Detecting Communities in Social Networks

Overview

- Introduction
- Definition of communities
- Evaluating communities
- Methods of community detection
- Tools for detecting communities

Introduction

- Relations of real-world entities are often represented as networks, such as social networks connected with friendships or co-authorships.
- Real social networks contain denser parts and sparser parts.
- Denser subnetworks correspond to groups of people that are closely connected with each other. Such denser subnetworks are called “communities”.

Contd.

Detecting communities from given social networks are practically important for the following reasons:

1. Communities can be used for information recommendation -members often have similar tastes and preferences. Membership of detected communities will be the basis of collaborative filtering.
2. Communities will help us understand the structures of given social networks. Communities are regarded as components of given social networks, and they will clarify the functions and properties of the networks.
3. Communities will play important roles when we visualize large-scale social networks. Relations of the communities clarify the processes of information sharing and information diffusions, and they may give us some insights for the growth the networks in the future.

Definition of Community

- The word “community” intuitively means a subnetwork whose edges connecting inside of it (intracommunity edges) are denser than the edges connecting outside of it (intercommunity edges).

- Definitions of community can be classified into the following three categories.
  - Local definitions
  - Global definitions
  - Definitions based on vertex similarity.

Local Definitions

The attention is focused on the vertices of the subnetwork under investigation and on its immediate neighborhood.

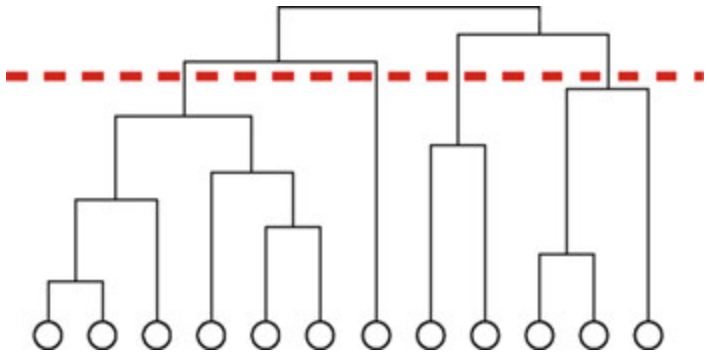
- Self referring ones – subnetwork
  - Clique - a maximal subnetworks where each vertex is adjacent to all the others
  - n-clique - a maximal subnetwork such that the distance of each pair of vertices is not larger than n
  - k-plex – a maximal subnetwork such that each vertex is adjacent to all the others except at most k of them
- Comparative ones - compares mutual connections of the vertices of the subnetwork with the connections with external neighbors.
  - LS set - a subnetwork where each vertex has more neighbors inside than outside of the subnetwork
  - Weak community - the total degrees of the vertices inside the community exceeds the number of edges lying between the community and the rest of the network

Global Definitions

- Global definitions of community characterize a subnetwork with respect to the network as a whole.
- These definitions usually starts from a null model,
  - a network which matches the original network in some of its topological features, but which does not display community structure.
  - Linking properties of subnetworks of the initial network are compared with those of the corresponding subnetworks in the null model. If there is a wide difference between them, the subnetworks are regarded as communities.
  - It consists of a randomized version of the original network, where edges are rewired at random, under the constraint that each vertex keeps its degree.
  - This null model is the basic concept behind the definition ofmodularity, a functionwhich evaluates the goodness
  - of partitions of a network into communities.

Definitions Based on Vertex Similarity

- Based on an assumption that communities are groups of vertices which are similar to each other.
- Hierarchical clustering is a way to find several layers of communities that are composed of vertices similar to each other. Repetitive merges of similar vertices based on some quantitative similarity measures will generate a structure called dendograms.



