Data Mining, Lecture 12

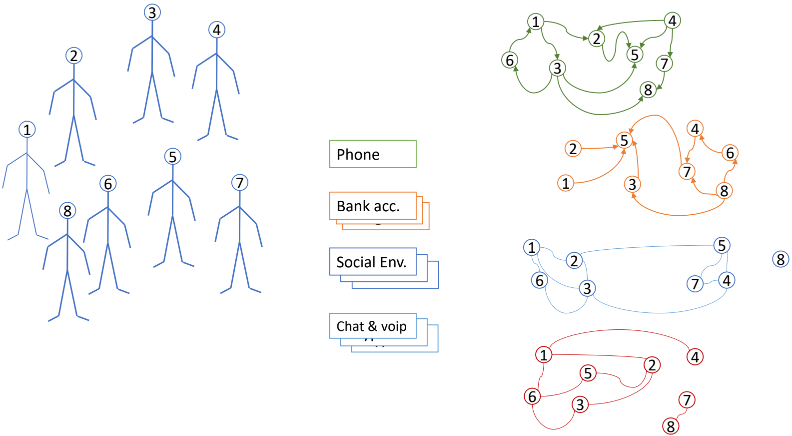
Social Networks Analysis

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

03.12.2019

Indiviaduals → Graphs



Preliminaries and properties

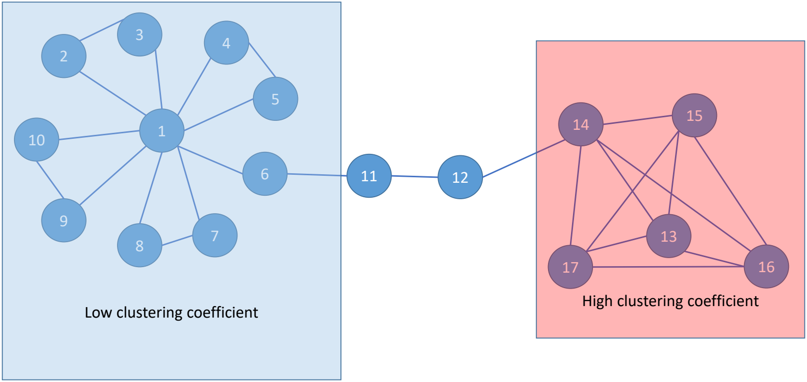
- Let us assume that social networks may be structured as a graph, $G = (N, A)$ where N is the set of nodes and A is the set of edges. Each individual in the networks is represented by a node in N and referred as *actor*. The edges represent connections between the actor.
- Assume that G is undirected.
- In some cases nodes may have content associated with them.
- Usually each node is associated with an actor (human individual).

Key properties

- *Homophily (Assortative mixing)*: nodes that are connected to one another are more likely to have similar properties.
- *Triadic closure:* If two individuals in a social network have a friend in common, then it is more likely that they are either connected or will eventually become connected in the future. **Observe the reference to some dynamics**. Implies an inherent correlation in the edge structure of the network
- *The clustering coefficient of the network* is the measure of the inherent tendency of a network to cluster. Similar to the Hopkins statistic for multidimensional data.
- Let $S_i \subseteq N$ be the subset of nodes connected to the node $i \in N$, let the cardinality of S_i be n_i. The local clustering coefficient is defined as follows:

$$\eta(i) = \frac{|\{j,k\} \in A : j \in S_i, k \in S_i|}{\binom{n_i}{2}}$$

Clustering Coefficient



Associations

- **Associates**: Basic level, do not share any interests.
- **Useful friends**: Information sharing.
- **Fun friends**: Socialise together, no emotional connection.
- **Favor friends**: may help each other, no emotional connection.
- **Help mates**: Combination of two previous.
- **Comforters**: Help mates with emotional connection.
- **Confidants**: Share personal emotional information, socialize together, unable to help each other.
- **Soulmates** : Most probably the tightest type of connection.
- number of stronger associations is always smaller than number of weak associations.

