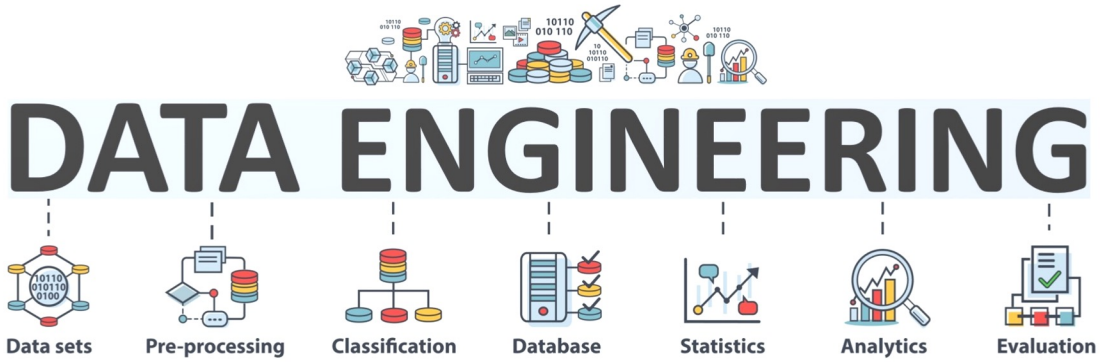


Lecture 8: Data Driven Applications (2) --

Social Network Analysis and Anomaly Detection

CS5481 Data Engineering

Instructor: Yifan Zhang

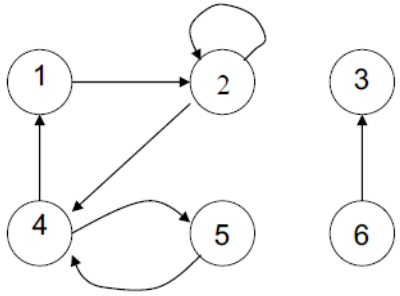
Outline

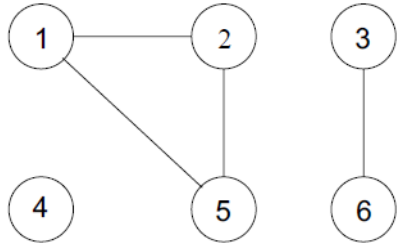
1. Social network analysis
2. Anomaly detection

Networks

What is a Network?

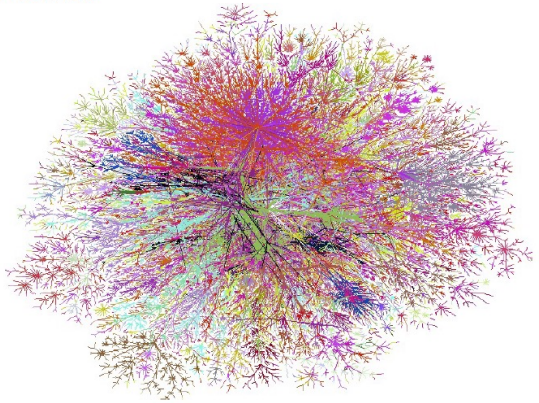
- Network = Graph
- Informally a graph is a set of nodes joined by a set of lines or arrows.



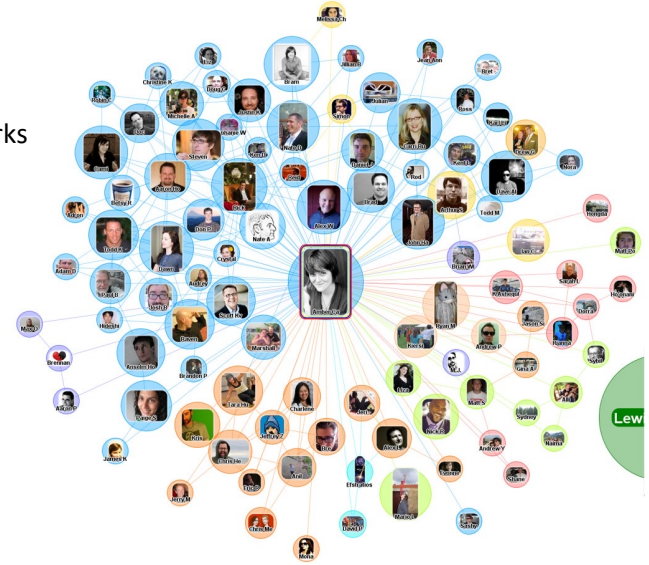


Real-World Networks

- The Internet



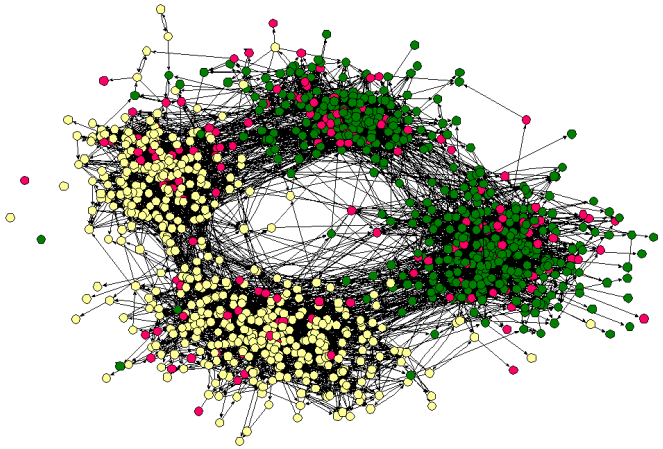
- Social networks



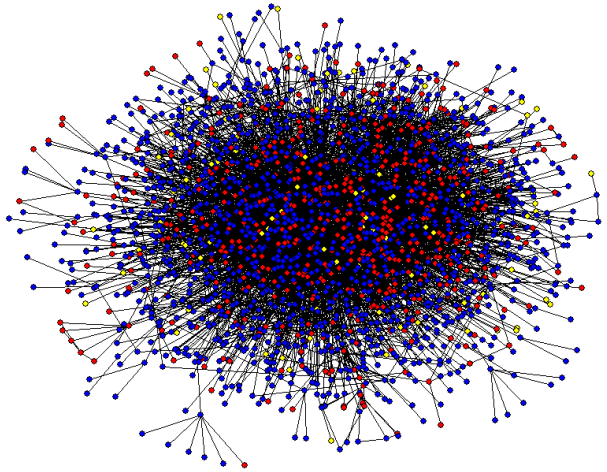
- Transportation networks

More real-world networks

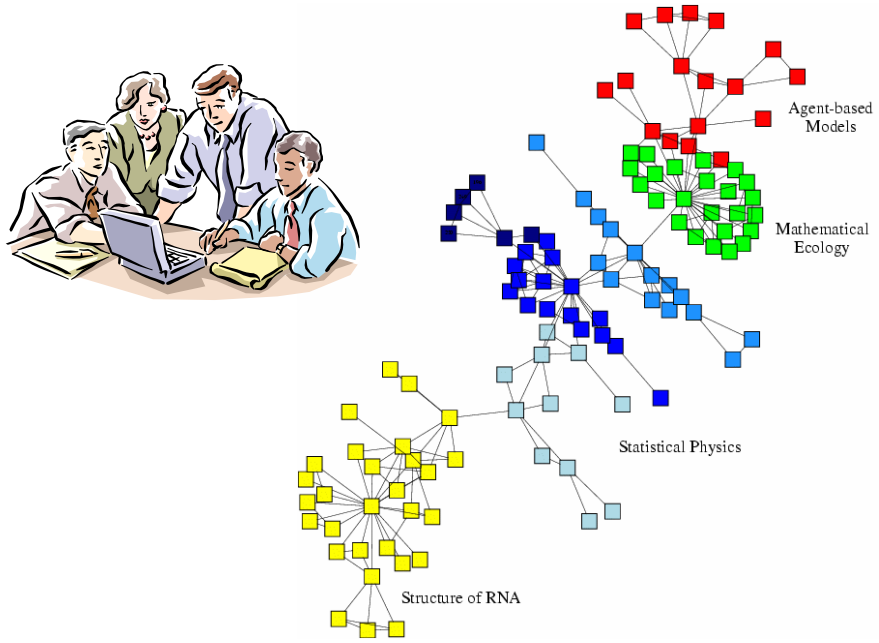
- Friendship network



- Protein-protein interaction networks



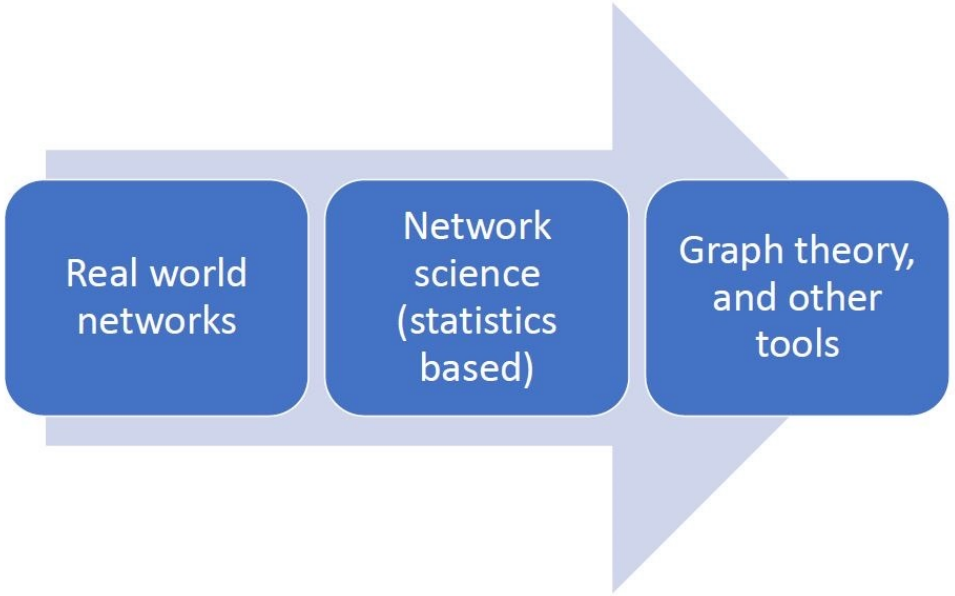
- Scientific collaboration network



Social network analysis – graph theory

Graph theory is a branch of discrete mathematics concerned with proving theorems and developing algorithms for arbitrary graphs (e.g. random graphs, lattices, hierarchies).

