In The Name of Allah

Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

Complex Networks

Hamid R. Rabiee
Mostafa Salehi

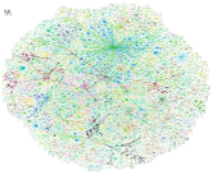
Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Agenda

- Introduction
- Network Characteristics
- Network Properties
- Questions to address
- Complex networks and pattern recognition
  - Community Detection
  - Link Prediction
  - Network Completion

Introduction

- **Complex networks is the name given to a multidisciplinary area (physics, mathematics and statistics, computer science and the social sciences)**
- **It is an area of research that has been of increasing importance over the past decade or so, as data acquisition systems have enabled large datasets of real networks to be gathered.**
- **Some Examples:**
  - **The Internet (a network of routers)**
  - **The World Wide Web (a network of web pages)**
  - **Social networks (a network of people)**
  - **Telephone call network (a network of telephone company subscribers)**
  - **The cell – a network of protein–protein interactions characterizing a cell;**

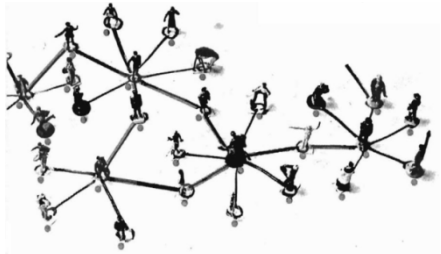


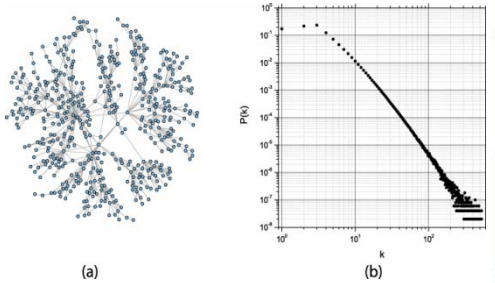
Network Characteristics

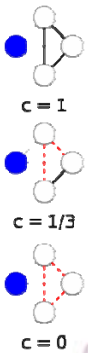
- Generally, the datasets describing the networks may be described by a **hypergraph** (a graph where the edges connect any number of vertices.)
- The datasets typically have the following characteristics:
- **Large size:** Networks can have millions of nodes (e.g., Facebook has 800+ millions users).
- **Sparseness.** The networks are sparse, most nodes having few connections to other nodes.
- **Weighted edges.** The connections between nodes are often weighted that describe the degree of interaction between the nodes (e.g., the weights could describe the amount of email traffic between people).
- **Temporal dependence of weights.** The strength of a connection between nodes is time varying (e.g, weights change with time as friendships form).
- **Growing/shrinking network (dynamic).** Nodes join and leave a network at different times (e.g., new Internet sites are created and old ones deleted.)
- **Attributed graphs.** Associated with each node may be a set of attributes or features (e.g., the age of a person, gender, location, qualifications).

Network Properties

- There are some properties of real-world networks that are common to networks of different types:
- The **small-world effect**: It refers to the property that in most, often large, networks, pairs of vertices are often connected by a short path length.
- **Scale-Free**: For most large networks, there are nodes whose values of the degree are far greater than the mean. Specifically, the distributions follow a power law in their tails. Networks with power law degree distributions are referred to as scale-free networks.
- **Clustering**: In social networks, groups or communities form according to factors such as the interests, age or occupation of the members. This tendency to form communities is not only a property of social networks and the identification of communities in networks is an important research topic. The degree to which nodes in a network cluster together to form a community can be quantified by the clustering coefficient.
- **Degree correlations**: The correlation between the degree of neighbouring nodes. (Are high degree nodes preferentially attached to other high degree nodes, or low degree nodes?)







Questions to address

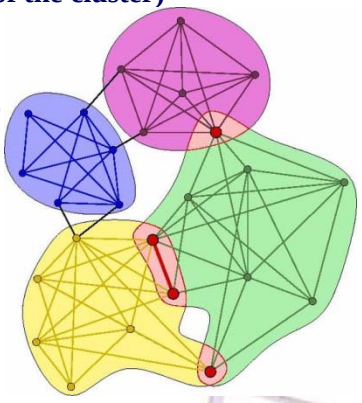
- **There are many questions that have been posed of datasets. Listed below are some of the properties of a network that we may wish to discover.**
- **Anomalous nodes (edges). Do any of the nodes (edges) have an unusual pattern of behaviour?**
- **Significant edges. These are connections between two nodes whose presence is vital to the network's function.**
- **Influential nodes. These are nodes whose removal results in a significant change in network behaviour. For example, an influential node may be the only common node to two communities.**
- **Community structure. This is the clustering of nodes into groups or communities for which there is high inter-community connectivity and low between-community connectivity.**

