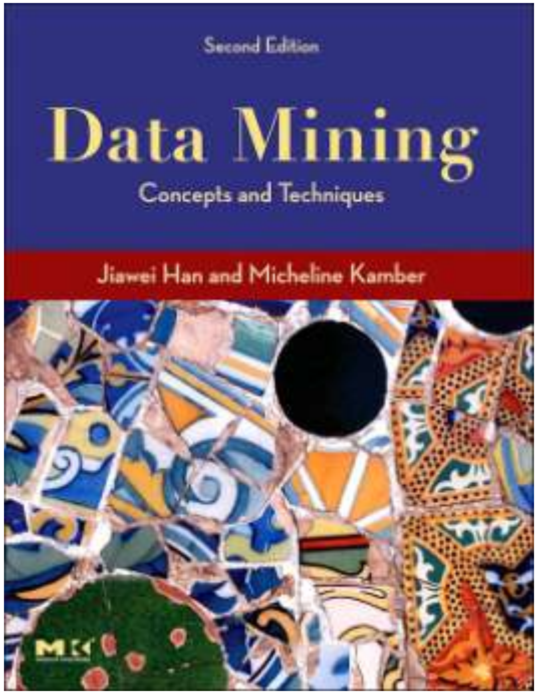
Graph Mining and Social Network Analysis

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

Daniele Loiacono

References

- Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems (Second Edition)
  - Chapter 9



Graph Mining

Graph Mining Overview

- Graphs are becoming increasingly important to model many phenomena in a large class of domains (e.g., bioinformatics, computer vision, social analysis)
- To deal with these needs, many data mining approaches have been extended also to graphs and trees
- Major approaches
  - Mining frequent subgraphs
  - Indexing
  - Similarity search
  - Classification
  - Clustering

Mining frequent subgraphs

- Given a labeled graph data set

$$D = \{G_1, G_2, \dots, G_n\}$$

- We define *support(g)* as the **percentage of graphs in D where *g* is a subgraph**
- A **frequent** subgraph in D is a subgraph with a support greater than *min_sup*
- How to find frequent subgraph?
  - Apriori-based approach
  - Pattern-growth approach

AprioriGraph

- Apply a level-wise iterative algorithm
  1. Choose two **similar size-k frequent** subgraphs in *S*
  2. **Merge** two similar subgraphs in a **size-(k+1)** subgraph
  3. If the new subgraph is **frequent** add to *S*
  4. Restart from 2. until all similar subgraphs have been considered. Otherwise restart from 1. and move to k+1.
- What is subgraph size?
  - Number of vertex
  - Number of edges
  - Number of edge-disjoint paths
- Two subgraphs of size-k are similar if they have the same size-(k-1) subgraph
- AprioriGraph has a big computational cost (due to the merging step)

PatternGrowthGraph

- Incrementally extend frequent subgraphs
  1. Add to S each frequent subgraphs g_E obtained by extending subgraph g
  2. Until S is not empty, select a new subgraph g in S to extend and start from 1.
- How to extend a subgraph?
  - Add a vertex
  - Add an edge
- The same graph can be discovered many times!
  - Get rid of duplicates once discovered
  - Reduce the generation of duplicates

Mining closed, unlabeled, and constrained subgraphs

- Closed subgraphs
  - *G is **closed** iff there is no proper supergraph G′ with the same support of G*
  - Reduce the growth of subgraphs discovered
  - Is a more compact representation of knowledge
- Unlabeled (or partially labeled) graphs
  - Introduce a special label Φ
  - Φ can match any label or only itself
- Constrained subgraphs
  - Containment constraint (edges, vertex, subgraphs)
  - Geometric constraint
  - Value constraint

Graph Indexing

- Indexing is basilar for effective search and query processing
- How to index graphs?
- **Path-based** approach takes the **path** as indexing unit
  - All the path up to *maxL* length are indexed
  - Does not scale very well
- **gIndex** approach takes **frequent** and **discriminative subgraphs** as indexing unit
  - A subgraph is frequent if it has a support greater than a threshold
  - A subgraph is discriminative if its support cannot be well approximated by the intersection of the graph sets that contain one of its subgraphs

