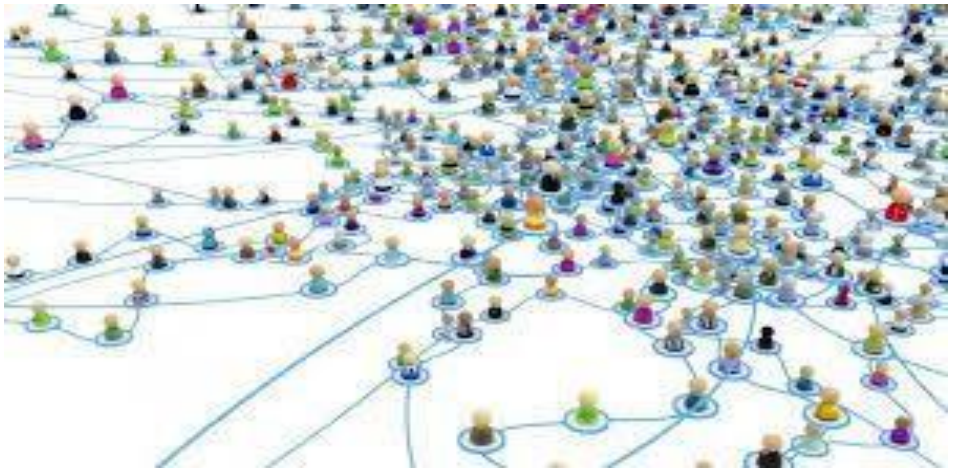
Social Network Analysis

Community Detection in SN



Course Instructor: Dr.V.S.Felix Enigo

Social Network Analysis

- Anetwork community mining problem (NCMP) is finding communities in a given network

Example:

- VLSI circuit board - processes frequently communicate with each other
- Image Segmentation - Segments of an image
- Web Page Clustering – Web pages related to common topics ……..…

- Allows to discover hidden patterns
- Enables to understand the structural and/or functional characteristics of networks to efficiently utilize them

Social Network Analysis

Community mining algorithms classified into two main categories:

- *Optimization based algorithms*
- *Heuristic based algorithms*

Optimization based algorithms – transforms NCMP to optimization problem and tries to find an optimal solution WRT a pre-defined objective function

Objective function can be:

- various cut criteria adopted by Spectral methods
- evaluation function introduced by the Kernighan–Lin algorithm
- network modularity

Social Network Analysis

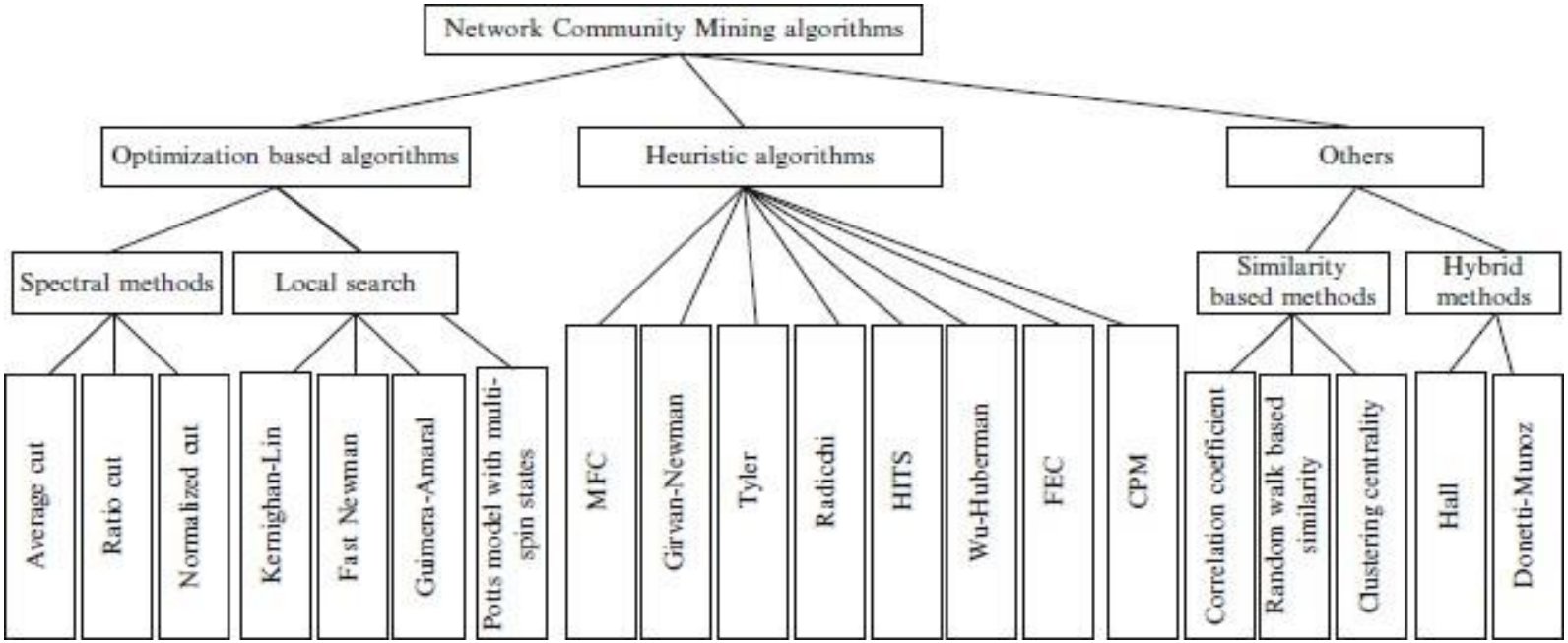
- Heuristic Algorithms solve an NCMP based on certain intuitive assumptions or heuristic rules.

