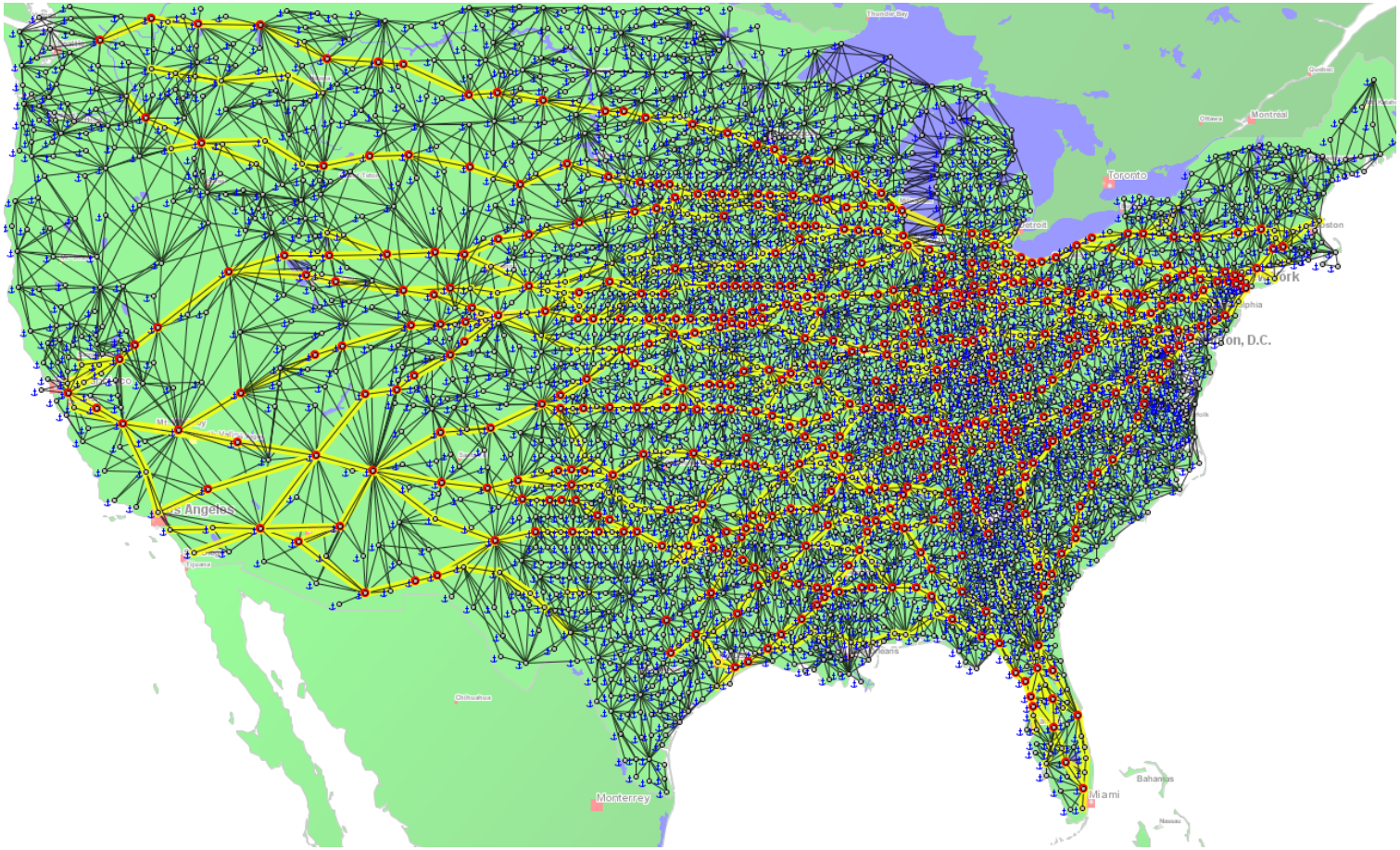
Texas Road Network:



Graphs in Land Networks

Problems:

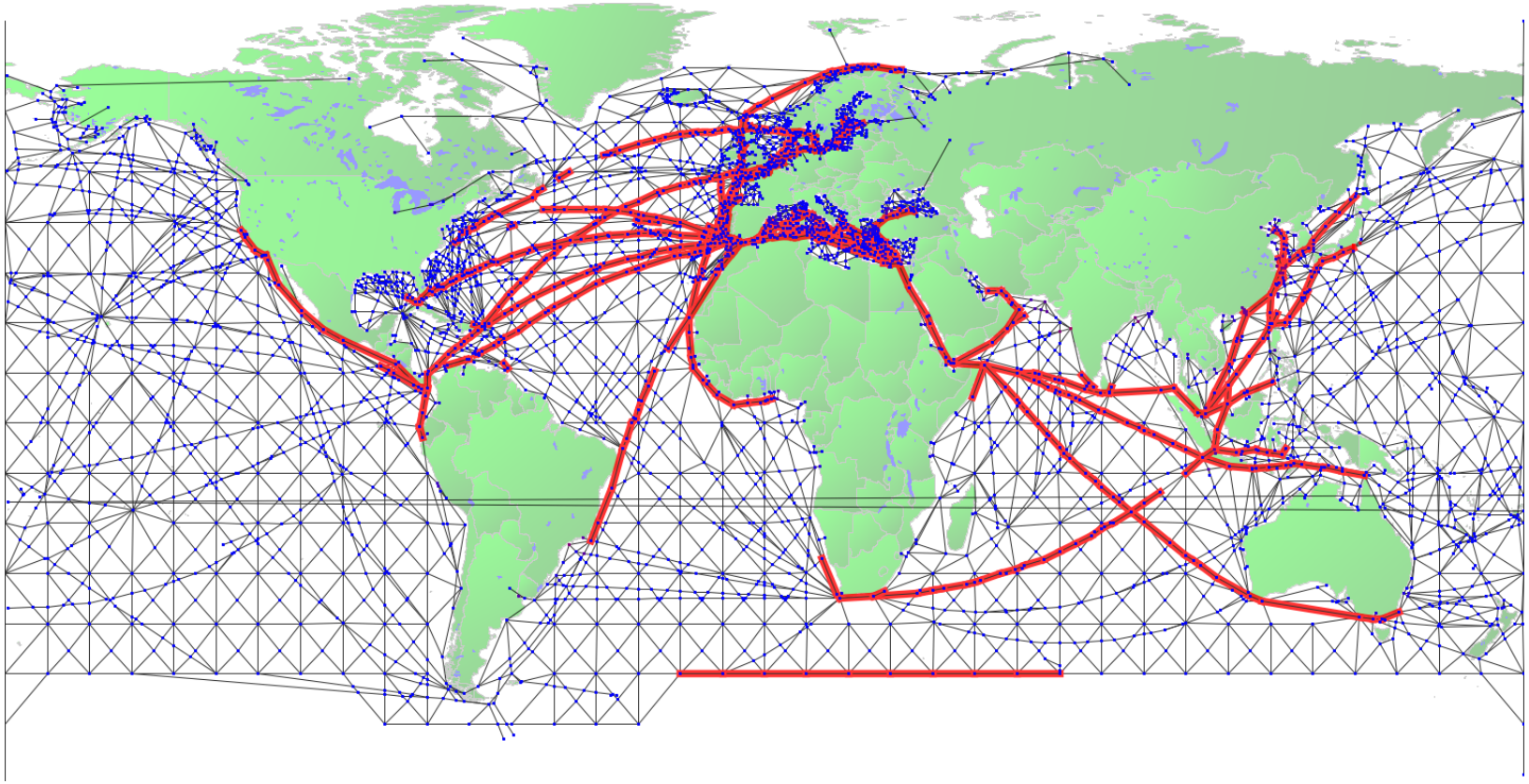
- **Analysis of flow, contagion: Using population statistics, network structure, and centrality metrics.**



Graphs in Maritime Logistics

Problems:

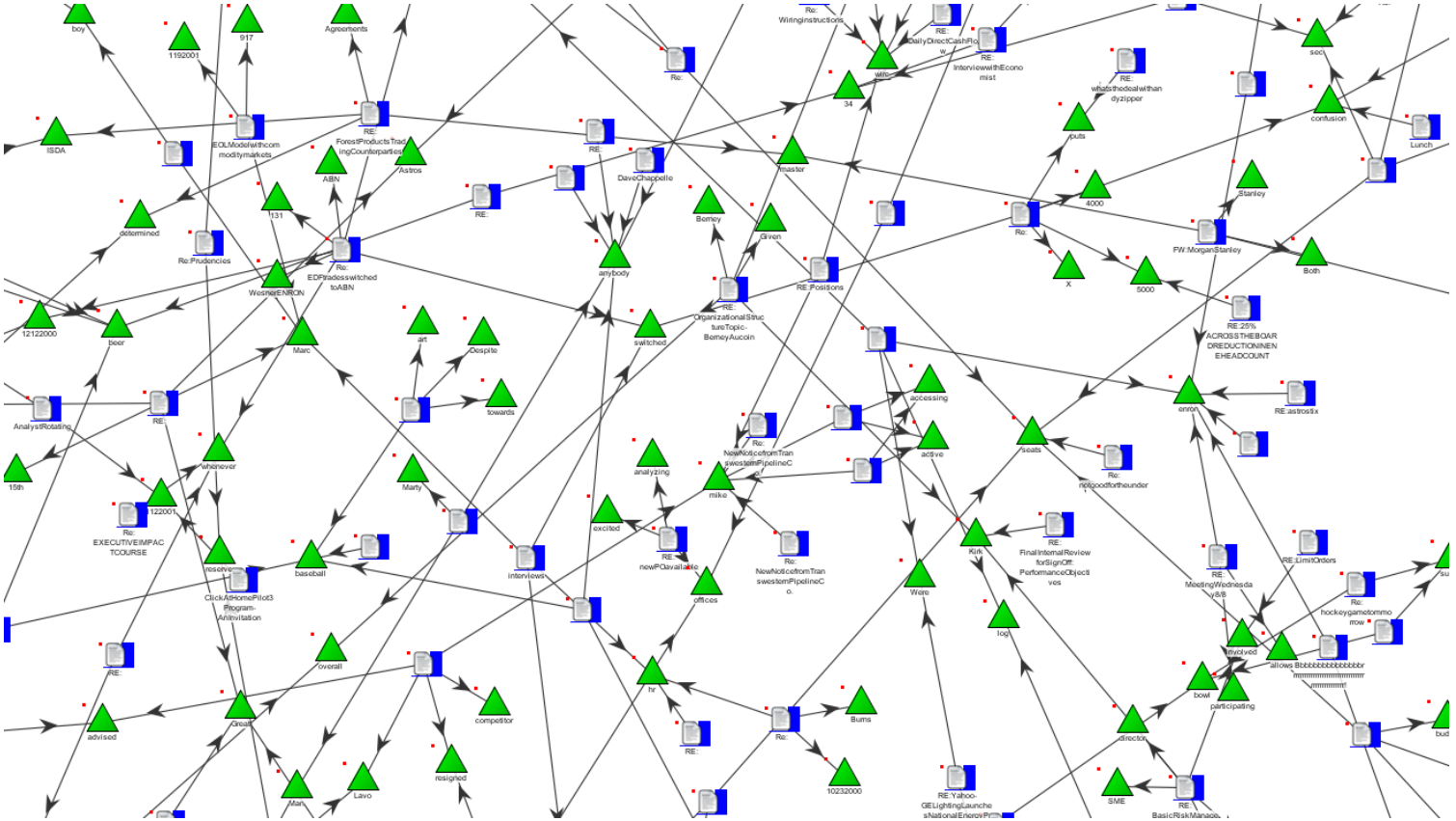
- Determination of high centrality routes
- Determining impact of closure of sea lanes (Suez canal for example)



Graphs in Unstructured Text Analysis

Problems:

- **Inferring relationships between documents, communication**
- **Resolving identity of an actor in different data set**