The rule of 5-15-50-150-500

- Internal circle 5
- Sympathy group 12 -15
- More or less regular group 50
- Stable social group 150 (Dunabar's value).
- Weak associations 500

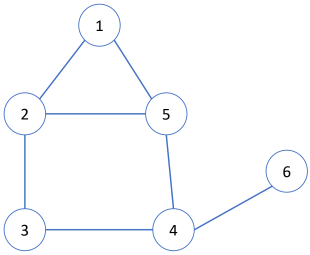
Key properties: Dynamics of Network Formation

- Preferential attachment: In a growing network, the likelihood of a node receiving new edges increases with its degree. Highly connected individuals will typically find it easier to make new connections.
- Let $\pi(i)$ is the probability a newly added node attaches itself to an existing node i. The model of $\pi(i)$

$$\pi(i) \propto \mathrm{Degree}(i)^{\alpha}$$

  where the value of the parameter α depend on domain where form the network is drawn.
- *Small world property:* Most real networks are assumed to be small world. Average path growth is $log(n(t))$.

Degrees and frequencies



Node	Deg.
1	2
2	3
3	2
4	3
5	3
6	1

Deg.	Frequency
1	1/6
2	2/6
3	3/6

Key properties: Dynamics of Network Formation II

- *Densification:* Almost all real-world networks add more nodes and edges over time than are deleted.

  $$e(t) \propto n(t)^{\beta}$$

  where $e(t)$ is the number of edges, exponent of β is the value between 1 and 2.
- *Shrinking diameters:* In most real-world networks, as the network densifies, the average distances between the nodes shrink over time.
- *Giant connected component:* As the network densifies over time, a giant connected component emerges.
- *Power-Law Degree Distributions:* a small minority of high-degree nodes continue to attract most of the newly added nodes:

  $$P(k) \propto k^{-\gamma}$$

  where γranges between 2 and 3;larger values of γ lead to more small degree nodes.

Key properties

- *Degree Centrality and Prestige* The degree centrality $C_D(i)$ of a node i of an undirected network is equal to the degree of the node, divided by the maximum possible degree of the nodes.

$$C_D(i) = \frac{\mathrm{Degree(i)}}{n-1}.$$

- *Degree prestige* is defined for directed networks only.

$$P_D(i) = \frac{\mathrm{InDegree(i)}}{n-1}.$$

- *The gregariousness of a node:* (extension of the centrality to outdegeree):

$$G_D(i) = \frac{\mathrm{OutDegree(i)}}{n-1}.$$

  The gregariousness of a node defines a different qualitative notion than prestige because it quantifies the propensity of an individual to seek out new connections.

Closeness Centrality and Proximity Prestige

- *Closeness centrality:* for undirected and connected networks. The average shortest path distance, starting from node i, is denoted by $\mathrm{AvDist}(i)$:

  $$\mathrm{AvDist}(i) = \frac{\sum_{j=1}^{n} \mathrm{Dist}(i,j)}{n-1}.$$

  The closeness centrality is the inverse of the average distance of other nodes to node i.

  $$C_C(i) = 1/\mathrm{AvDist}(i)$$

  ranges between 0 and 1.

Closeness Centrality and Proximity Prestige

- *Proximity prestige:* defined for the directed networks. $\mathrm{Influence}(i)$ corresponds to all recursively defined "followers" of i.

$$\mathrm{AvDist}(i) = \frac{\sum_{j \in \mathrm{Influence}(i)} \mathrm{Dist}(i,j)}{|\mathrm{Influence}(i)|}.$$

- Nodes that have less influence should be penalized.

$$\mathrm{InfluenceFraction}(i) = \frac{\mathrm{Influence}(i)}{n-1}.$$

- Proximity prestige may be defined as follows:

$$P_P(i) = \frac{\mathrm{InfluenceFraction}(i)}{\mathrm{AvDist}(i)}.$$

Betweenness Centrality

- Closeness centrality does not account the degree of importance (criticality) of the node with respect to the number of shortest paths goes through it.
- Let $q_{j,k}$ denotes the number of shortest paths between nodes j and k. Let $q_{j,k}(i)$ be the the number of shortest paths goes through the node i.
- denote by $f_{jk}(i)$ the fraction of pairs that pass through the node i

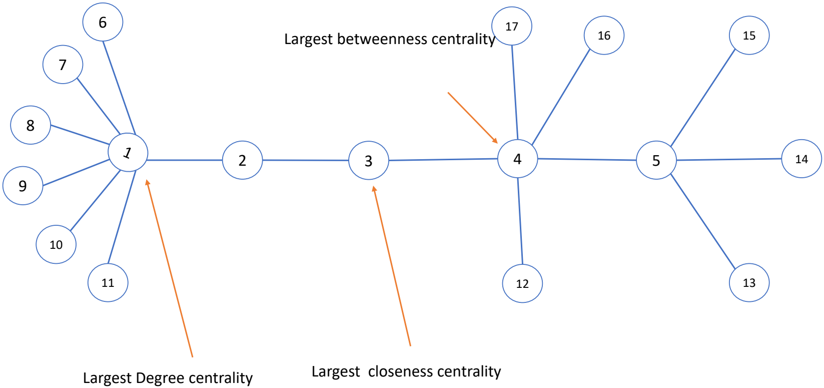
$$f_{jk}(i) = \frac{q_{j,k}(i)}{q_{j,k}}.$$