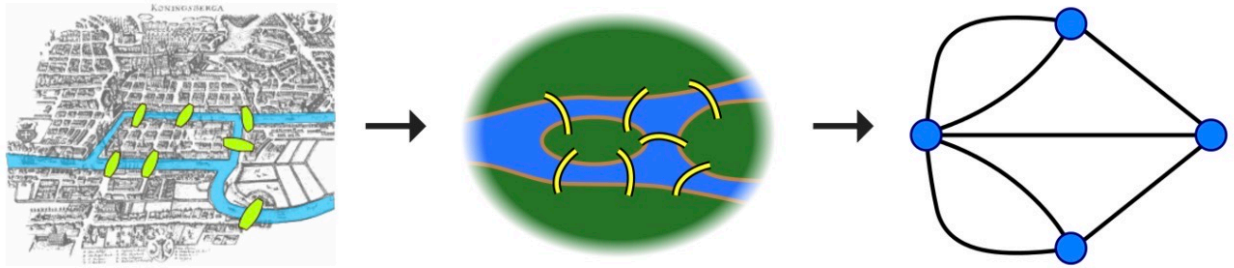
Graph theory and network science



Social network analysis – graph

What makes a problem graph-like?

- There are two components to a graph
  - Nodes and edges
- In graph-like problems, these components have natural correspondences to problem elements
  - Entities are nodes and interactions between entities are edges
- Most complex systems are graph-like



Map of Konigsberg in Euler's time showing the actual layout of the seven bridges, highlighting the river Pregel and the bridges. Graphs can be used to remove unnecessary (geographical) details, in order to highlight the connectivity patterns between different areas. Taken from Wikipedia.

Graph theory

G is an ordered triple G:=(V, E)

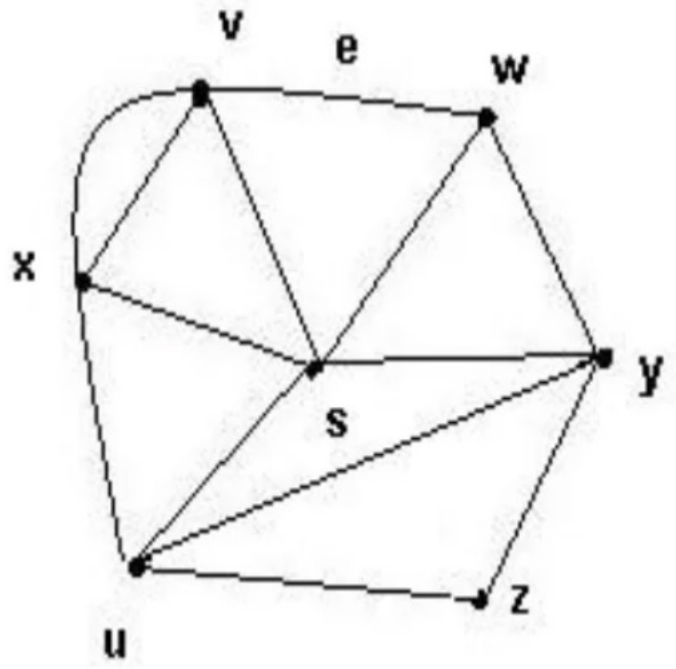
- V is a set of nodes, points, or vertices.
- E is a set, whose elements are known as edges, arcs, or lines.

Node (Vertex)

- Basic Element
- Drawn as a node or a dot.
- Vertex set of G is usually denoted by V(G), or V

Edge (Arc, link)

- A set of two elements
- Drawn as a line connecting two vertices, called end vertices, or endpoints.
- The edge set of G is usually denoted by E(G), or E.



- V = {s, u, v, w, x, y, z}
- E = {(x,s), $(x,v)_1$, $(x,v)_2$, (x,u), (v,w), (s,v), (s,u), (s,w), (s,y), (w,y), (u,y), (u,z),(y,z)}

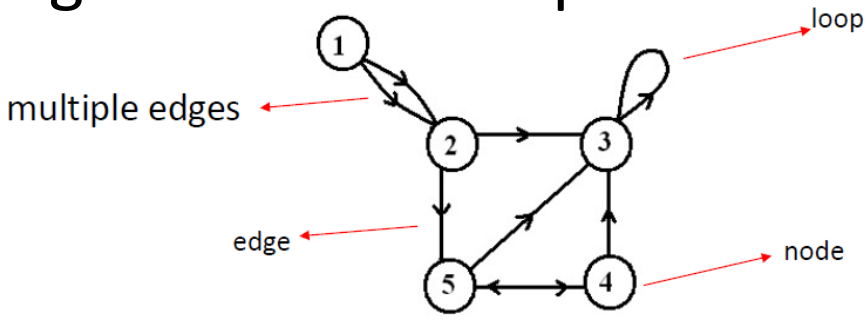
Undirected, directed, weighted networks

Undirected Network

In an undirected graph or network, the edges are reciprocal—so if A is connected to B, B is by definition connected to A.

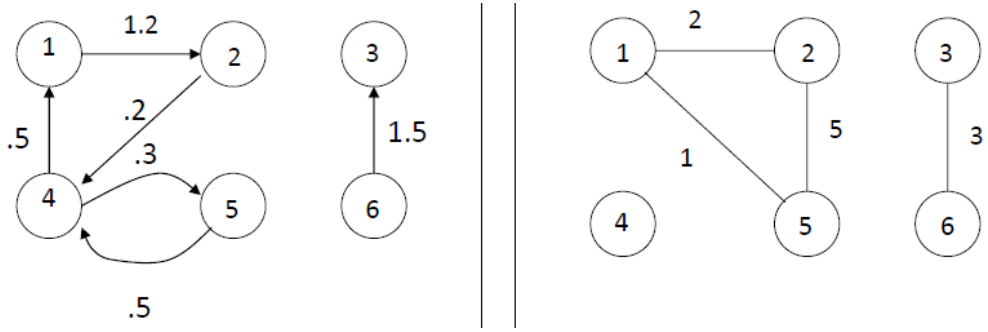
Directed Network

Edges have directions: an edge is an ordered pair of nodes



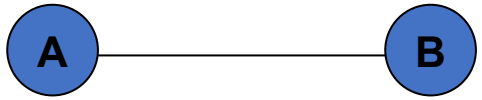
Weighted Network

Weighted network is a graph for which each edge has an associated weight, usually given by a weight function w: E->R.



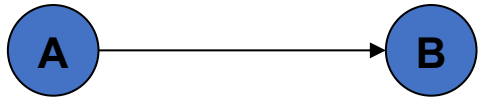
Examples of undirected, directed, weighted networks

With whom do you work?



= undirected relationship

Who do you ask for help/Who do you give advice?



= directed relationship

How often? Eg never, monthly, weekly, daily?

= strength of relationships

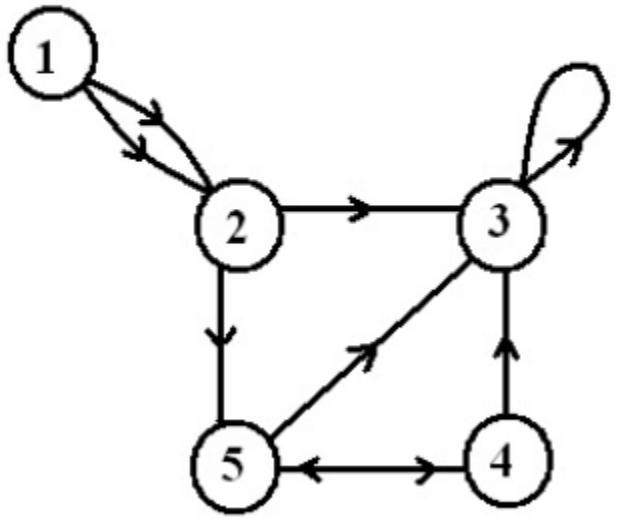
Q1: With whom do you work?

Q2: With whom do you have a drink?

= multiplexity of relationships

In- and out-degrees

- In-degree: number of edges entering.
- Out-degree: number of edges leaving.
- Degree = in-degree + out-degree



outdeg(1)=2
indeg(1)=0

outdeg(2)=2
indeg(2)=2

outdeg(3)=1
indeg(3)=4

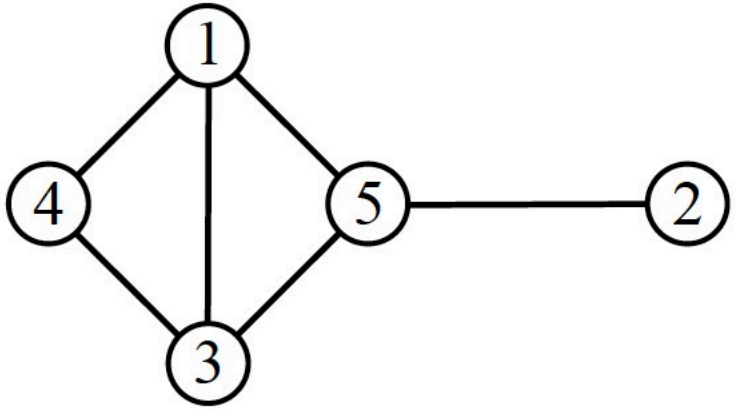
Graph representation

Adjacency matrix

Use a matrix to represent the graph vertices and edges

- An n*n matrix with n vertices.
- The (i,j) entry denotes the edges
  - Undirected/Directed: 1/0 denotes with/without an edge for vertices i and j.
  - Weighted: a positive value denotes the weight

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



Social network analysis

Social Network:

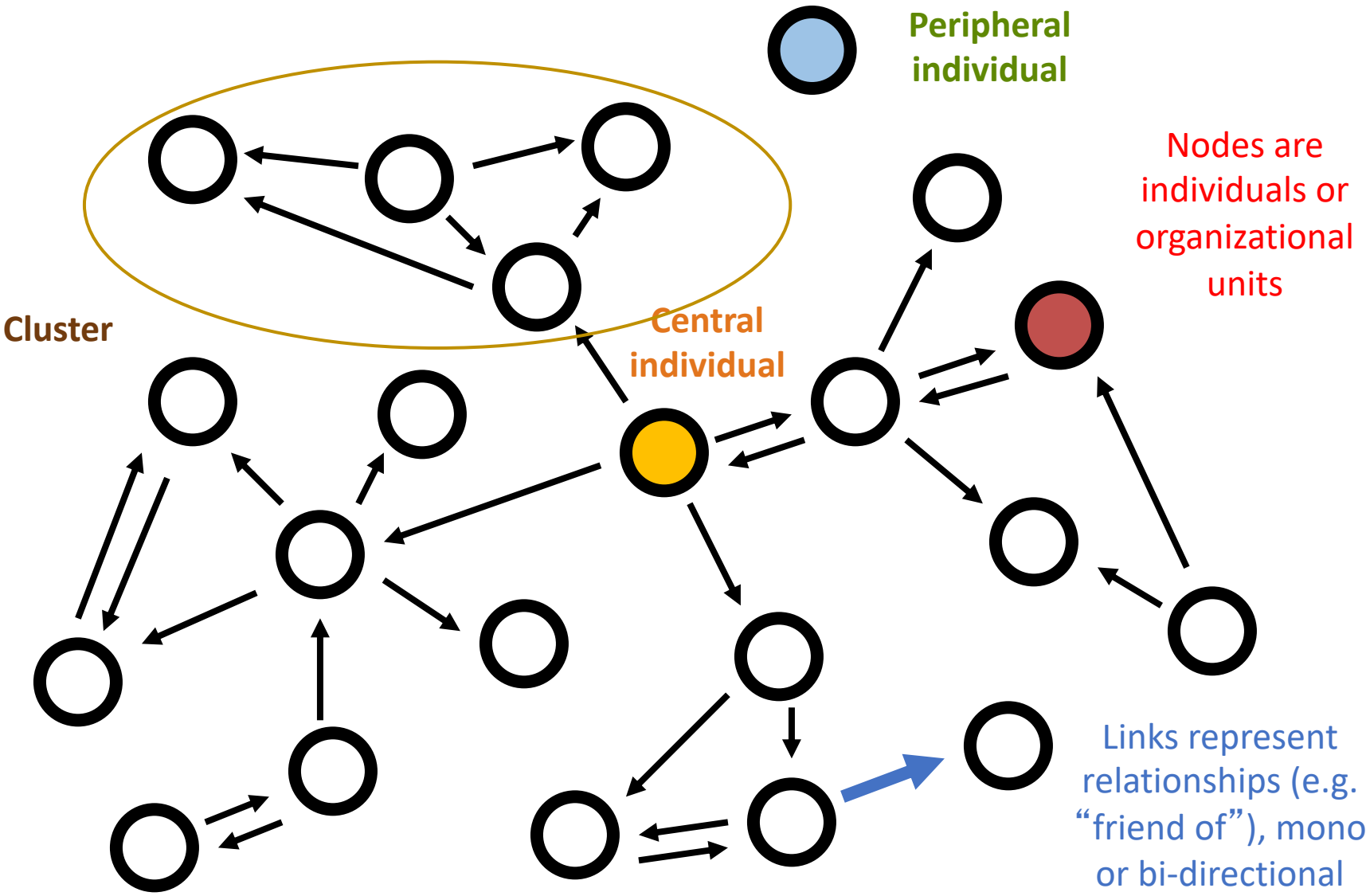
A set of individual actors (= people) connected through social relationships

Social Network Analysis (SNA):

is a diagnostic method for collecting and analysing data about the patterns of relationships among people in groups. A systematic approach to make the invisible flows within an organisation visible

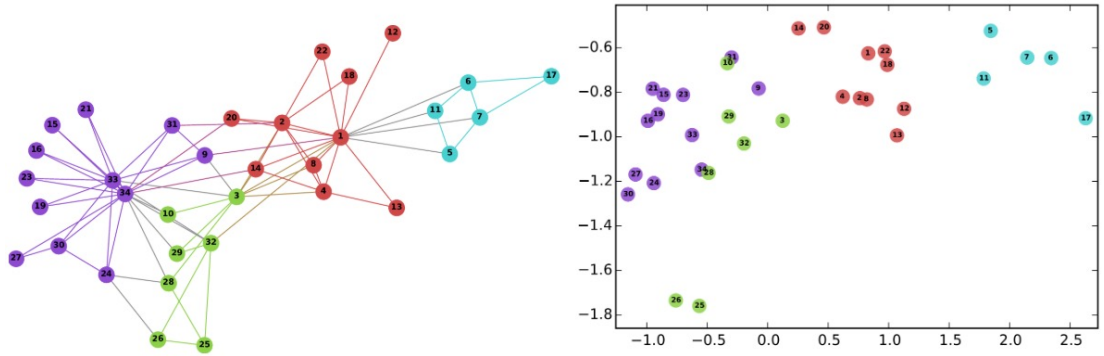
Sociograms

A **sociogram** is a graphic representation of social links that a person has. It is a graph drawing that plots the structure of interpersonal relations in a group situation.

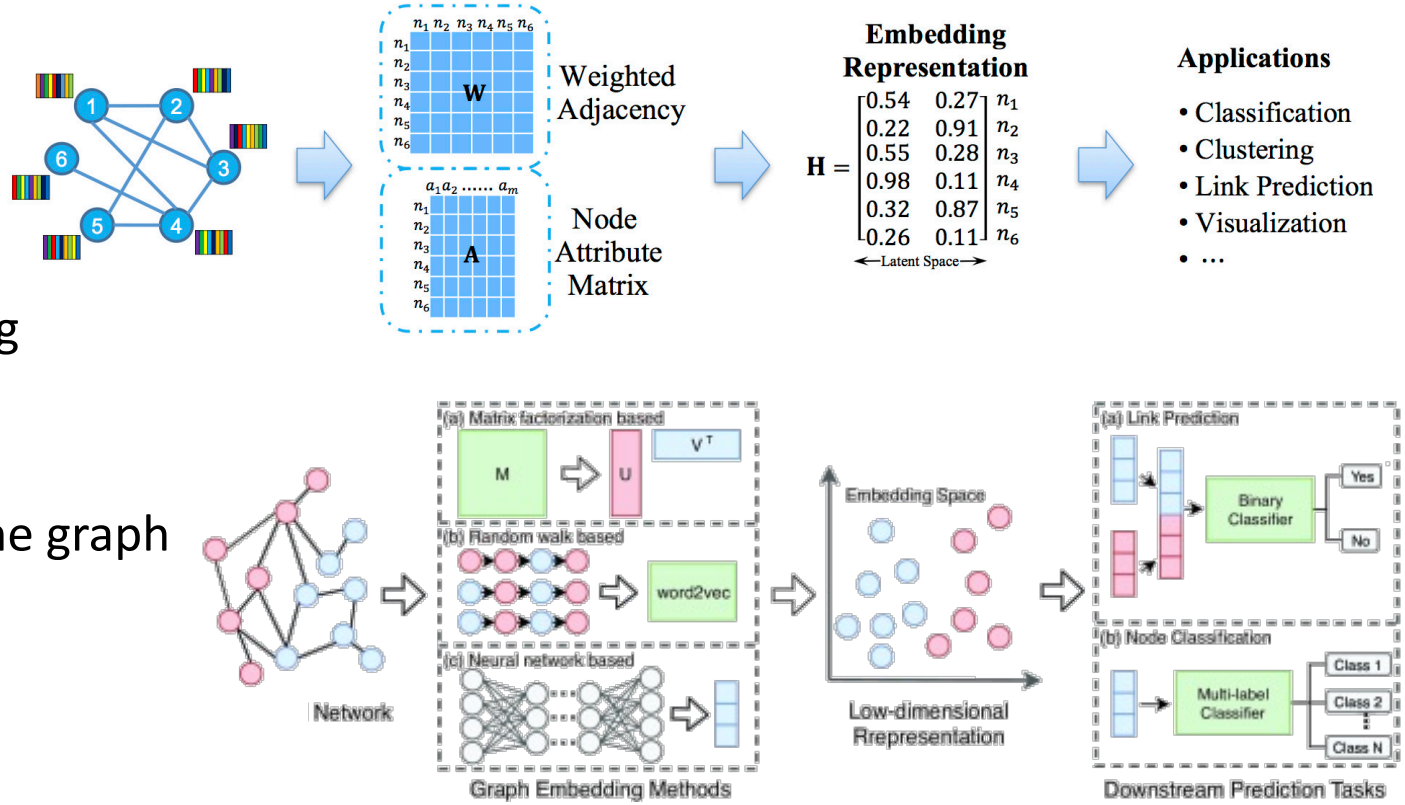


Network embedding

- **Network embedding:** mapping graph nodes to vectors of real numbers in a multidimensional space.