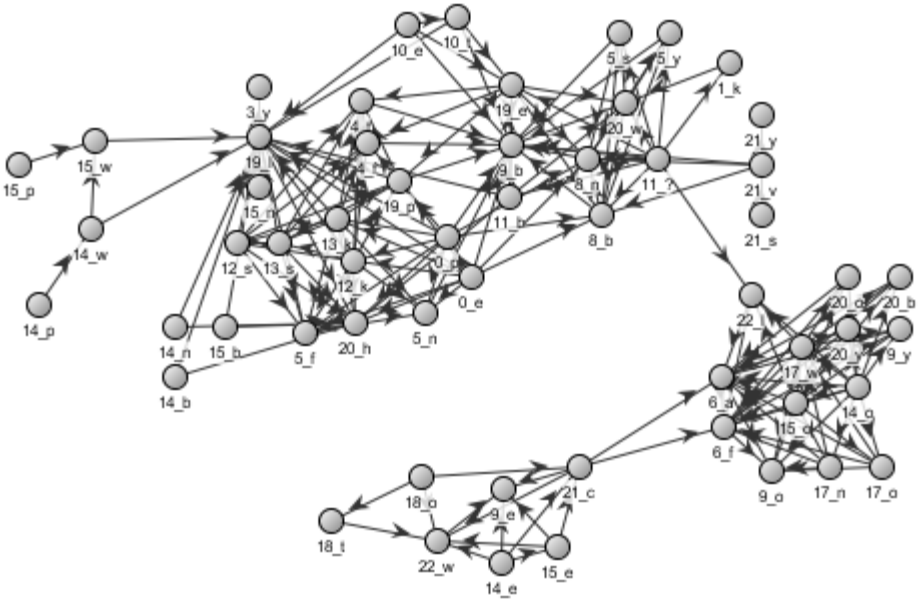


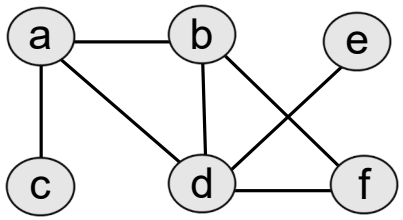
Graphs in High Dimensional Probability Models

Problems:

- **How can we build a probability distribution for an attribute rich data set ?**

- **Answer lies in structure of relationships in attribute correlations**





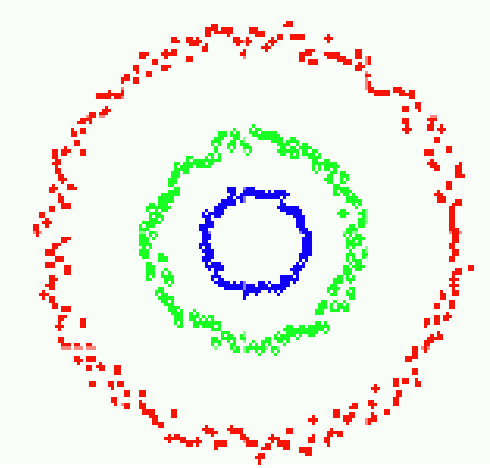
$$P(a,b,c,d,e,f) = \frac{P(a,c)P(a,b,d)P(b,d,f)P(e,d)}{P(a)P(b,d)P(d)}$$

Graphs in Discovering Topological Clusterings

Clustering of data that has complex topological structure is often challenging because there are no clear boundaries and clusters do not represent dense regions.

Finding clusters for these types of problems requires an understanding of connectivity of points and connectivity is very much a "graph problem"

No simple boundary that separates clusters



Clusters are not necessarily dense regions

Spectral methods that we will talk about can be used to solve such complex clustering problems by mapping them on to graph partitioning problem.

Graphs in High Dimensional Correlation Analysis

Problems:

- Determination of highly correlated gene clusters

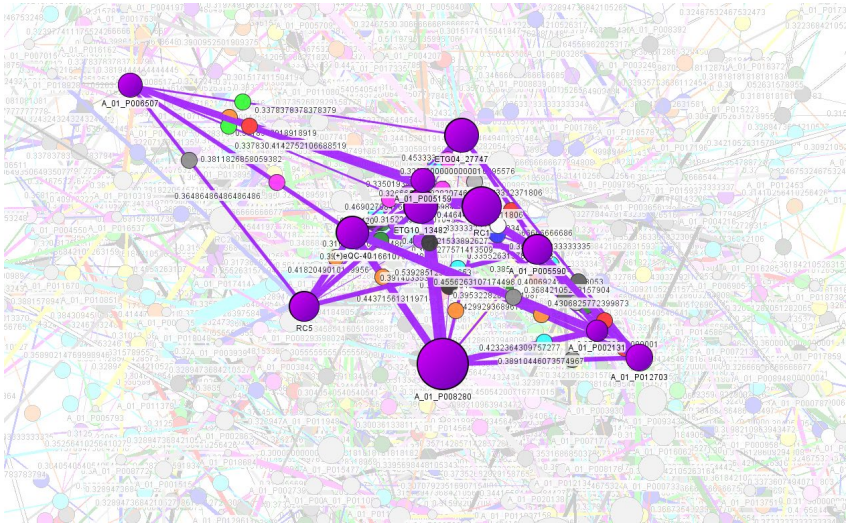
Individual x

~20,000 attributes

A_01_P00	-1.35987	-1.52555	-0.92857	-1.1348	-0.9825	-1.5151	-1.4506	-1.70545	-0.99555	-1.83013	-2.00012	-2.0591
A_01_P00	2.15291	0.134629	0.110145	0.196639	-0.63537	-0.91269	-0.99328	-1.14591	-0.62431	-0.2948	-0.93138	-0.39628
A_01_P01	3.738447	3.903645	4.413672	4.842012	4.722735	4.205534	4.398612	4.842676	2.53738	4.025587	3.279281	0.349596
A_01_P00	-0.83936	0.440726	0.571087	0.919746	0.442425	0.337388	0.056067	-0.35081	-0.1381	-0.26245	0.144998	-0.02933
A_01_P00	-1.09223	-0.60996	-0.6629	-0.66946	0.188534	-0.35304	1.094064	-0.7313	-0.06844	-0.21758	-0.0939	0.083113
A_01_P01	1.303442	0.94778	0.251063	0.46765	1.220471	0.970915	1.329262	0.876055	0.748621	0.93433	1.122345	0.967031
A_01_P00	-4.52012	-4.3658	-3.97049	-3.70677	-3.57729	-3.91595	-4.54643	-4.224	-4.05446	-3.89812	-4.4134	-3.00194
A_01_P00	-0.4995	-1.19408	-1.61685	-1.39306	0.135696	-0.59302	-0.0672	-1.15481	-1.69117	-0.30605	-0.44107	-0.54338
A_01_P00	0.493103	0.971945	1.261047	1.500173	1.000508	1.288398	0.326134	1.732595	2.950797	1.055105	1.52478	2.030461
A_01_P01	2.562247	2.992177	3.012286	3.567378	2.811593	2.595047	4.001796	3.943029	4.33555	1.6303	2.721643	2.89847
A_01_P00	0.098297	0.300304	0.280796	-0.10335	0.821121	0.536207	0.56873	0.789619	1.253181	0.627159	1.217073	1.285887
A_01_P00	0.841885	0.705344	0.778045	0.857809	0.872313	0.635193	0.543369	1.031077	0.548762	1.281323	0.650392	1.552624
A_01_P00	0.064046	0.969997	0.407577	0.92618	1.465986	1.337728	-0.60284	-0.00947	-0.20148	0.048855	0.375238	1.415226
A_01_P00	1.473942	0.184218	0.503189	0.382367	1.035612	-0.43869	1.233864	1.308009	0.314015	0.678408	0.720136	0.214763
A_01_P01	0.05485	1.110387	-0.10845	0.367485	0.870483	0.6913	0.377316	0.508018	0.40821	0.416664	1.187635	-0.02858
A_01_P01	0.154476	-1.26831	-1.0759	0.268849	-0.90477	-1.40298	-0.89749	-1.92523	-2.21591	-1.49256	-1.88889	-2.15686
A_01_P01	1.512725	1.080667	0.658108	-0.26278	0.884653	-0.56233	0.465169	0.516965	1.498214	1.715135	0.683246	1.521567
A_01_P01	-1.20148	-1.13265	-1.1326	0.710732	0.22051	-0.27244	-2.23788	-1.83116	-1.34426	-0.14557	-0.96512	-1.55381
A_01_P00	-0.42071	0.215487	0.679026	0.279455	-0.33898	-0.06006	-0.22242	1.236499	1.960618	0.506195	0.391243	0.282076

...

Correlation Graph with 100+ Million Links!