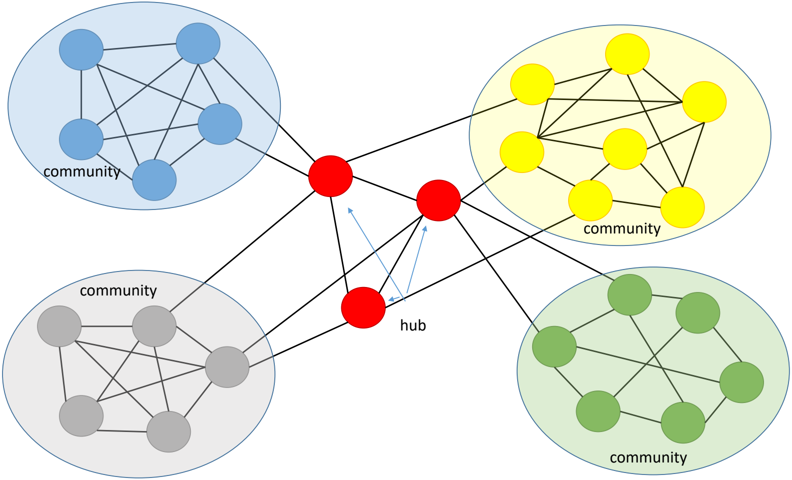
- *The betweenness centrality:* is defined as follows:

$$C_B(i) = \frac{\sum_{j<k} f_{jk}(i)}{\binom{n}{2}}.$$

May be generalized for disconnected networks. May be redesigned for edges.

Centrality and Prestige




community
community
community
community
hub

Communities

- Giant (gigantic) component - connected component (where all or nearly all the vertexes are connected), which weight in the network is constant.

$$\lim_{N \to \infty} \frac{N_1}{N} = c > 0.$$

- Community structure. Connected components and weak associations between them.

Community detection

- Community detection is an approximate synonym for clustering in the context of social network analysis.
- Methods of "graph partitioning" may b e applied.
- k-means and other non specific clustering algorithm may not be easily applied here.
- Different parts of the social network have different edge densities. In other words, the local clustering coefficients in distinct parts of the social network are typically quite different.
- KernighanLin Algorithm.
- GirvanNewman Algorithm.
- Multilevel Graph Partitioning: METIS
- Spectral Clustering

Collective Classification

- Iterative Classification Algorithm.
- Label Propagation with Random Walks.
- Supervised Spectral Methods.

Link Prediction

- Structural measure. Structural measures typically use the principle of triadic closure to make predictions.
- Content-based measures. In these cases, the principle of homophily is used to make predictions.

Neighbourhood based measures

- Common neighbour based measure between nodes i and j.

$$C_N(i,j) = |S_i \cap S_j|.$$

The major weakness of the common-neighbor measure is that it does not account for the relative number of common neighbors between them as compared to the number of other connections.

- Jaccard Measure:

$$J_M(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}.$$

Main drawback is that it does not adjust well to the degrees of their intermediate neighbors.

- AdamicAdar Measure:

$$A_A(i,j) = \sum_{k \in S_i \cap s_j} \frac{1}{\log(|S_k|)}$$

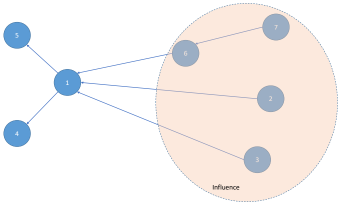
Neighbourhood based measures

- *Katz measure*. Effective when the number of shared links is small.

$$K_M(i,j) = \sum_{t=1}^{\infty} \beta^t n_{i,j}^t$$

Influence

- For the oriented graphs one may define prestige in the context of the close neighborhood.
- Influence I_i of the vertex v_i is the subset of the vertexes such that they are terminal for at least one path with origin in the vertex v_i.



Practice

- Install "igraph" package for "R".
- The practice is based on https://kateto.net/networks-r-igraph .
- Download the data set from kateto.net/netscix2016 .