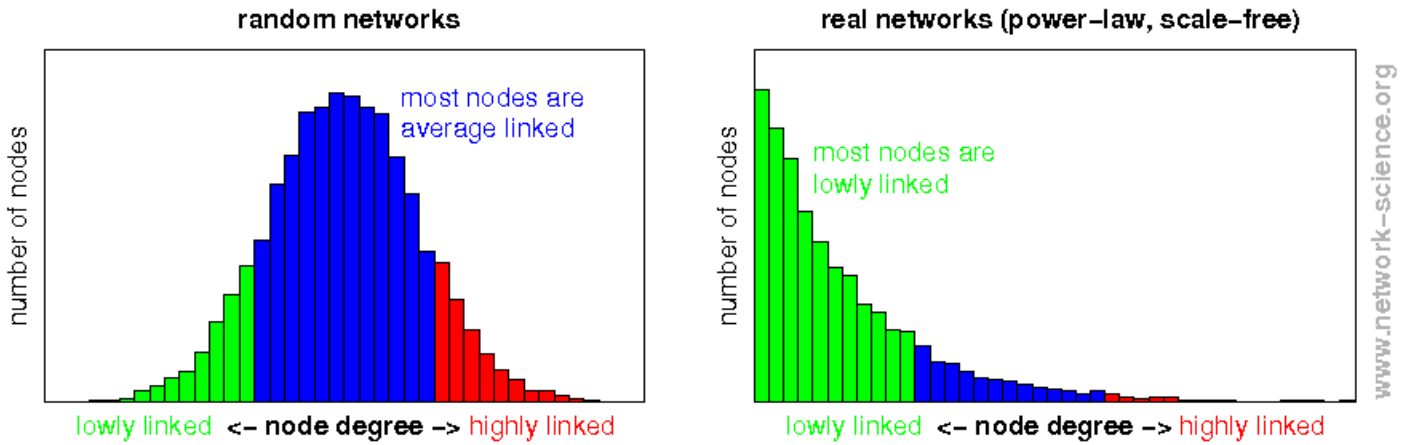


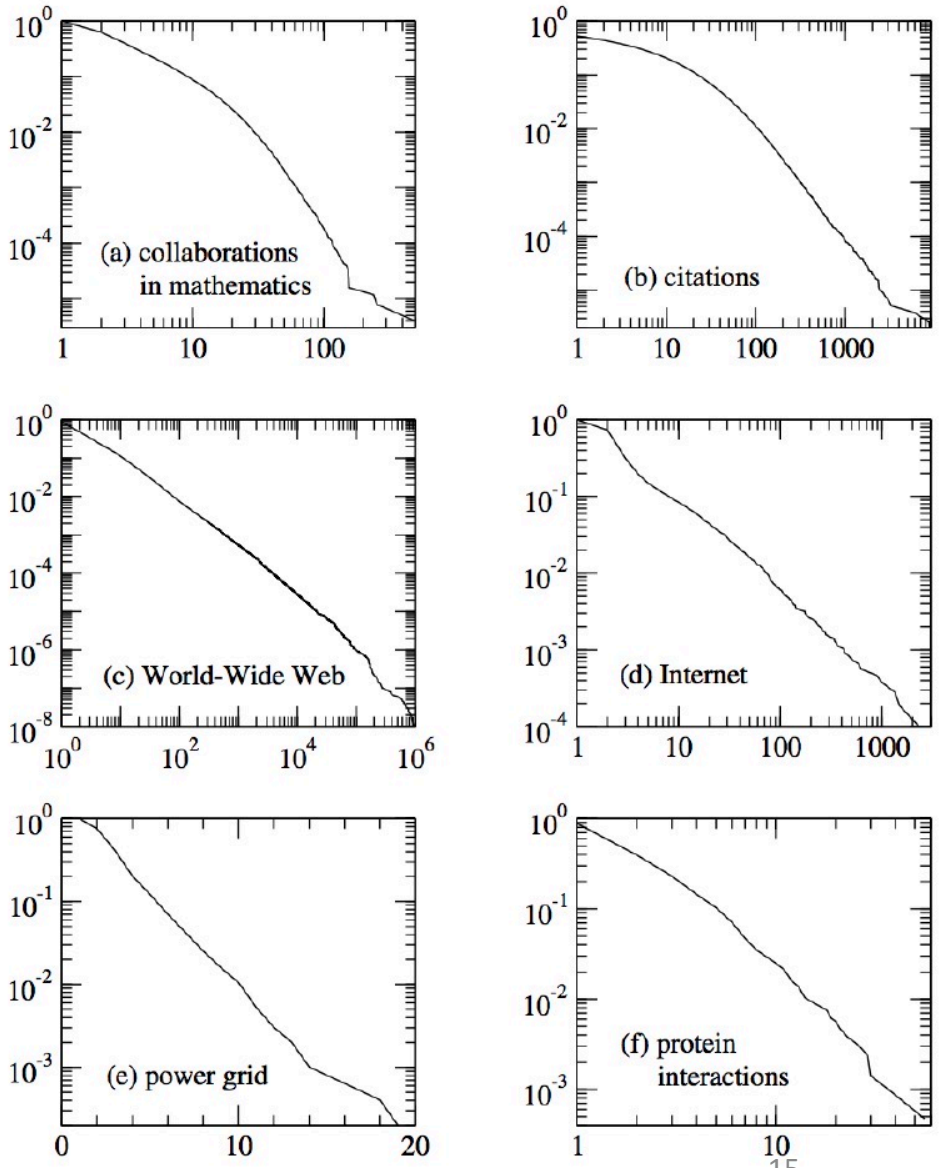
- **Methods**
  - Spectrum or factorization based (analysing adjacency matrix)
  - Random walk based (random walk over the graph to generate sequences)
  - Neural network based (learning on graph structure, e.g., GNN)



Social network analysis – distributions

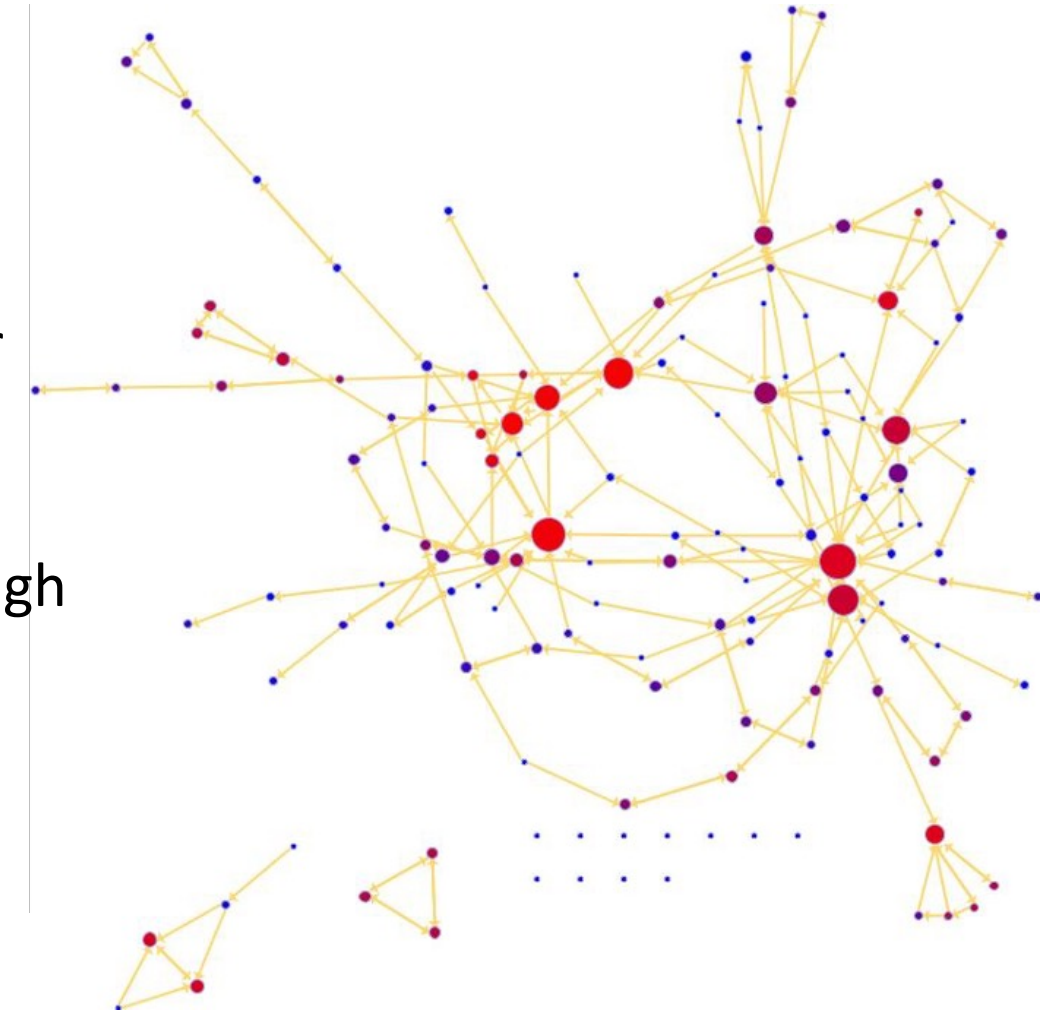
- Distribution of degrees
  - Horizontal axis: degrees; vertical axis: number of vertices
  - Common distributions: power law with long-tail, Gaussian (bell shape)

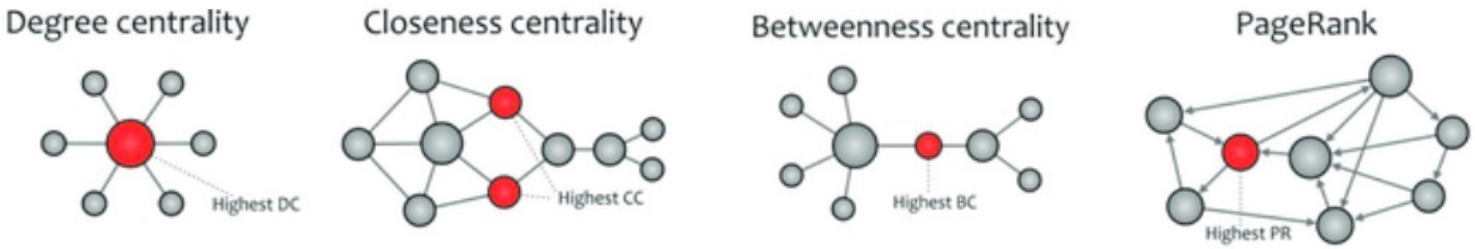




Social network analysis – centrality

- Centrality measures aim to quantify the importance of nodes in a network.
  - Degree centrality: based on degrees
  - Closeness centrality: based on distance from other nodes.
  - Betweenness centrality: based # paths going through
  - Importance centrality: based on weights of nodes, such as PageRank.