Types of analysis involving graphs

General areas of graph analytics:

- **Modeling graphs that capture properties of real world graph features**
- **Importance/centrality of graph nodes and links**
- **Identification of communities/groups/clusters**
- **Inferring links using evidence of indirect connections**
- **Measuring global properties of graphs (such as resiliency against attacks)**
- **Understanding processes (such as contagion) on graphs**
- **Mathematical methods that rely on graphs (such as probabilistic graphical models)**

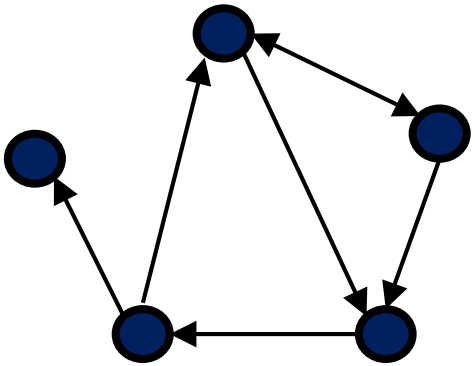
Methods used:

- **Linear algebra**
- **Probability and Statistics**
- **Graph algorithms (such as shortest paths, depth first, bread first search)**

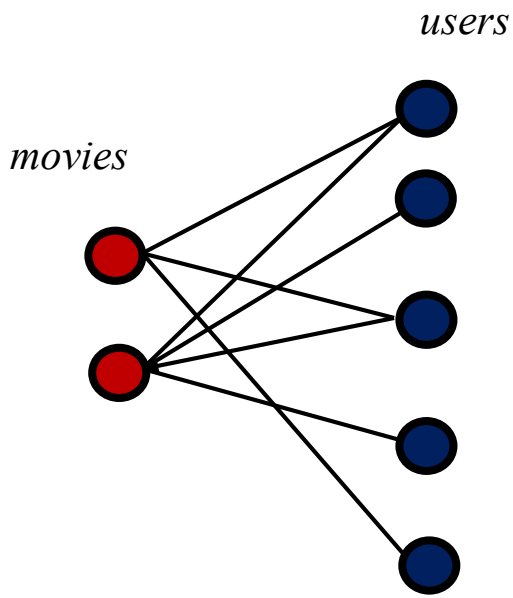
STRUCTURE

Types of Networks

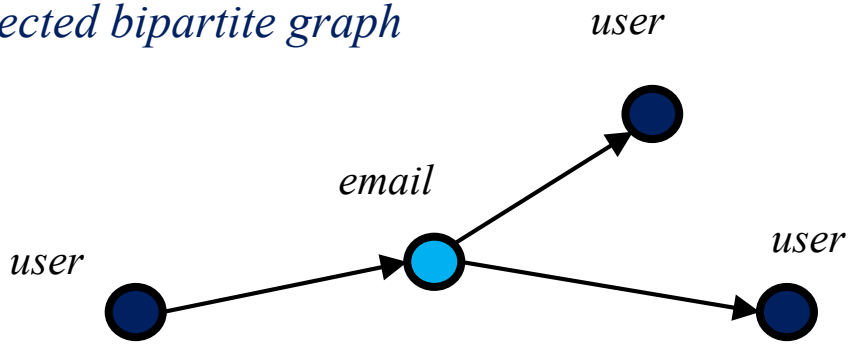
Directed or undirected
Graph of homogeneous nodes



Graph of nodes belonging to different types: bipartite graph



A directed bipartite graph



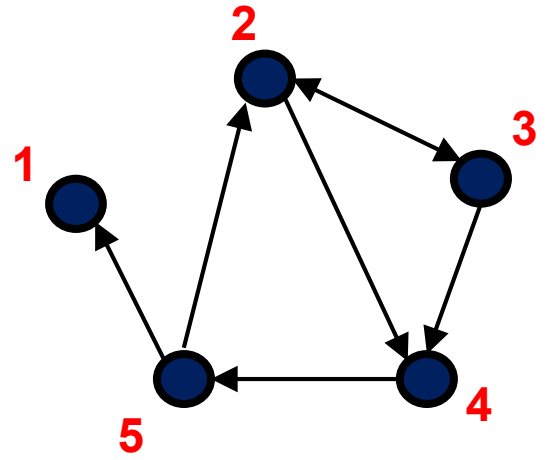
Data Structure for Connectivity

Adjacency List

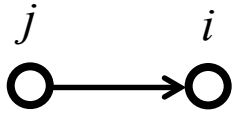
- *1: (), (5)*
- *2: (3,4), (5)*
- *3: (2,4), (2)*
- *4: (5), (2,3)*
- *5: (1,2), (4)*