Graph Classification and Clustering

- Mining of frequent subgraphs can be effectively used for classification and clustering purposes
- Classification
  - **Frequent** and **discriminative** subgraphs are used as **features** to perform the classification task
  - A subgraph is discriminative if it is frequent only in one class of graphs and infrequent in the others
  - The threshold on frequency and discriminativeness should be tuned to obtain the desired classification results
- Clustering
  - The mined frequent subgraphs are used to define **similarity** between graphs
  - Two graphs that **share a large set of patterns** should be considered **similar** and grouped in the same cluster
  - The threshold on frequency can be tuned to find the desired number of clusters
- As the mining step affects heavily the final outcome, this is an intertwined process rather tan a two-steps process

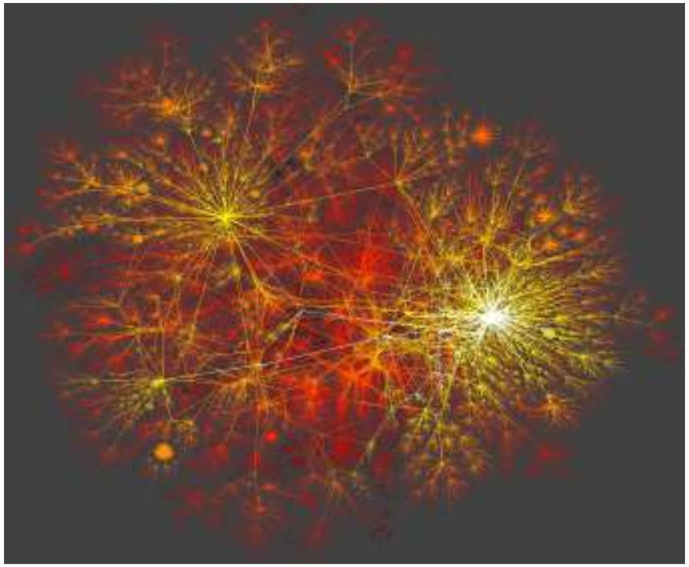
Social Network Analysis

Social Network

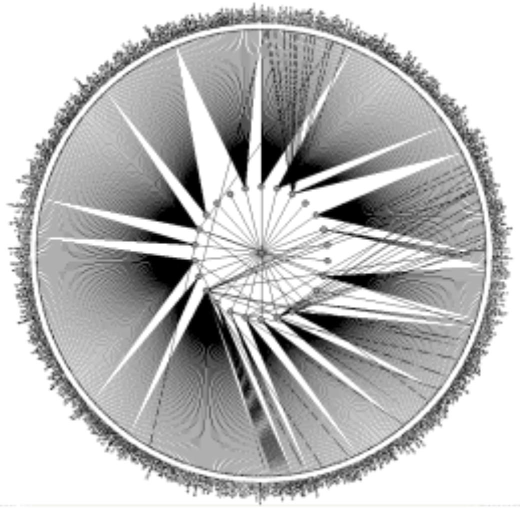
- A social network is an **heterogeneous** and **multirelational** dataset represented by a graph
  - Vertexes represent the **objects** (entities)
  - Edges represent the **links** (relationships or interaction)
  - Both objects and links may have **attributes**
  - Social networks are usually very large
- Social network can be used to represents many real-world phenomena (not necessarily social)
  - Electrical power grids
  - Phone calls
  - Spread of computer virus
  - WWW

Small World Networks (1)

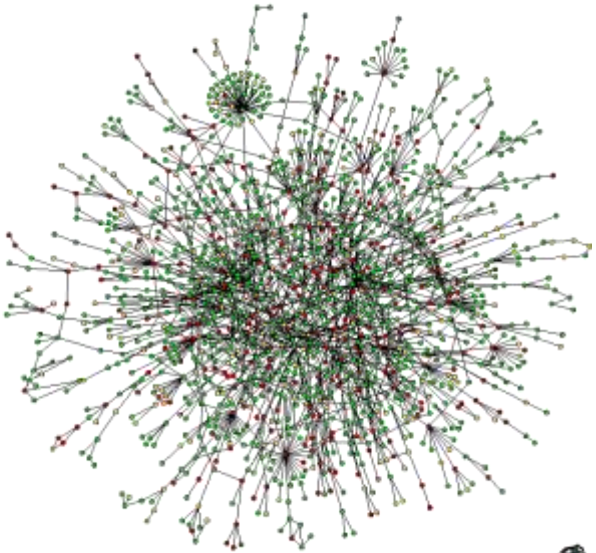
- Are social networks random graphs?
- NO!