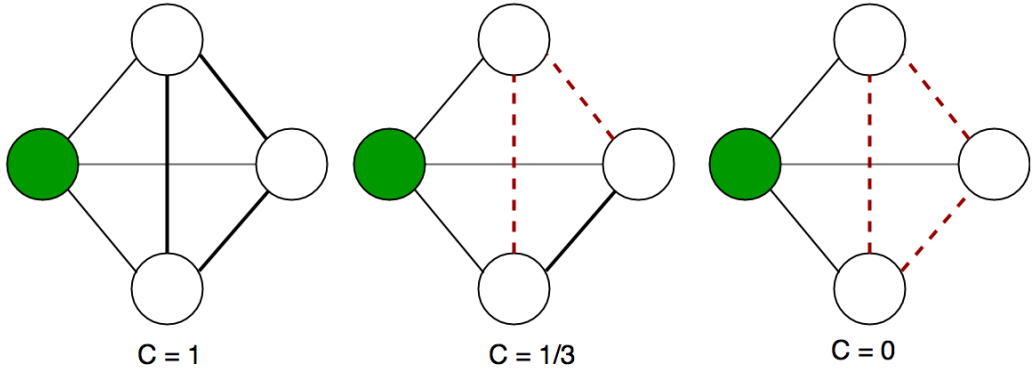


Social network analysis – Local cohesion

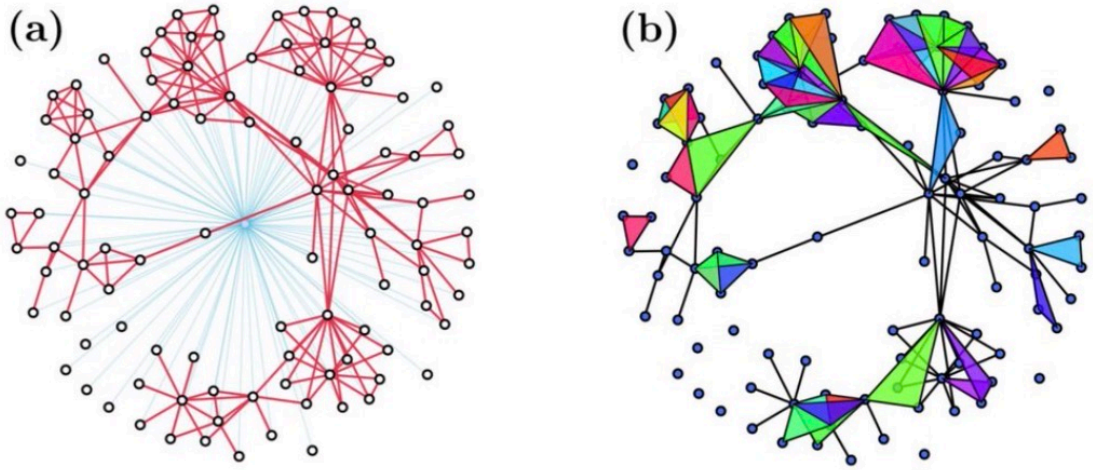
- Clustering coefficient
  - The amount of triangles in a network is quantified by the clustering coefficient of a node.

$$C_i \equiv \frac{\text{number of triangles including the } i\text{th node}}{k_i(k_i-1)/2}$$

possible triangles connecting to node i



- Local cohesion: the clustering coefficient measures the cohesion level
  - E.g., if all three are friends (a triangle), then they are cohesive; if two friends are enemies (not connected), then they are not cohesive.

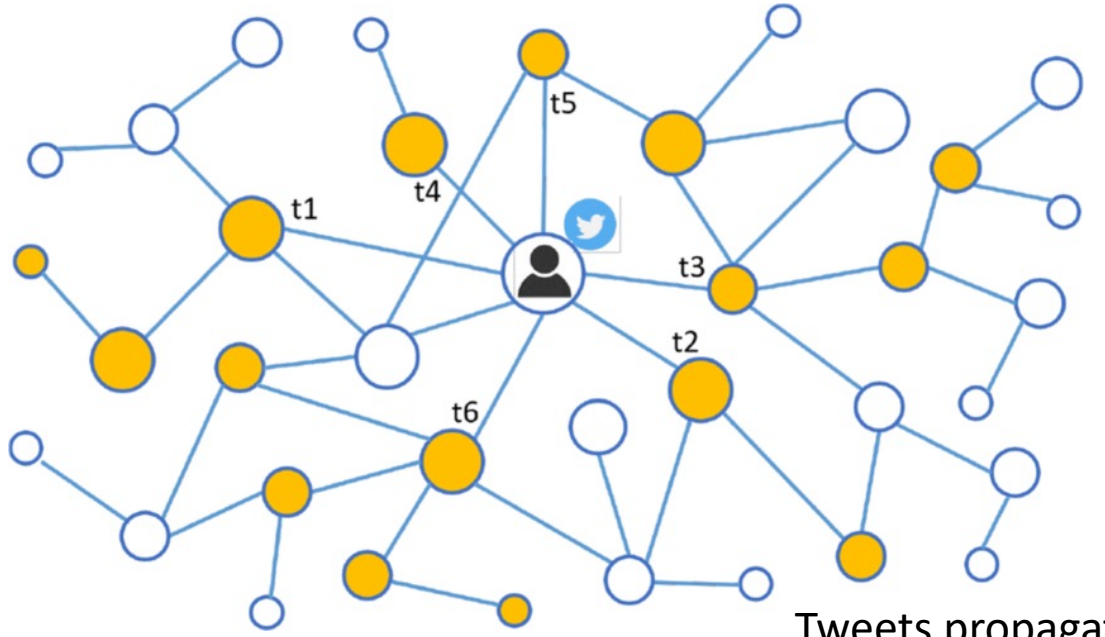


Social network analysis – community detection

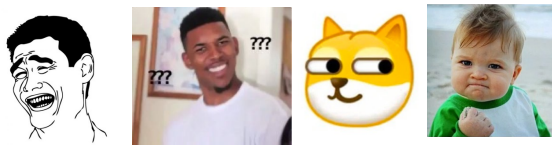
- **Community detection**, also called graph partition, helps us to reveal the hidden relations among the nodes in the network.
  - If the graph has the **community structure**, then its nodes can be easily grouped into sets of nodes such that each set of nodes is densely connected internally.
- Social networks include community groups based on common location, interests, occupation, etc. E.g., citation networks form communities by research topics.
  - Being able to identify these sub-structures within a network can provide insight into how network function and topology affect each other.
  - Communities often have very different properties than the average properties of the networks. Thus, only concentrating on the average properties usually misses many important and interesting features inside the networks.

Social network analysis – network propagation

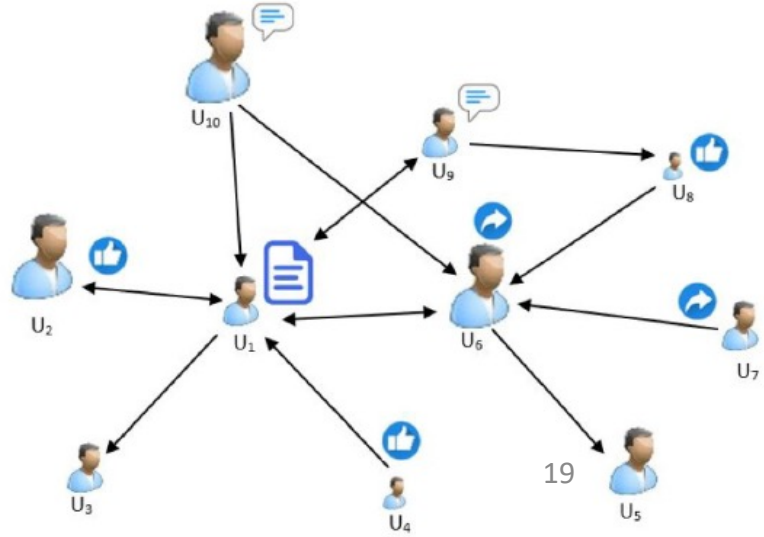
- **Network propagation** refers to a class of algorithms that integrate information from input data across connected nodes in a given network.
- Study the dynamics of network propagation, i.e., how data/information is spread across networks, and how networks are influenced in the meantime.
  - Epidemic/Virus spread model
- Applications
  - News/Tweets/Memes/Trends propagation
  - Rumer/Fake news detection



Tweets propagation



Memes = social genes



Fake news detection

Challenges in social network analysis

- Heterogeneity of social networks: size, data, structure, etc.
- Scalability: how to deal with large-scale social networks?
- Missing, incomplete, and unstructured data
- Model accuracy

Anomaly detection overview

Anomaly detection: the identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behavior.

Historically, the field of statistics tried to find and remove outliers as a way to improve analyses.

There are now many fields where the outliers / anomalies are the objects of greatest interest.

- The rare events may be the ones with the greatest impact, and often in a negative way.

Causes of Anomalies

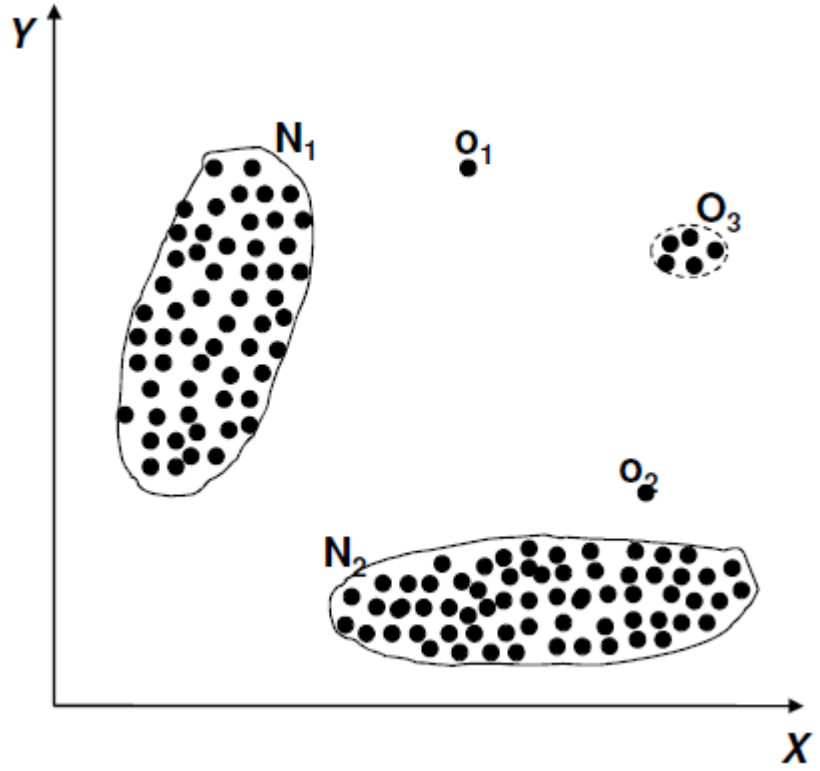
- Data from different class of object or underlying mechanism
    - disease vs. non-disease
    - fraud vs. not fraud
- Natural variation
    - tails on a Gaussian distribution
- Data measurement and collection errors

