FINDING COMMUNITY STRUCTURE IN SOCIAL NETWORK ANALYSIS - OVERVIEW

Jan Rupnik
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773419; fax: +386 1 4251038
e-mail: jan.rupnik@ijs.si

ABSTRACT

This paper is an overview of some of the basic concepts in the theory of social network analysis. We start with defining social networks and basic structures such as walks and components. We introduce different measures of centrality, prestige, and a measure of modularity used in detecting community structure. We conclude with an overview of different approaches to finding communities in networks.

1 INTRODUCTION

A social network is a set of people or groups each of which has connections of some kind to some or all of the others. In the language of social network analysis, the people or groups are called ‘‘actors’’ and the connections ‘‘ties’’. Both actors and ties can be defined in different ways depending on the questions of interest. An actor might be a single person, a team, or a company. A tie might be a friendship between two people, a collaboration or common member between two teams, or a business relationship between companies. The realationships between actors can be symmetric (for example: communicates with) or directional (for example: lent money to). The corresponding mathematical models are un-drirected graph and directed graph.

2 COLLABORATION NETWORKS

A common type of social networks are the affiation networks. An affiliation network is a network of actors connected by common membership in groups of some sort, such as clubs, teams, or organizations. They are suitable for statistical analysis since membership to a group can usually be determined precisely. They can get very large, since they can be extracted automatically. interviews or questionnaires unnecessary. A network of movie actors, for example, and the movies in which they appear has been compiled using the resources of the Internet Movie Database , and contains the names of nearly half a million actors.

A good example of an affiliation network is the scientific collaboration network in which the link between two scientists is established by their coauthorship of one or meny articles. If a paper has k coauthors we get a k-clique. groups of coauthors of a single paper.
This network is in some ways more truly a social network than many affiliation networks; it is probably fair to say that most pairs of people who have written a paper together are genuinely acquainted with one another, in a way that movie actors who appeared together in a movie may not be.

3 WALKS, TRAILS, PATHS

A **walk** is a sequence of actors and relations that begins and ends with actors. A *closed walk* is one where the beginning and end point of the walk are the same actor. Process of the monetary exchange is an example where a dollar bill travels from one person to another with no limitations.

A **trail** between two actors is any walk that includes a given relation no more than once (the same other actors, however, can be part of a trail multiple times. The length of a trail is the number of relations in it. An example of a trail is the gossip process, where story is moving through an informal network. It never travels between the same pair of actors, but can reach the same actor multiple times.

A **path** is a walk in which each other actor and each other relation in the graph may be used at most one time. The single exception to this is a closed path, which beginsand ends with the same actor. Length of a path is the number of it’s links. Two paths are independent if they share no nodes (except beginning and the end). They are line independent if they share no lines. An example of a path is the viral infection which activates effective immunological response (every actor is infected at most once).

A **chain** is a walk in a directed graph, that ignores is not restricted by direction of edges.

4 COMPONENTS

A subset of vertices in a network is called ***a strongly connected component*** if (taking directions of lines into account) from every vertex of the subset we can reach every other vertex belonging to the same subset. Between any two vertices from the same connected component there always exists a walk. We say that they are strongly connected.

If direction of lines is not important (where we consider the network to be undirected), such a subset is called a ***weakly connected component***. Between any two vertices from the same weakly connected component there always exists a chain. We say that the two vertices are weakly connected.

5 COMMON MEASURES

Distance
Geodesic distance between two nodes is the length of the shortest path between them. Diameter of a network is the maximum of all possible distances in that network.

Centrality
The centrality of a node in a network is a measure of the structural importance of the node. A person's centrality in a social network affects the opportunities and constraints that they face. There are three important aspects of centrality: degree, closeness, and betweenness.

- **Degree** - Degree centrality is the number of nodes that a given node is connected to. In general, the greater a person's degree, the more potential influence they have on the network, and vice-versa.

- **Closeness** - Closeness centrality is defined as the total graph-theoretic distance to all other nodes in the network. When a node has a low closeness score (i.e., is highly central), it tends to receive anything flowing through the network very quickly.

- **Betweenness** - A node highly is central, if it lies on several shortest paths among other pairs of nodes. More precisely, if g_{ij} is the number of geodesic paths from i to j and g_{ikj} is the number of paths from i to j that pass through k, then g_{ikj}/g_{ij} is the proportion of geodesic paths from i to j that pass through k. The sum $c_k = \sum g_{ikj}/g_{ij}$ for all i,j pairs is betweenness centrality of node k.