Example:

- Maximum flow community (MFC) algorithm assumes that “flows” through inter-community links should be larger than those of intra-community links

- Heuristic rule in GN algorithm states that the ***edge betweenness*** of inter-community links should be larger than that of intra-community links

Social Network Analysis



Social Network Analysis

Optimization Based Algorithms

Two Broad Classes: Spectral methods and local search based methods

Spectral methods optimize certain pre-defined cut criteria using the quadratic optimization technique

Cut of a bipartition of a network is number of inter-group links

An optimal bipartition of a network is the one with the minimum cut but leads to bias partition as it counts the number of inter-group link

In order to avoid this problem, other criteria such as *average cut*, *ratio cut*, *normalized cut*, and their variants are used to compute the density

Social Network Analysis

Problems of finding different optimal cuts have been proven to be NP-complete

Spectral methods finds approximately optimal cut by transforming the problem into a constraint quadratic optimization problem represented as $min(X^T MX) / (X^T X)$

Here X denotes the indicator vector of a bipartition

Minimizing the **Average cut**, M corresponds to the Laplacian matrix of a given graph

The **Ratio cut, Normalized cut, or others** a variant of Laplacian matrix

Optimal solution is found out by computing the second smallest eigenvector of M

Social Network Analysis

- Spectral graph theory, looks at the eigenvalues of the graph Laplacian, to say whether a graph is connected and also how well it's connected

- The graph Laplacian is the matrix $L = D - A$

where D is the diagonal matrix whose entries are the degrees of each node
A is the adjacency matrix

- The smallest eigenvalue of L, λ_1, is always 0, if the graph has two disconnected components
- The second smallest eigenvalue λ_2 tells you about how well a **graph is connected**.

- if λ_2 is **small**, this suggests the graph is **nearly** disconnected, that it has two components that are not very connected to each other

Social Network Analysis

- Spectral methods are bipartition methods that try to split a graph into two, with a balanced size and the minimum cut

- If a network contains multiple communities, one can find all of the communities with a hierarchical structure in a recursive way until a pre-defined stopping criterion is satisfied

Social Network Analysis

Local search based optimization

- Kernighan–Lin algorithm
- Fast Newman algorithm
- Guimera–Amaral algorithm

- Adopt similar idea in finding a neighbour of the current solution in the problem space during each iteration

- Differs in optimization objectives and different strategies for regulating the local search

Social Network Analysis

The Kernighan-Lin algorithm (or KL)

