Detecting Communities in Social Network

Dr. V.S. Felix Enigo, DCSE, SSNCE

Introduction

- Importance of Detecting communities in social networks:
- Used for collaborative filtering in recommendation –members have similar tastes and preferences
- Understand the structures of given social networks – functions and properties of network
- Visualize large-scale social networks - information sharing and diffusions, growth

Definition of Community

- Community – sub-network with denser intra-community edges than inter-community edges
- Definitions of community can be classified as:

- Local definitions

- Global definitions

- Based on vertex similarity

Local Definitions

- Focused on vertices of subnetwork under investigation and 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

Contd...

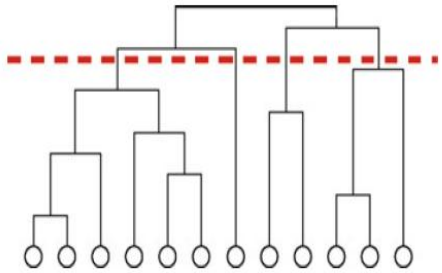
- Comparative ones - compares mutual connections of vertices of subnetwork with connections with external neighbors
- LS set - a subnetwork where each vertex has more neighbors inside than outside of the subnetwork
- Weak community - the total degrees of vertices inside community > number of edges lying between the community and the rest of network

Global Definitions

- It characterizes a subnetwork with respect to the network as a whole
- It starts from a null model (Newman and Girvan)
- Null model – a network that matches original network in some topological features, but no community structure
- To design a null model - introduce randomness in the distribution of edges among vertices of original network
- Link properties of subnetworks to the original network, If wide difference wrt subnetwork, then it is a community
- Null model is a way to evaluate goodness of partition of network into community (Modularity)

Definitions Based on Vertex Similarity

- Based on assumptions that communities are groups of vertices which are similar to each other
- Similarity between each pair of vertices done quantitatively
- In Hierarchical clustering - layers of communities composed of vertices similar to each other
- Ex. dendrogram, highly similar vertices found in lower part
- Subtrees got by cutting dendrogram with horizontal line correspond to communities
- Communities of different granularity got by changing position of horizontal line



Evaluating Communities

- Many ways to partition network into communities
- A quality function needed to evaluate goodness of a partition
- Modularity is the quality function proposed by Newman and Girivan:

$$Q = \sum_{s=1}^{n_m} \left[\frac{l_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right]$$

- n_m is the number of communities
- l_s is the total number of edges joining vertices of community s
- d_s is the sum of the degrees of the vertices of s
- upper term in each summand represents fraction of edges of network inside community
- lower term represents the expected fraction of edges in random network with same degree for each vertex (null model)
- i.e. comparison between real and expected edges

Contd...

- Newman and Girivan formula implies:
- Subnetwork is a community, if no.edges inside > expected number in modularity's null model (if true more tightly connected)
- Large +ve Q indicates good partitions
- Modularity of whole network, taken as single community, is zero
- Modularity is always smaller than one, but can be –ve
- Modularity optimization is a popular method for community detection
- Modularity optimization fails for communities smaller than scale (depends on size of network and resolution limit – degree of interconnectedness of community)

Methods for Community Detection

- Naive methods - graph partitioning, hierarchical clustering, and k-means clustering (no. of clusters or their size given in advance)

- Methods for detecting communities:
- Divisive algorithms
- Modularity optimization
- Spectral algorithms

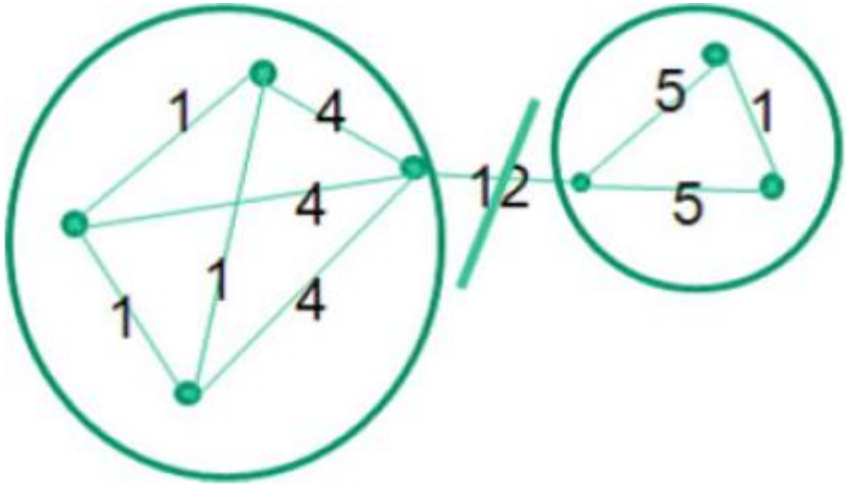
Divisive Algorithms

- To identify communities in a network - detect the edges that connect vertices of different communities and remove them
- Girvan and Newman proposed community detection algorithm based on edge betweenness centrality
- Edge betweenness centrality is defined as the number of the shortest paths that go through an edge in a network
- Steps in the algorithm
- (1) Computation of the centrality of all edges,
- (2) Removal of edge with largest centrality
- (3) Recalculation of centralities on the running network
- (4) Iteration of the cycle from step (2).

Contd...

- Intercommunity edges has larger edge betweenness value, as it serves as shortest path for many communities

- Ex. Edge 12 in the below figure



Modularity Optimization

- Numerous way partition a network can be done, so need best modularity optimization Q
- As NP hard problem, approximations algorithms that produce result in reasonable time is used
- Most popular modularity optimization is CNM algorithm
- Others greedy algorithms and simulated annealing

Spectral Algorithms

- Spectral algorithms cuts given network into pieces so that the number of edges to be cut will be minimized
- Basic algorithm is spectral graph bi-partitioning
- Laplacian matrix L of given network nxn symmetric matrix is used
- Laplacian matrix is defined as L=D-A
- where A is the adjacency matrix
- D is the diagonal degree matrix
- All eigen values of L are real and non-negative
- L has a full set of n real and orthogonal eigenvectors

- To minimize the above cut, vertices are partitioned based on the signs of the eigenvector that corresponds to the second smallest eigen value of L

- Community detection based on repetitive bi-partitioning is relatively fast.

Other Algorithms & Tools

- Random walk, and the ones searching for overlapping cliques
- Tools for large scale networks:
  - CNM algorithm of community detection based on modularity optimization
  - Works for few million vertices
- Tools for Interactive Analysis:
- JUNG, Netminer, Pajek, igraph, SONIVIS, Commetrix, NetworkWorkbench, visone, Cfinder etc.