Fast look up of neighbors

More work to figure out if 2 nodes are connected



A_{ij} :



Edge List

- *5,1*
- *5,2*
- *4,5*
- *2,4*
- *2,3*
- *3,2*
- *3,4*

Easy to check if 2 nodes are connected

Natural fit for a sparse matrix representation

More work to look up neighbors

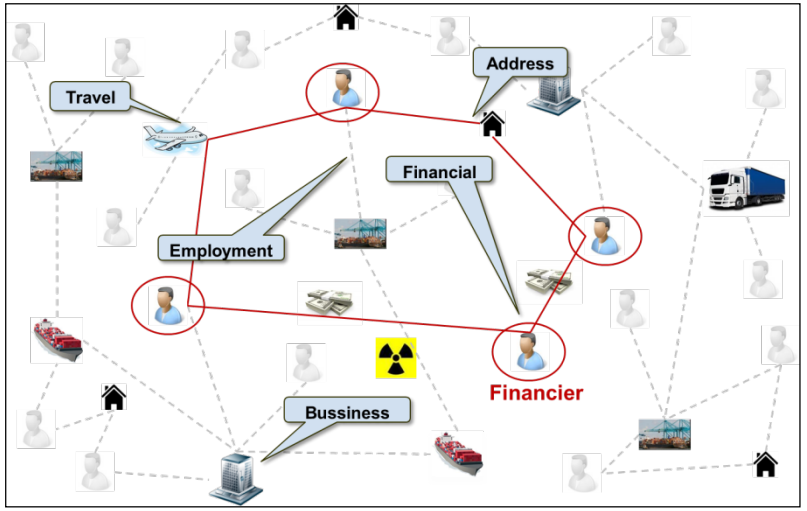
A =

	1	2	3	4	5
1	0	0	0	0	1
2	0	0	1	0	1
3	0	1	0	0	0
4	0	1	1	0	0
5	0	0	0	1	0

Property Graphs and their role in Large Scale Data Management & Analysis

Graph as a Data Structure:

- A flexible structure with power in expressing relationships
- Facilitates integration of new data without having to deal with complex schema
- Facilitates enriching data with analytics during and after data ingest- *Question Focused Data (QFD)*

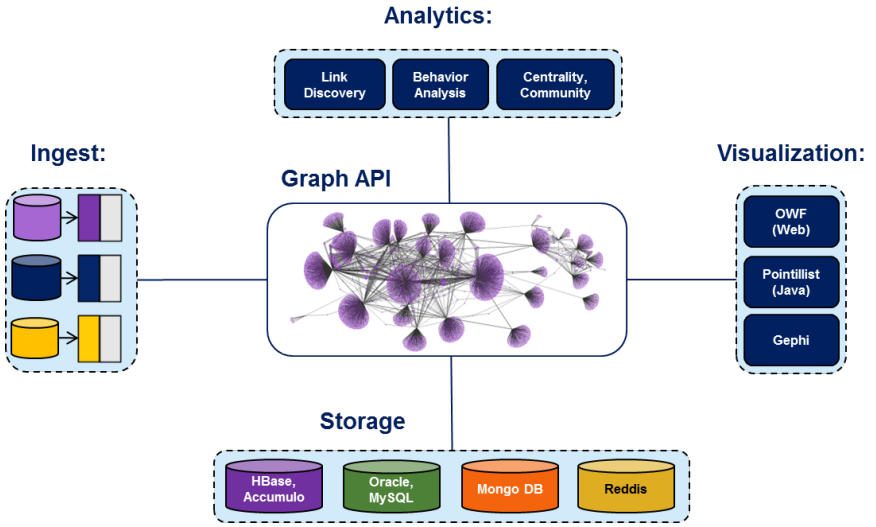


Graph as an Interface:

De-couple Apps from underlying structure of data

A standard interface that unifies access to non-relational as well as relational data bases

Allows higher level analysis



Subjectivity of Property Graph Structure

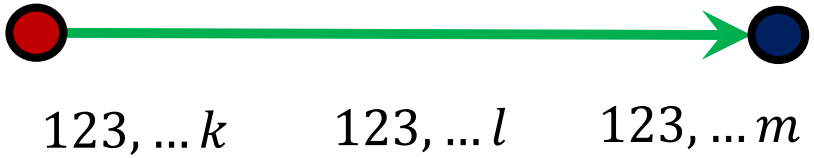
One of the underappreciated aspects of Graph analytics is how subjective graph structure is and how many different graphs can be constructed from the same data set. Graph structure is analogous to "schema" in a relational database.