Evaluating Communities

- we need a quality function for evaluating how good a partition is.
- The most popular quality function is the modularity of Newman and Girivan

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

- where the sum runs over all pairs of vertices,
- A is the adjacency matrix,
- ki is the degree of vertexi i
- m is the total number of edges of the network.
- The element of A*ij* of the adjacency matrix is 1 if vertices i and j are connected, otherwise it is 0.
- The delta-function yields 1 if vertices i and j are in the same community, 0 otherwise.

Methods for Community Detection

- Naive methods for dividing given networks into subnetworks, such as
  - graph partitioning, hierarchical clustering, and k-means clustering - needs to provide the numbers of clusters or their size in advance.
- The methods for detecting communities classified into the following categories:
  - (1) divisive algorithms,
  - (2) modularity optimization,
  - (3) spectral algorithms, and
  - (4) other algorithms.

Divisive Algorithms

- A simple way to identify communities in a network is to detect the edges that connect vertices of different communities and remove them, so that the communities get disconnected from each other.
- Algorithm is that proposed by Girvan and Newman.
- In this algorithm, edges are selected according to the values of measures of edge centrality, estimating the importance of edges according to some property on the network.
- The steps of the algorithm are as follows:
  - (1) Computation of the centrality of all edges,
  - (2) Removal of edge with largest centrality,
  - (3) Recalculation of centralities on the running network, and
  - (4) Iteration of the cycle from step (2).

Modularity Optimization

- Modularity is a quality function for evaluating partitions.
- The partition corresponding to its maximum value on a given network should be the best one.
- there are currently several algorithms that are able to find fairly good approximations of the modularity maximum in a reasonable time.
- One of the famous algorithms for modularity optimization is CNM algorithm proposed by Clauset et al.
- Another examples of the algorithms are greedy algorithms and simulated annealing.

Spectral Algorithms

- Spectral algorithms are to cut given network into pieces so that the number of edges to be cut will be minimized. One of the basic algorithm is spectral graph bipartitioning.
- In general, community detection based on repetative bipartitioning is relatively fast.