- Aims to minimize an evaluation function defined as the difference of the numbers of intra-community links and inter-community links
- In each iteration, KL moves or swaps nodes between communities in order to decrease the evaluation function
- Iterative process stops when the evaluation function remains unchanged

Social Network Analysis

- KL runs moderately fast with the time complexity of $O(n^2)$
- During the local search, KL only accepts better neighbor solutions and rejects all worse ones, finding local optimal than global
- KL needs prior knowledge such as the number and average size of, communities to generate initial partition.
- KL is also sensitive to initial partitions bad one causes slow convergence and poor solution

Social Network Analysis

Faster Newman algorithm (or FN)

- Proposed by Newman for detecting community structures with the time complexity of O(mn)
- FN is also a local search based optimization method.
- Starting from an initial state in which each community only contains a single node
- FN repeatedly joins communities together in pairs by choosing the best merge, until only one community is left
- In this bottom-up way, the *dendrogram* of community structure is constructed

Social Network Analysis

- To choose best merge in each iteration, a new metric *modularity* is used
- Modularity quantitatively measures how well-formed a community structure is
- The modularity of a given network in terms of a Q-function is defined as follows:

$$Q = \sum_i e_{ii} - a_i^2$$

- e ij is the weighted links that connects the nodes in Community ***i*** to nodes in Community *j and* $a_i = \sum_j e_{ij}$.
- Large Q-values is better partition

Social Network Analysis

Guimera and Amaral Algorithm (or GA for short)

- Finds a partition of a network with the maximum modularity
- GA adopts simulated annealing (SA) to regulate the local search processing order to obtain a better solution
- Simulated Annealing is **a stochastic global search optimization algorithm**.
- It makes use of randomness as part of the search process
- This makes the algorithm appropriate for nonlinear objective functions where other local search algorithms do not operate well.
- Starting from an initial partition GA generates, evaluates, accepts or rejects a new neighbor partition from the current one in each iteration

Social Network Analysis

- To generate a new neighbor partition, GA moves or swaps nodes between groups, divides a group or merges two groups

- GA evaluates the new partition by calculating its modularity and decides whether or not to accept it by using the *metropolis* criterion based on the current system temperature

$$p = \begin{cases} 1, & C_{t+1} \le C_t \\ e^{-(C_{t+1}-C_t)/T}, & C_{t+1} > C_t \end{cases}$$

$C_t = -Q_t$ where p is the probability of accepting the solution at time t + 1

T is the system temperature at time t+1

Social Network Analysis

- GA offers good performance as it finds globally optimal solution
- Efficiency of GA depends by the convergent speed of SA which is slow and sensitive to initial parameters
- Parameters are: initial layout, the strategies of finding a neighbor solution, and the cooling system temperature
- GA outputs a partition of a network without a hierarchical structure, and does not require prior knowledge (no. of communities)

Social Network Analysis

Potts Model

Proposed by Reichardt and Bornholdt

Network is considered as a multiple- state Potts model, in which each node is a spin with q values

Best network partition corresponds to the most stable state of the Potts model (minimum energy)

In the stable state the node that spins with the same values constitute one community

Distribution of spin values is found by minimizing a pre-defined energy function using a Monte Carlo optimization method + simulated annealing algorithm

Social Network Analysis

Heuristic Methods

- Maximum flow community (MFC) algorithm
- Girvan–Newman algorithm (GN)
- Hyperlink Induced Topic Search algorithm (HITS)
- Wu–Huberman algorithm (WH)
- Clique Percolation Method (CPM)

Social Network Analysis

Maximum Flow Community Algorithm

MFC algorithm was proposed by Flake et al

It is based on the Max Flow-Min Cut theorem in graph theory

Idea of MFC is maximum flow through a given network is decided only by the capacity of network ***bottle- necks***

Bottlenecks refers to the capacity of the Min-Cut sets and the sparse inter-community links

Inter-community links is discovered by calculating the Min-Cut sets

By iteratively removing ***bottle- necks*** links, involved communities will be gradually separated from each other