Consider an n dimensional data (assume categorical attributes). And assume each record is a node-link-node triplet. Nodes and links can be defined by (k,l,m) combinations of these attributes .

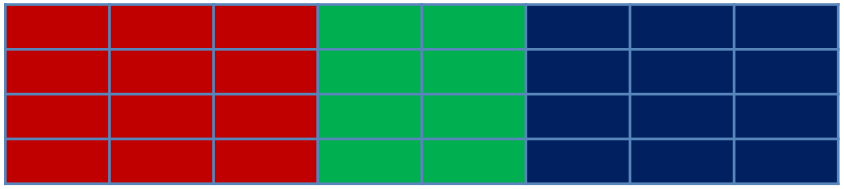
Number of ways to define a graph structure for a given set of k,l,m

$$\frac{n!}{k!\,l!\,m!}$$

$$n = k + l + m$$



Total number of ways to define a graph structure for attribute rich data can grow rapidly!



Human understanding of the data often reduces these options to a handful of options.

CLASS PLAN

Topics

Introduction (M1)

- *Introduction to Graphs & Applications*
- *Working with graphs using Java and Python [*]*
- *Graph query and visualization [*]*

Graph Matrices and Linear Algebra (M2)

- *Matrix properties, eigenvalue equation*
- *Adjacency and Laplacian Matrices and their properties*

Statistical properties (M3)

- *Conditional Probability, distributions, Bayesian analysis*
- *Degree distribution, Clustering coefficient, Assortativity*

Random Graphs (M4)

- *Configuration model, Erdos Renyi,*
- *Giant component, Lattice percolation*

Random Graphs (M5)

- *Barabasi Albert, Watts-Strogatz*

Graph Clustering: Identifying tight-knit groups (M6)

- *Components, Clans, Cores, Plexes, Cliques, Trusses, Trapezes*

[*] Optional material

Centrality: Measuring importance of graph elements (M7)

- *Distance, Bonacich, Eigenvector, Page Rank, Hub-Authority*
- *Link Cohesion, Betweenness, Weighted Betweenness,*

Graphs in High Dimensional analysis (M8)

- *Correlation and Mutual Information graphs, Clique Tree,*
- *High dimensional indexing and clustering*

Link Inference (M9)

- *Inferring links using structural statistics*
- *Recommendation systems*

Partitioning (M10 & M11)

- *Kernighan-Lin, Girvan-Newman*
- *Modularity, Peer pressure, Markov, Spectral*
- *Spectral Clustering*

Resiliency (M12)

- *Graph Resistance, Bandwidth, Flow*
- *Number of Spanning Trees*

Dynamic Processes: (M13)

- *Opinion formation*
- *Diffusion*

Final Project

Final project ideas:

- **Analysis of a data set of interest to you using graph analysis algorithms**
- **Implementation and presentation of a graph analysis algorithm (preferably one that is not covered in class)**
- **Implementation and analysis of a random graph generation algorithm not covered in class**

Data Sets & Reference

Your class website provides some example graphs we will use in graphml and graphson formats as well as raw data that was used to create them.

Some useful sites (among many) with graph data sets:

- **http://snap.stanford.edu/data/**
- **http://networkdata.ics.uci.edu/**
- **http://www.cise.ufl.edu/research/sparse/matrices/**

Some useful references (among many others):

- **Textbook: "Networks, An Introduction". M. Newman**
- **"Networks, Crowds, and Markets", J. Kleinberg**
- **"Graph Algorithms in the Language of Linear Algebra" J. Kepner, J. Gilbert**
- **"Graph Spectra for Complex Networks" Piet Van Mieghem**