Internet Map



Science Coauthorship

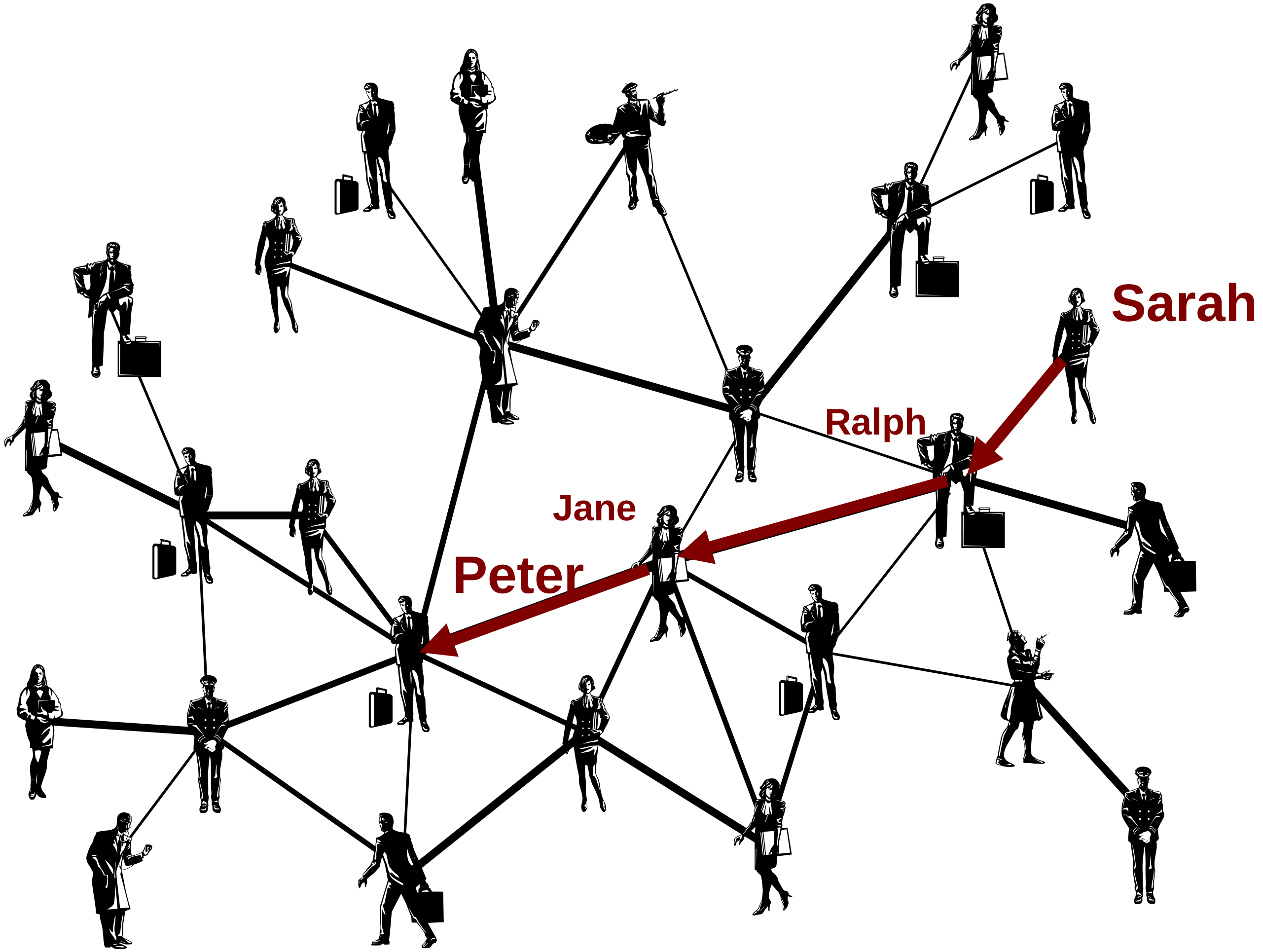


Protein Network

High degree of local clustering

Few degrees of separation

Small World Networks (2)