Complex networks and pattern recognition

- At the intersection of the research areas of theory and algorithms for the analysis of graphs and pattern recognition, there are broadly two main themes:
  1. The use of graph-based methods in pattern recognition applications. (Ref: D.J. Marchette. Random Graphs for Statistical Pattern Recognition. John Wiley & Sons, Ltd, 2004.)
  2. The application of data analysis methods to data comprising measurements of interactions between entities. (Ref: A.R. Webb. Statistical Pattern Recognition, 3nd ed. Wiley, Chichester, 2011.)
- Here, the focus is on the latter theme: what patterns (structure) are present in the network? How can we find it?
- Next, we introduce some applications of data analysis methods (in pattern recognition) to the data arising from the networks:
  - Community detection, which largely employs unsupervised pattern recognition techniques to discover communities in networks;
  - Link prediction and Network Completion, which is supervised.

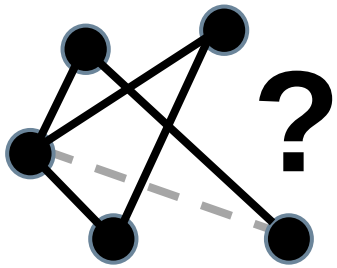
Community Detection

- Detecting the presence of communities – clusters of nodes with a high intracluster connectivity and a low intercluster connectivity – is important in many application areas such as protein discovery, marketing, and sociology,.
- The identification of high order structures in the data provides insights into the network's organisation leading, for example, to the discovery of functionally related proteins, groups of people sharing common interests or potential markets for new products.
- The term graph clustering is also used to refer to the task of grouping vertices in a graph, taking account of graph structure.
- Two approaches for identifying communities: 1) global approaches that use the entire vertex set and 2) local approaches that use a particular node as a seed to identify a local community.
- Community detection methods:
  - Clustering methods
    - Hierarchical methods (uses the distance measures between a node and any member of the cluster)
    - k-medoids method (corresponding to the k-means clustering algorithm)
    - Spectral methods (they are based on an eigen decomposition of the graph Laplacian)
  - Girvan–Newman algorithm
  - Modularity approaches
  - Local modularity
  - Clique percolation



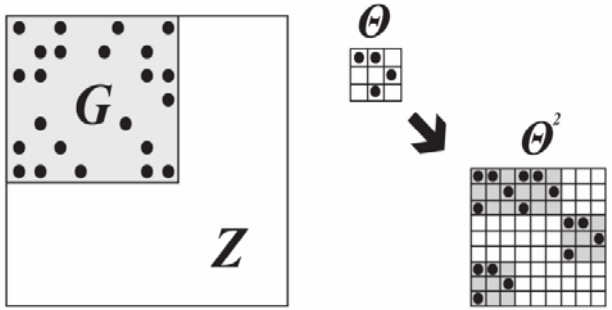
Link Prediction

- **Link prediction is the problem of predicting a link between two nodes in a network based on the existing observed links in the network and the properties of the nodes.**
- **Reasons for link prediction include:**
  - **to extract missing information.**
  - **to identify anomalous interactions.**
  - **to evaluate models of network evolution.**
- **Link prediction may be used in many diverse application areas. Examples include:**
  - **Social networks – predicting friendships.**
  - **Marketing – identifying potential markets for products.**
  - **Security – predicting anomalous links.**
- **Approaches to link prediction: One of the approaches to link prediction is to view the problem as one in binary classification. The two classes comprise pairs of nodes that are linked and pairs of nodes that are not linked, taken from a network at time t, or sampled from part of a larger network. Attributes of these pairs of nodes are used as inputs to a classifier which is then tested on the network at a later time or on an unseen part of a large network.**
- **One of the difficulties with a classification approach is that there is a very large class skew, with many more examples of nonlinked pairs, and this imbalance grows as a network evolves and becomes larger. This can lead to poor model performance with the prior probability of a link usually being very small.**



Network Completion

- **Many times the collected network data is incomplete with nodes and edges missing.**
- **Commonly, only a part of the network can be observed and we would like to infer the unobserved part of the net work.**
- **Network Completion Problem: Given a network with missing nodes and edges, can we complete the missing part?**
- **In order to capture the connectivity patterns in the observed part of the network and use this knowledge to complete the unobserved part of the network, one inherently requires a model of the structure of real-networks.**
- **Based on the Kronecker graphs model, we naturally cast the problem of network completion into the Expectation Maximization (EM) framework where we aim to estimate the Kronecker graphs model parameters as well as the edges in the missing part of the network.**



M. Kim and J. Leskovec, The Network Completion Problem: Inferring Missing Nodes and Edges in Networks, SIAM Conference on Data Mining, 2011.

References

- **A.R. Webb. Statistical Pattern Recognition, 3nd ed. Wiley, Chichester, 2011.**

Any Question

End of Lecture

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/