Prestige measures

Prestige measures are usually computed for directed networks only, since for this measures the direction is important property of the relation. (Example: Persons, who are chosen as friends by many others have a special position - prestige in the group.) Prestige becomes salient especially if positive choices are not reciprocated, for instance if people tend to express positive sentiments towards prestigious persons but receive negative sentiments in return. In these cases, social prestige is connected to social power and the privilege not to reciprocate choices.

- **Input degree -** According to meaning of relation they represent support or influence.Degree is a very restricted measure of prestige because it takes into account direct choices only.

- **Influence domain -** The influence domain of a vertex in a directed network is the number or proportion of all other vertices which are connected by a path to this vertex. It is an extension of prestige to indirect choices. This measure is usefull for networks that are not strongly connected (in which case input closeness centrality is used) .

- **Mean distance -** Computing the mean distance of a vertex from vertices in it's influence domain gives a numeric index, so that direct choices contribute more to prestige than indirect. (Problem: vertices with small influence domains can have low average distance, but are not very influential)

- **Proximity prestige -** We compute proximity prestige (PP) of selected vertex by dividing the influence domain of a vertex (expressed as a proportion) by the average distance from all vertices in the influence domain. A larger influence domain and a smaller distance yield a higher proximity prestige score.

- **Hubs and authorities –** A measure of prestige commonly used for web graphs (directed network of homepages). In directed networks we can usually identify two types of important vertices: hubs and authorities. Each home page describes something (is an authority) and because of that other pages point to it. But on the other hand each page points to some other pages (is a hub). Vertex is a good hub, if it points to many good authorities, and is a good authority, if it is pointed to by many good hubs. For each vertex v we compute weights $0 \leq x_v, y_v \leq 1$, which tell us the strength of authority and hub of a given vertex. Weights are computed according to network by solving the eigenvector problem of matrices AA^T

(hubs) and AA^T (authorities), where A is the adjacency matrix.

Reach
Given an integer k, reach of node v is defined as number of nodes with the length of the shortest path to v less or equal to k. Some studies have shown that key paths in most social networks are one or two (rarely three) steps long.

Network Centralization
Network centralization is the variance of computed node centralities. Network, where a low number of nodes have much higher centrality than other nodes is highly centralized. A very centralized network is dominated by one or a few very central nodes. If these nodes are removed or damaged, the network quickly fragments into unconnected sub-networks. A highly central node can become a single point of failure. A network centralized around a well connected hub can fail abruptly if that hub is disabled or removed. Hubs are nodes with high degree and betweenness centrality.

Clustering coefficient
The clustering coefficient is a measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'. Given a node v, it is calculated as the proportion between all triangles that contain v and all connected triplets (with v in the middle) in the network. Network clustering coefficient is defined as the arithmetic mean of clustering coefficients of every node.

Network cohesion
Refers to the degree to which actors are connected directly to each other by cohesive bonds. Groups are identified as ‘cliques’ if every actor is directly tied to every other actor, or ‘social circles’ if there is less stringency of direct contact. Structural cohesion level k corresponds to the mathematical term k-connectivity. A graph is k-connected if there exists a set of k vertices whose removal renders G disconnected, and which can not be achieved with a smaller set.

Modularity
Modularity is a property of a network and a specific proposed division of that network into communities. It measures when the division is a good one, in the sense that there are many edges within communities and only a few between them. For a division with g groups, we define a $g \cdot g$ matrix e whose component e_{ij} is the fraction of edges in the original network that connect vertices in group i to those in group j. Then the modularity is defined to be

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki} = \mathrm{Tr}\,\mathbf{e} - \left\| \mathbf{e}^2 \right\|,$$

where $||x||$ is the sum of all the elements of x. A value of $Q = 0$ indicates that the community structure is no stronger than would be expected by random chance and values other than zero represent deviations from randomness.

6 COMMUNITY STRUCTURE

Many networks are inhomogeneous, consisting not of an undifferentiated mass of vertices, but of distinct groups. Within these groups there are many edges between vertices, but between groups there are fewer edges, producing a structure like that sketched in Fig. 1. Such groups are called communities. Communities in a web graph for instance might correspond to sets of web sites dealing with related topics, while communities in a biochemical network or an electronic circuit might correspond to functional units of some kind.