Society:
Six degrees
S. Milgram 1967
F. Karinthy 1929

WWW:
19 degrees
Albert *et al.* 1999

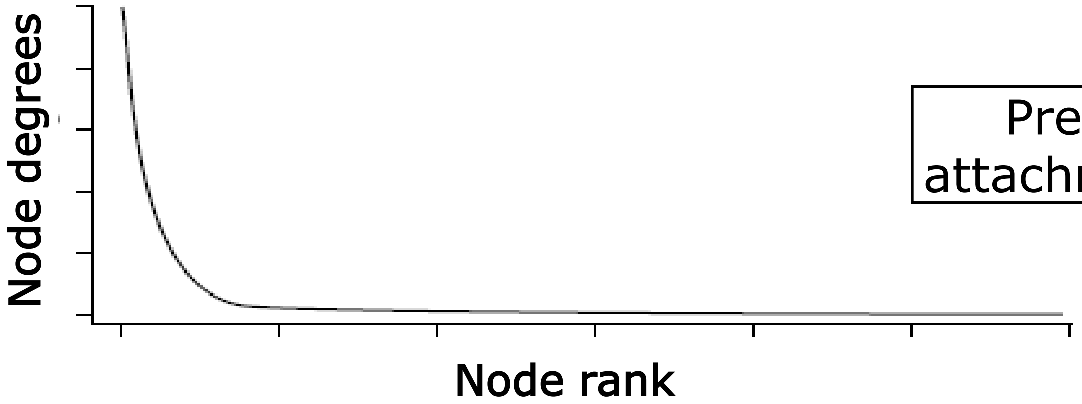
Small World Networks (3)

- **Definitions**
  - Node's **degree** is the number of incident edges
  - Network **effective diameter** is the max distance within 90% of the network
- **Properties**
  - **Densification power law**

$$e(t) = n(t)^{\alpha}$$

n: number of nodes
e: number of eges
$1<\alpha<2$

  - **Shrinking diameter**
  - **Heavy-tailed degrees distribution**



Preferential attachment model

Mining social networks (1)

- Several **Link mining** tasks can be identified in the analysis of social networks
- Link based object classification
  - Classification of objects on the basis of its attributes, its links and attributes of objects linked to it
  - E.g., predict topic of a paper on the basis of
    - Keywords occurrence
    - **Citations and cocitations**
- Link type prediction
  - Prediction of link type on the basis of objects attributes
  - E.g., predict if a link between two Web pages is an advertising link or not
- Predicting link existence
  - Predict the presence of a link between two objects

Mining social networks (2)

- Link cardinality estimation
  - Prediction of the number of links to an object
  - Prediction of the number of objects reachable from a specific object
- Object reconciliation
  - Discover if two objects are the same on the basis of their attributes and links
  - E.g., predict if two websites are mirrors of each other
- Group detection
  - Clustering of objects on the basis both of their attributes and their links
- Subgraph detection
  - Discover characteristic subgraphs within network

Challenges

- Feature construction
  - Not only the objects attributes need to be considered but also attributes of **linked objects**
  - **Feature selection** and **aggregation** techniques must be applied to reduce the size of search space
- Collective classification and consolidation
  - Unlabeled data cannot be classified independently
  - New objects can be **correlated** and need to be considered **collectively** to consolidate the current model
- Link prediction
  - The prior probability of link between two objects may be very low
- Community mining from multirelational networks
  - Many approaches assume an **homogenous relationship** while social networks usually represent **different communities** and **functionalities**