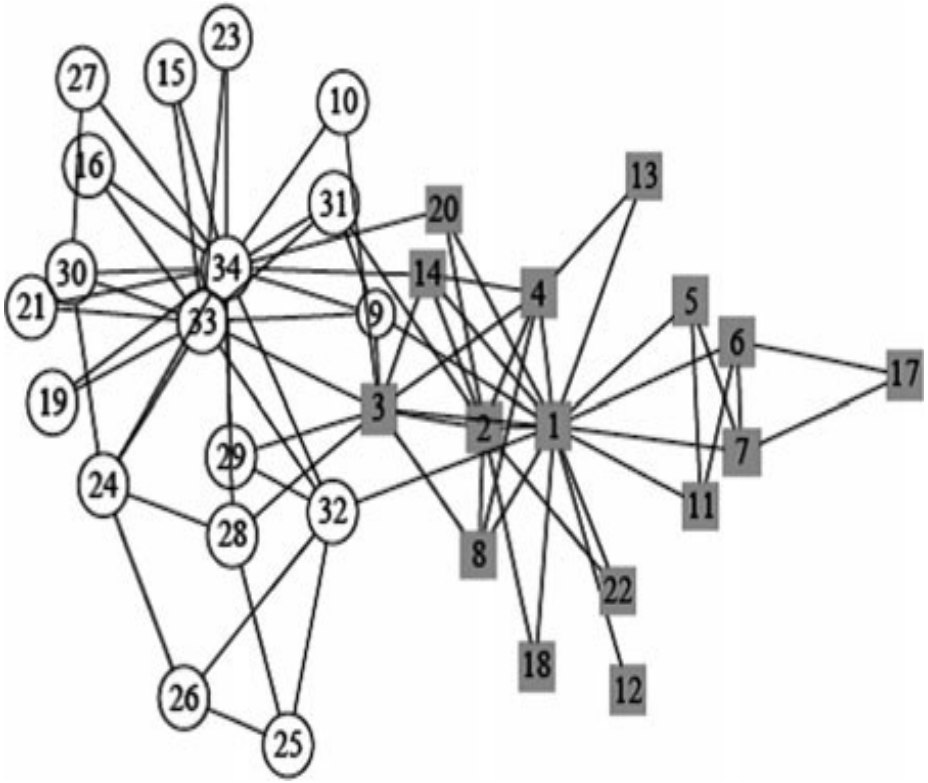
- Other algorithms

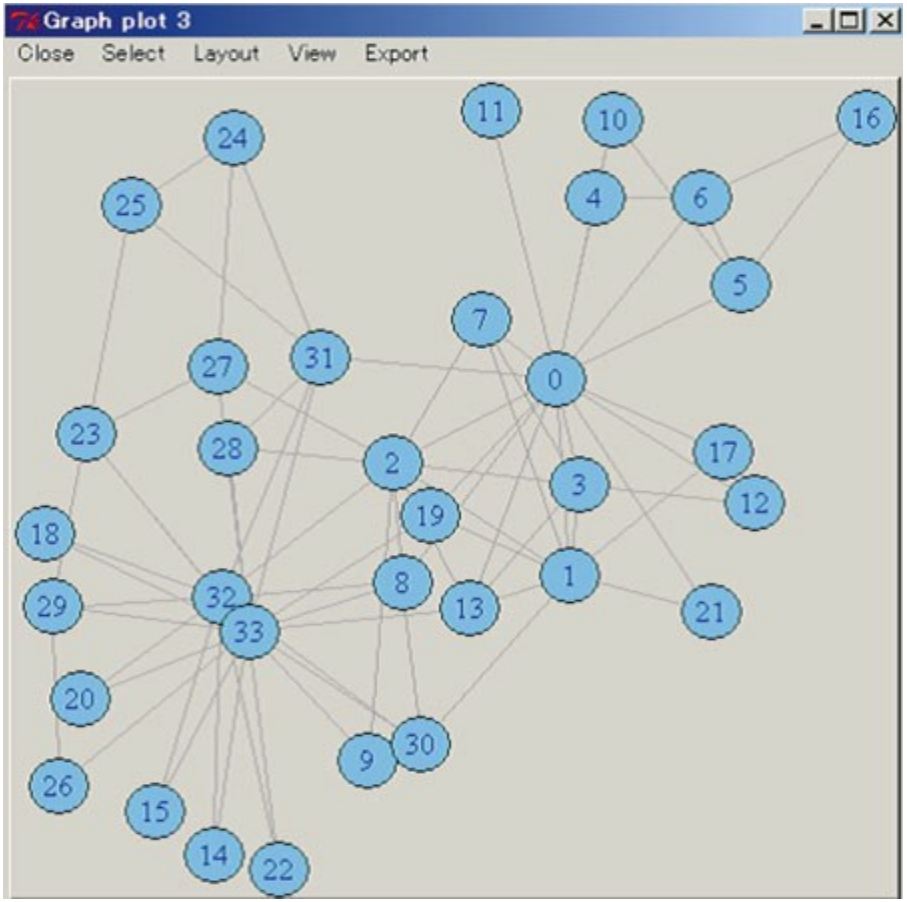
- Methods focusing on random walk, and the ones searching for overlapping cliques.

Tools for Detecting Communities

- Tools for Large-Scale Networks – CNM Algorithm – few million vertices
- Tools for Interactive Analysis
  - JUNG - (http://jung.sourceforge.net/)
  - Netminer - (http://www.netminer.com/NetMiner/overview 01.jsp)
  - Pajek - (http://vlado.fmf.unilj.si/pub/networks/pajek/)
  - igraph - (http://igraph.sourceforge.net/)
  - SONIVIS - (http://www.sonivis.org/)
  - Commetrix - (http://www.commetrix.de/)
  - NetworkWorkbench - (http://nwb.slis.indiana.edu/)
  - visone - (http://visone.info/)
  - Cfinder - (http://www.cfinder.org/)

Examples





Karate club network by two different tools