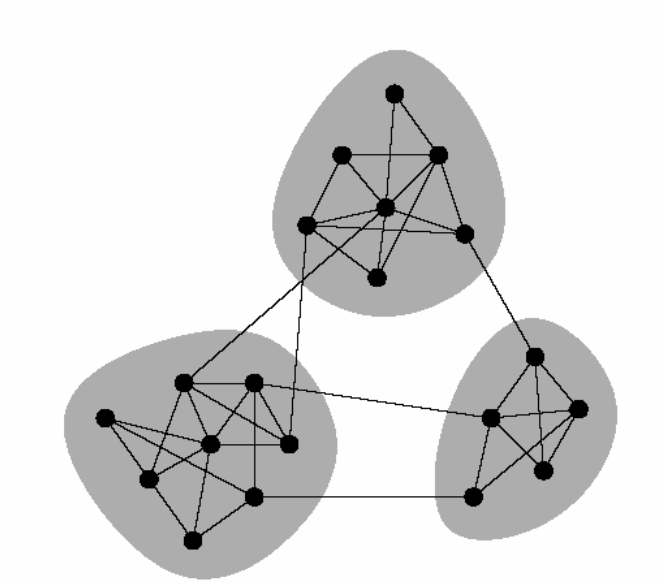

Figure 1: *Small network with strong community structure. The network breaks into three communities.*

7 ALGORITHMS FOR DETECTING COMMUNITY STRUCTURE

7.1 DIVISIVE METHOD BASED ALGORITHMS

Girvan Newman algorithm [3]
Algorithm progressively removes edges with highest edge betweenness from the original graph. If a network contains communities or groups that are only loosely connected by a few inter-group edges, then all shortest paths between different communities must go along one of these few edges. Thus, the edges connecting communities will have

high edge betweenness. By removing these edges, we separate groups from one another and so reveal the underlying community structure of the graph. After removing an edge the algorithm recalculates betweenness centralities for every node which can be done in $O(n^2)$ on sparse networks (See Brandes [8]).

The algorithm of Radicchi et al. [4]
Algorithm removes edges that belong to a relatively low number of loops, for they are likely to be edges between communities.

7.2 AGGLOMERATIVE HIERARCHICAL CLUSTERING METHOD BASED ALGORITHMS

Modularity optimization algorithm [2]
Starting with a state in which each vertex is the sole member of one of n communities, we repeatedly join communities together in pairs, choosing at each step the join that results in the greatest increase (or smallest decrease) in modularity. This method can be applied to very large networks.

Single linkage methods
The idea behind this technique is to develop a measure of similarity between pairs of vertices, based on the network structure one is given. Many different such similarity measures are possible. Once one has such a measure then, starting with an empty network of n vertices and no edges, one adds edges between pairs of vertices in order of decreasing similarity, starting with the pair with strongest similarity. Structural equivalence is an example of a similarity measure. Two vertices are said to be structurally equivalent if they have the same set of neighbors.

7.3 OTHER METHODS

Spectral bisection algorithm [5]
A method based on the eigendecomposition of the Laplacian matrix (identity matrix minus adjacency matrix). Eigenvector, corresponding to the second lowest eigenvalue, determines a partition of nodes into two communities.

8 CONCLUSION

We have reviewed some of the basic indices for social network analysis and illustrated how we can apply some of them in more complex algorithms (GN algorithm, for example).

Acknowledgement
This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under NeOn (IST-4-027595-IP) and PASCAL Network of Excellence (IST-2002-506778). This publication only reflects the authors' views.

References

[1] Stanley Wasserman, Katherine Faust, Social Network Analysis: Methods and Applications, *United Kingdom: Cambridge University Press, 1994*
[2] Aaron Clauset, M. E. J. Newman, and Cristopher Moore: Finding community structure in very large networks*, Phys. Rev. E 70, 066111 (2004)*
[3] M. Girvan and M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA 99, 7821{7826 (2002).*
[4] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Defining and identifying communities in networks. *Preprint cond-mat/0309488 (2003)*
[5] A. Pothen, H. Simon, and K.-P. Liou, Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl. 11, 430-452 (1990).*
[6] M. E. J. Newman, The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci. USA 98, 404-409 (2001)*
[7] R. Hanneman, Introduction to Social Network Methods. *University of California, Riverside: Department of Sociology, 2001.*
[8] Ulrik Brandes:A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology 25(2):163-177, 2001.*