Applications

- Link Prediction
- Viral Marketing

Link prediction

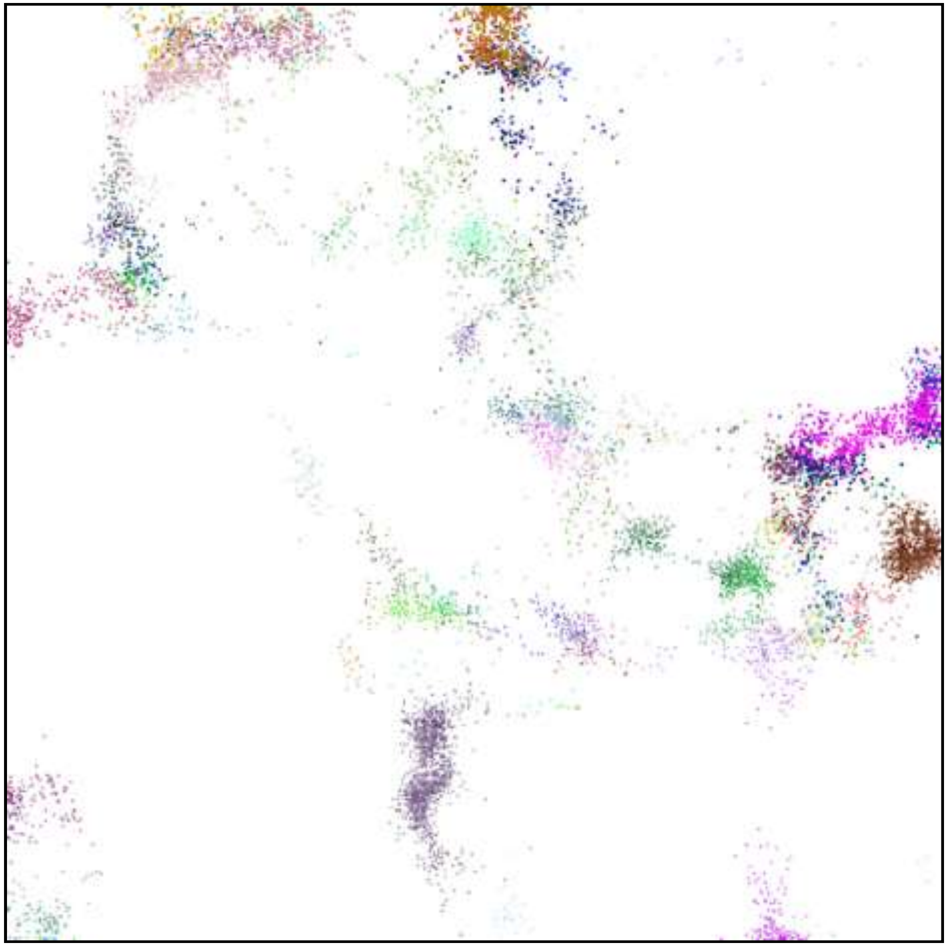
- What edges will be added to the network?
- Given a snapshot of a network at time *t*, **link prediction** aims to predict the edges that will be added before a given future time *t′*
- Link prediction is generally solved assigning to each pair of nodes a weight ***score(X,Y)***
- The higher the *score* the more likely that link will be added in the near future
- The *score(X,Y)* can be computed in several way
  - **Shortest path**: the shortest he path between X and Y the highest is their score
  - **Common neighbors**: the greater the number of neighbors X and Y have in common, the highest is their score
  - **Ensemble of all paths**: weighted sum of paths that connects X and Y (shorter paths have usually larger weights)

Viral Marketing

- Several marketing approaches
  - Mass marketing is targeted on specific segment of customers
  - Direct marketing is target on specific customers solely on the basis of their characteristics
  - **Viral marketing** tries to exploit the **social connections** to maximize the output of marketing actions
- Each customer has a specific **network value** based on
  - The number of connections
  - Its role in the network (e.g., opinion leader, listener)
  - Role of its connections
- Viral marketing aims to exploit the network value of customers to predict their influence and to maximize the outcome of marketing actions

Viral Marketing: Random Spreading

- 500 randomly chosen customers are given a product (from 5000).



Viral Marketing: Directed Spreading

- The 500 *most connected consumers are given a product*.