Structure of anomalies

- Point anomalies
- Contextual anomalies
- Collective anomalies

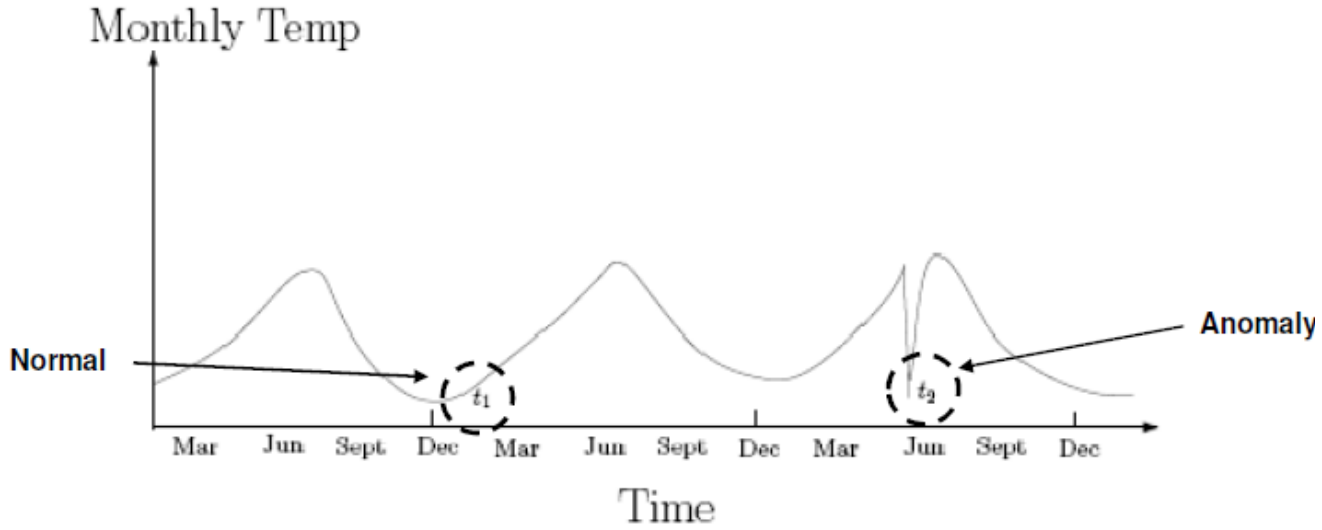
Point anomalies

- An individual data instance is anomalous with respect to the data



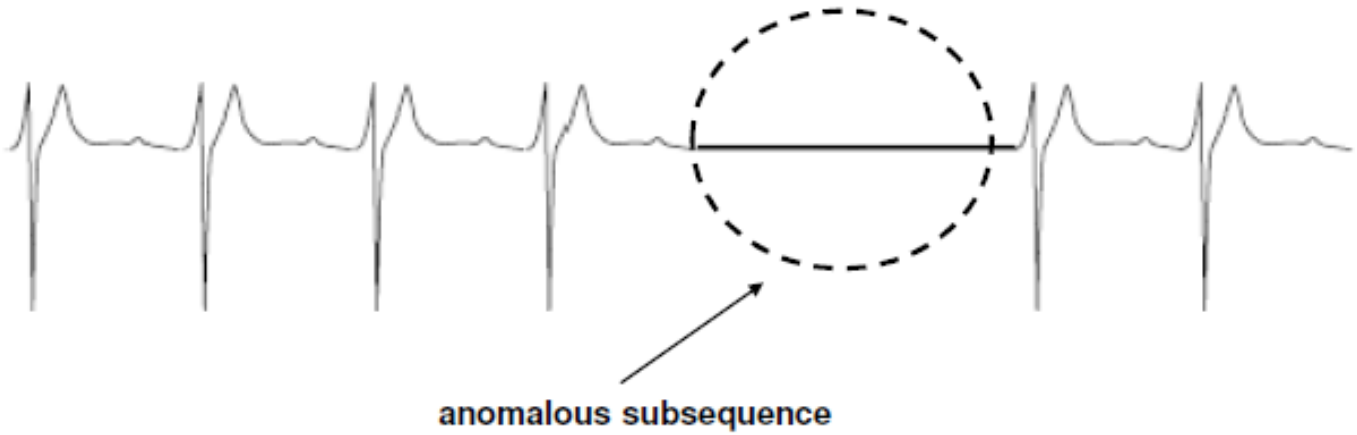
Contextual anomalies

- An individual data instance is anomalous within a context
- By combining with contexts: circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood, such as time, location, and related events.
- Also referred to as conditional anomalies



Collective anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential data
  - Spatial data
  - Graph data
- The individual instances within a collective anomaly are not anomalous by themselves



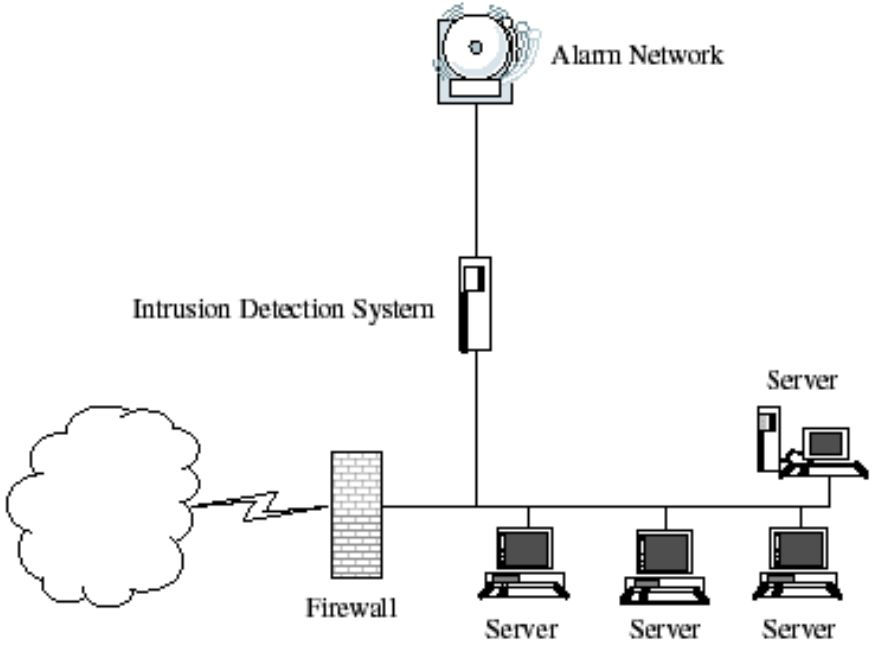
Applications of anomaly detection

- Network intrusion
- Insurance / credit card fraud
- Healthcare informatics / medical diagnostics
- Industrial damage detection
- Image processing / video surveillance
- Novel topic detection in text mining
- …

Network intrusion detection

Intrusion Detection

- Monitor events occurring in a computer system or network and analyze them for intrusions
- Intrusions defined as attempts to bypass the security mechanisms of a computer or network



Challenges

- Traditional intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
- Substantial latency in deployment of newly created signatures across the computer system

Anomaly detection can alleviate these limitations

Credit card fraud detection

Fraud Detection

Detection of criminal activities occurring in commercial organizations.

Malicious users might be:

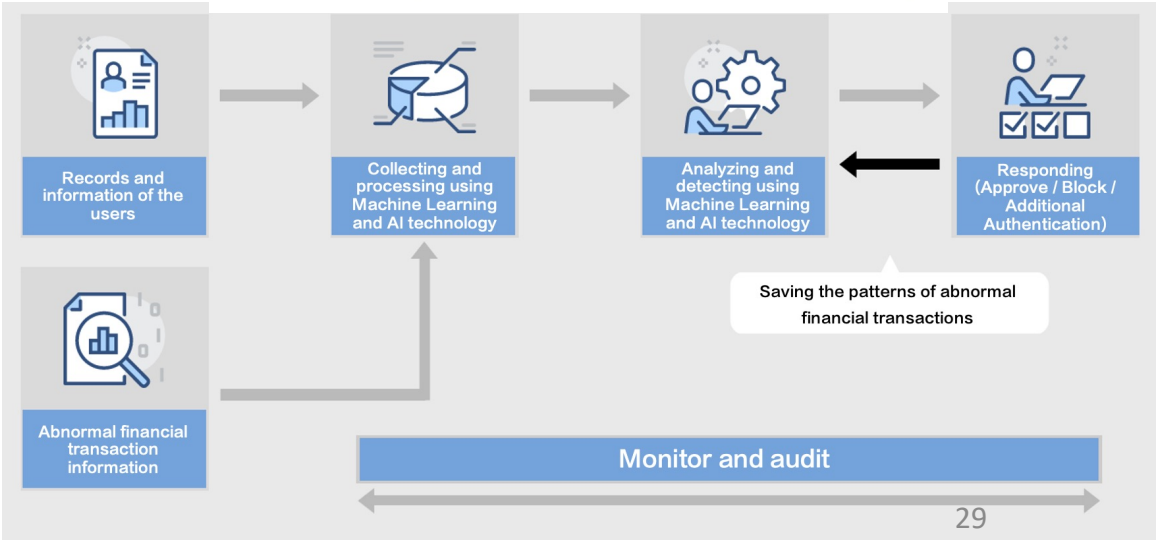
- Employees
- Actual customers
- Someone posing as a customer (identity theft)

Types of fraud

- Credit card fraud
- Insurance claim fraud
- Mobile / cell phone fraud
- Insider trading