Social Network Analysis

Girivan Newman Algorithm

- GN algorithm detects all communities by recursively breaking inter-community links
- The heuristic rule introduced in GN is that the inter-community links are those with the maximum ***edge betweenness***
- GN algorithm is a hierarchical method and can produce a dendrogram of community structure in a top-down fashion
- Its time complexity is $O(m^2n)$ which makes it not suitable for large-scale networks

Social Network Analysis

- To speed up the basic GN algorithm, several improvements have been proposed
- Tyler et al. introduced a statistical technique into the basic GN algorithm
- Monte Carlo method is used to estimate an approximate edge betweenness value for a selected link set
- Improvement in speed is gained obtained at the price of a reduction in accuracy

Social Network Analysis

- Radicchi et al. defined a new metric, called *link clustering coefficient*, to replace edge betweenness as it is time-consuming
- Idea is inter-community links are unlikely to belong to a short loop, such as triangles and squares
- Heuristic rule defined is link clustering coefficient - the number of triangles or squares in which a link is involved
- In each iterative step, links with the minimum link clustering coefficient will be cut off
- The average time complexity for computing the link clustering coefficient of all links is O (M^3 / M^2) < for computing edge betweenness, which is O(mn)

Social Network Analysis

Hyperlink Induced Topic Search algorithm [**HITS]** (by Kleinberg)

- It aims to discover hyperlink- based Web communities (authority-hub communities from the Web)
- The basic assumption behind HITS is that there exist authorities and hubs on the Web
- Authorities are often pointed to by hubs that preferentially point to authorities
- Based on the mutually reinforcing relationship between authorities and hubs, an iterative method for inferring such authority-hub communities from the Web is developed
- HITS computes the principal eigenvectors of two special matrices in terms of the adjacency matrix of the Web
- Asearch engine based on HITS can return the most topic-related pages to users

Social Network Analysis

WH algorithm

Proposed by Wu and Huberman

A network is modelled as an electrical circuit by allocating one unit resistor on each link

Selects two nodes from two distinct communities as the positive and negative poles

The idea is that the resistance within communities will be much less than that between communities as intra-community links are much denser than inter-community links

Social Network Analysis

- Heuristic is voltage difference of distinct communities should be more significant
- WH algorithm separates the group with a high voltage and the group with a low voltage from a network
- The node sequence sorted by their respective voltage values to find two maximum gaps
- Determines the final division by considering the co-occurrence of nodes in such separated groups
- Algorithm is very fast with a linear time in terms of the size of a network
- WH algorithm depends heavily on its prior knowledge, which is hard
- It needs to identify two “poles” belonging to different communities, needs the approximate size of each community in order to find multiple communities

Social Network Analysis

CPM algorithm

- Proposed by Palla and his colleague to discover an overlapping community structure
- Network community is made of “adjacent” k cliques, which share at least k - 1 nodes with each other
- Heuristic is each clique uniquely belongs to one community, but cliques within different communities may share nodes
- CPM is able to find the overlaps of communities
- For a given K, CPM first locates all k cliques (k <= K) from a given network
- Build a clique-clique overlap matrix to find out communities in terms of different k.

Social Network Analysis

Other Methods

Clustering in bottom-up approach repetitively joining pairs of current groups based on their similarities

Similarities are computed based on correlation coefficients and random walk similarities in terms of linkage relations

Proposed by Hall - Transforming an NCMP into a clustering problem in a vector space

Allocating coordinate to each node, and then cluster such spatial points using any typical spatial clustering algorithm, such as k-mean

Social Network Analysis

- Donetti and Munoz proposed a method for solving NCMPs based on quite a similar idea

- Mapping a network into a k-dimensional vector space using the k smallest eigenvectors of the Laplacian matrix before clustering spatial points

Social Network Analysis

Summary

- Community detection allows to discover hidden patterns, understand structural and/or functional properties of a network

Community mining algorithms classified into two main categories:

- *Optimization based algorithms*
- *Heuristic based algorithms*

Optimization based algorithms – transforms NCMP to optimization problem and finds an optimal solution wrt pre-defined objective function

Heuristic Algorithms solve an NCMP based on certain intuitive assumptions or heuristic rules.