Challenges

- Fast and accurate real-time detection
- Misclassification cost is very high



Healthcare diagnosis informatics

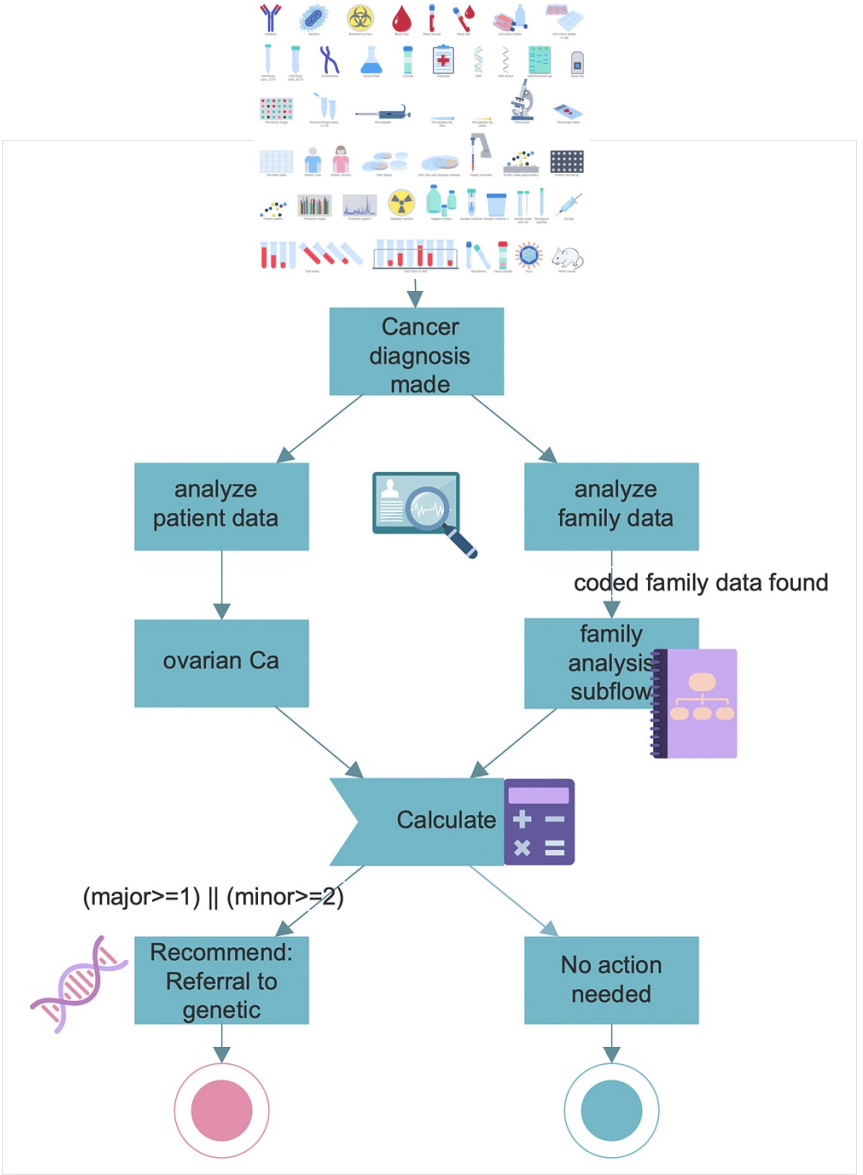
Detect anomalous patient records

- Indicate disease outbreaks,

  instrumentation errors, etc.

Key challenges

- Only normal labels available
- Misclassification cost is very high
- Data can be complex: spatio-temporal



Industrial damage detection

- Detect faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.

  Example: aircraft safety

  - anomalous aircraft (engine) / fleet usage
  - anomalies in engine combustion data
  - total aircraft health and usage management

- Key challenges
  - Data is extremely large, noisy, and unlabelled
  - Most of applications exhibit temporal behavior
  - Detected anomalous events typically require immediate intervention

Image processing

Detecting outliers in a image monitored over time.

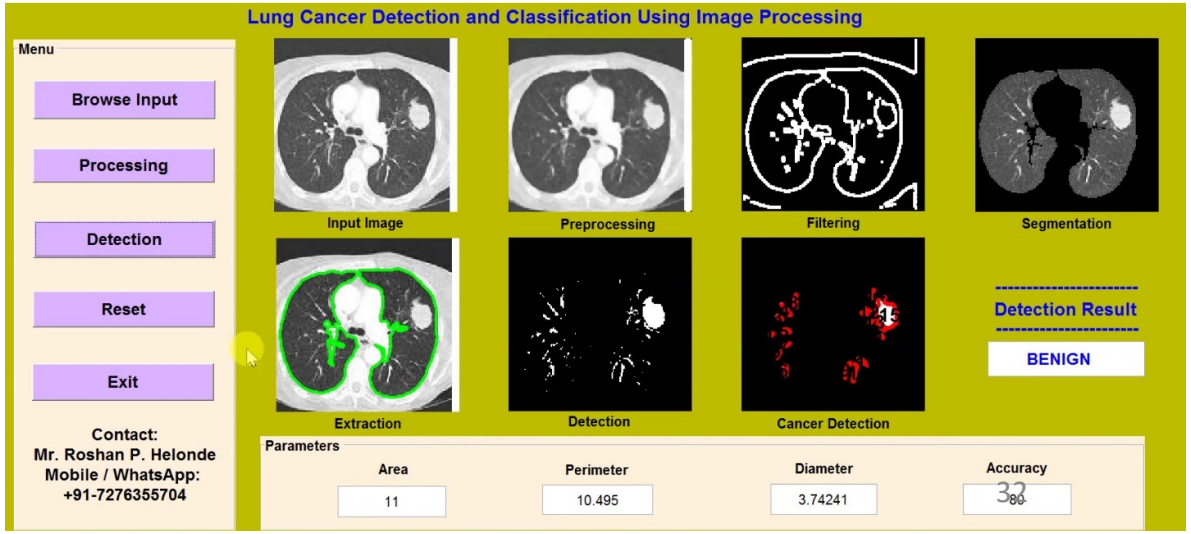
Detecting anomalous regions within an image.

Used in

- mammography image analysis
- video surveillance: e.g., traffic accident detection
- satellite image analysis

Key Challenges

- Detecting collective anomalies
- Data sets are very large



Output of anomaly detection

Label

- Each test instance is given a normal or anomaly label
- Typical output of classification-based approaches

Score

Each test instance is assigned an anomaly score

- allows outputs to be ranked
- requires an additional threshold parameter

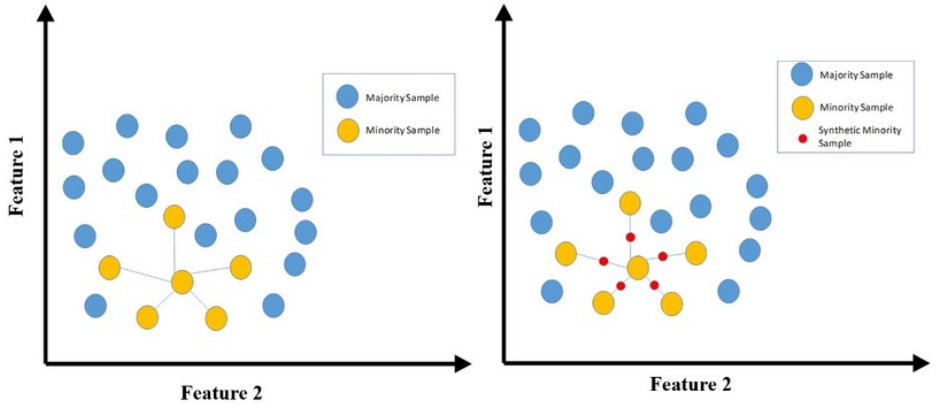
Variants of anomaly detection

- Given a dataset D, find all the data points $x \in D$ with anomaly scores greater than some threshold t.
- Given a dataset D, find all the data points $x \in D$ having the top-n largest anomaly scores.
- Given a dataset D, containing mostly normal data points, and a test point x, compute the anomaly score of x with respect to D.

Data labels for anomaly detection

Supervised anomaly detection

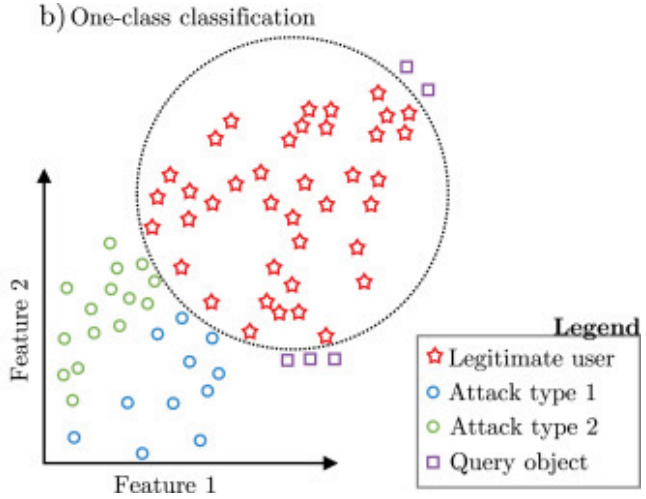
- Labels available for both normal data and anomalies
- Similar to classification with high class imbalance



- Data augmentation solution: SMOTE (Synthetic Minority Oversampling TEchnique) is an oversampling technique where the synthetic samples are generated for the minority class.

Semi-supervised anomaly detection

- Labels available only for normal data
- One class classification solution: only learning normal pattern



Unsupervised anomaly detection

- No labels assumed
- Based on the assumption that anomalies are very rare compared to normal data

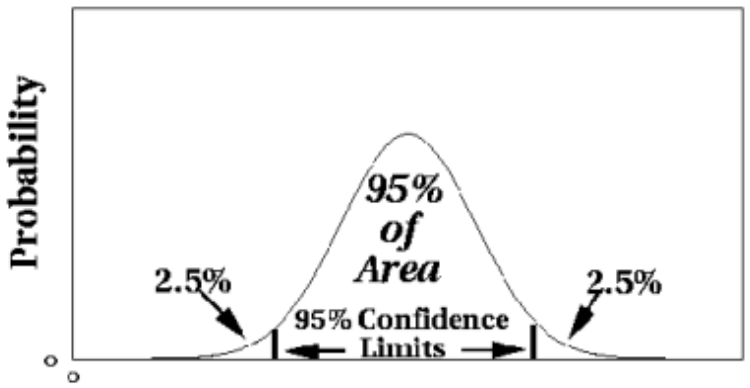
Unsupervised anomaly detection

- No labels available
- Based on assumption that anomalies are very rare compared to “normal” data
- General steps

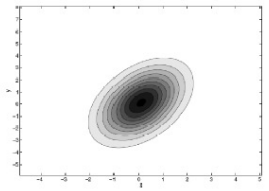
  - Build a profile of “normal” behavior
    - summary statistics for overall population
    - model of multivariate data distribution
  - Use the “normal” profile to detect anomalies
    - anomalies are observations whose characteristics differ significantly from the normal profile

Techniques for unsupervised anomaly detection (1)

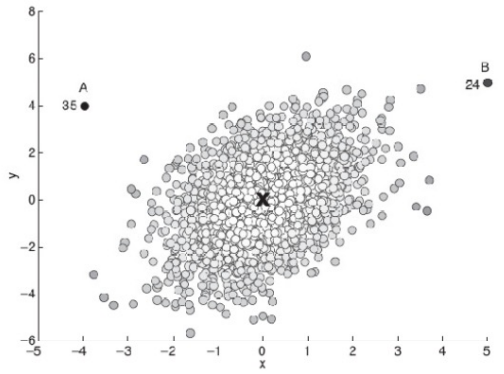
- Statistical: based on univariate or multivariate Gaussian distribution
  - Parameters of model (e.g., mean, variance)
  - Confidence limit (related to number of expected outliers)



Univariate Gaussian



	Distance	
	Euclidean	Mahalanobis
A	5.7	35
B	7.1	24



Multivariate Gaussian

Techniques for unsupervised anomaly detection (2)

- Proximity/Distance-based: anomalous data are those that are distant from most of other data
  - Outlier score is distance to *kth* nearest neighbor.
  - Score sensitive to choice of *k*.

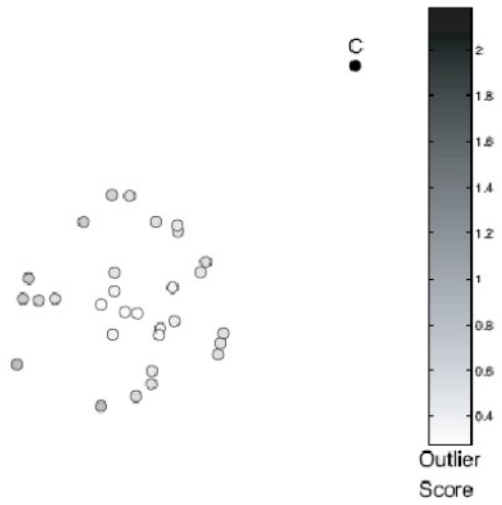


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

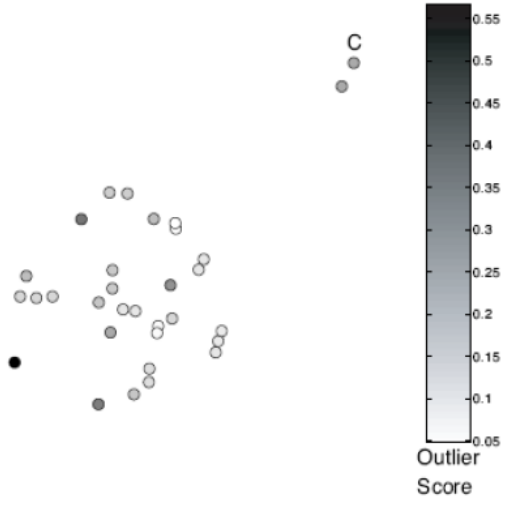


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

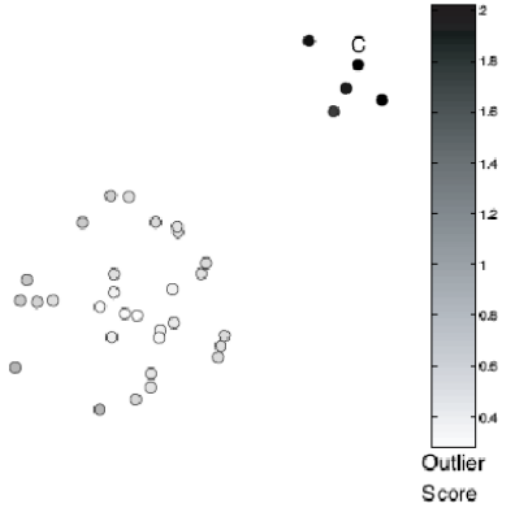


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

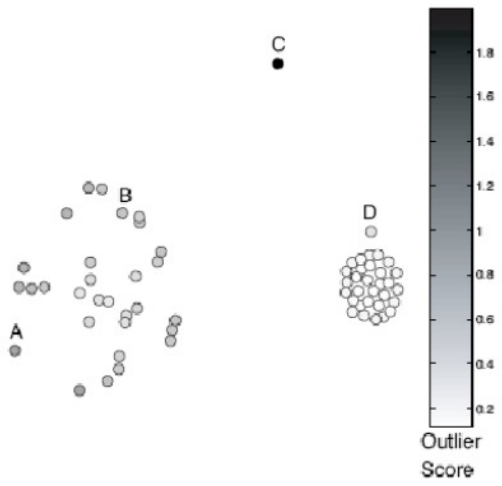
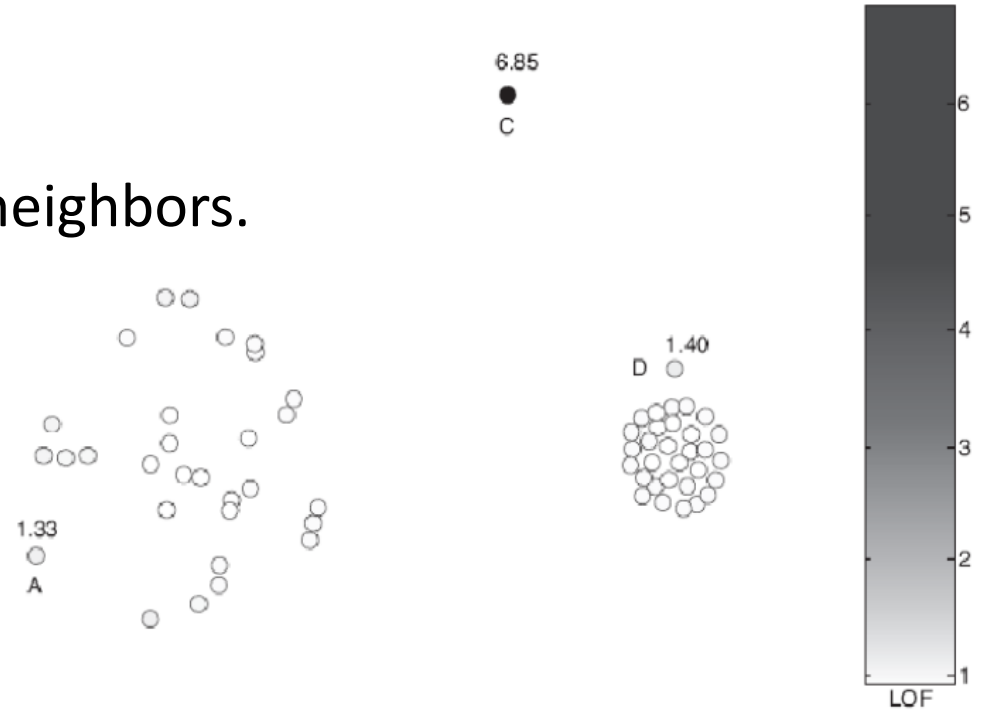


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Techniques for unsupervised anomaly detection (2)

- Density-based: anomalous data are those that are in regions of low density and are relatively distant from other data.
  - Outlier score is inverse of density around object.
  - Scores usually based on proximities.
  - E.g., based on average distance to k nearest neighbors.

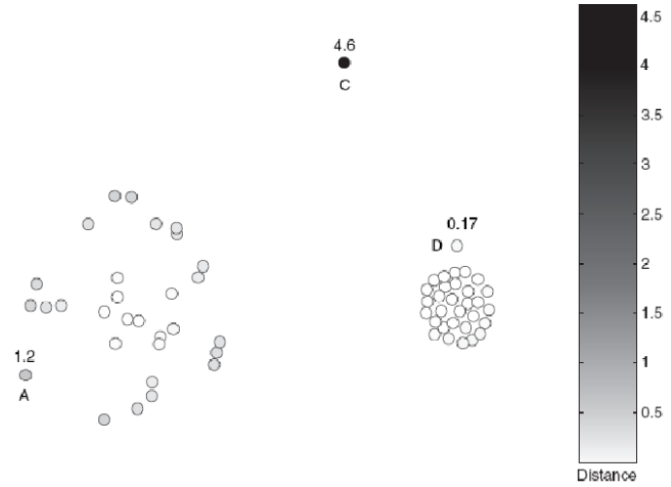
$$\text{density}(\mathbf{x}, k) = \left(\frac{1}{k} \sum_{\mathbf{y} \in N(\mathbf{x},k)} \text{distance}(\mathbf{x}, \mathbf{y}) \right)^{-1}$$



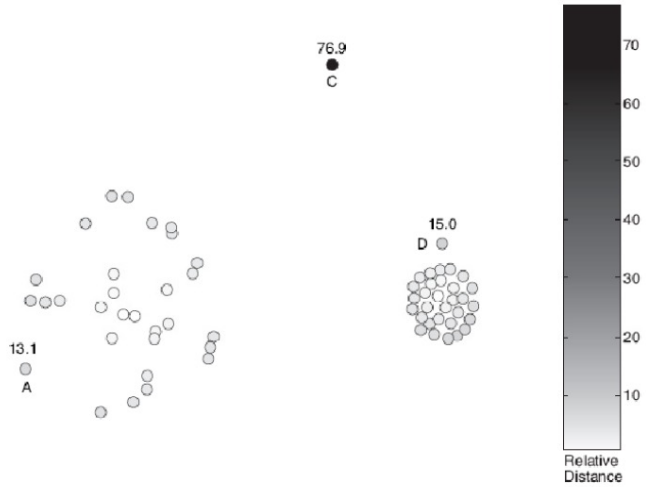
relative density (LOF) outlier scores

Techniques for unsupervised anomaly detection (2)

- Clustering-based: outliers are objects that do not belong strongly to any cluster.
  - Assess degree to which (distance or relative distance to cluster centers) object belongs to any cluster (e.g. through k-means).
  - Eliminate object(s) to improve objective function.
  - Discard small clusters far from other clusters. Sensitive to number of clusters chosen.



distance of points from nearest centroid



relative distance of points from nearest centroid

5. Real-world issues in anomaly detection

- Data often streaming, not static
  - Credit card transactions
- Anomalies can be bursty
  - Network intrusions
- How to employ contextual or structure information?
  - Contexts: time, location, other situational information
  - Graph/Networks and other correlation of data

Thanks for your attention!

Appendix

1. https://www.coursera.org/learn/social-network-analysis/home/week/

2. https://www.coursera.org/learn/python-social-network-analysis

3. Some slides taken or adapted from: “Anomaly Detection: A Tutorial” ArindamBanerjee, Varun Chandola, Vipin Kumar, Jaideep Srivastava, University of Minnesota Aleksandar Lazarevic, United Technology Research Center

4.http://courses.washington.edu/css581/lecture_slides/18_anomaly_detection.pdf

5. https://courses-archive.maths.ox.ac.uk/node/view_material/53485