Social Network Analytics

Week 04

Shivraj Kanungo

George Washington University

February 3, 2016

Outline

Network community

- A community is most often defined as a group of individuals living in the same geographical location.
- It can also be used to describe a group of people with a shared characteristic or common interest: a research community, conservative or liberal bloggers on the web, etc.
- There is also the approach that views communities as something socially and symbolically constructed, resting on a shared understanding - Benedict Anderson defined the nation state as an "imagined community" (1983).
- Using social networks analysis we define communities differently - by looking at how people are connected to each other, and clustering these into similar groups.

What is a cohesive subgroup?

- Cohesive subgroups have always represented an important construct for sociologists who study individuals and organizations.
- The interest may arise from the need to know how subgroups are formed
- Although cohesion and homogeneity are continuously sought as criteria to identify subgroups they are not observed all that frequently
- In cohesive subgroups actors (or nodes) tend to be connected via many direct, reciprocated choice relations, and who share information, achieve homogeneity of thoughts and behavior, and tend to act collectively.

The idea behind subgroups

- A cohesive subgroup is a subset of a network that displays some minimal level of cohesion
- Underlying idea behind the formation for subgroups is cohesion
- Understanding cohesion
  - Psychological concept: people's feelings of belonging to a group
  - Network view: ties linking actors / nodes
- Combine the two ideas ... and you tend to end up with actors in a network who are similar along some dimensions or have some underlying reasons to group together

Network theories that go with cohesive subgroups

- Homophily
  - a tendency of nodes to become connected to other nodes that are similar under a certain definition of similarity.
- Embeddedness
  - is a multidimensional construct relating generally to the importance of social networks for action. Embeddedness indicates that actors who are integrated in dense clusters or multiplex relations of social networks face different sets of resources and constraints than those who are not embedded in such networks

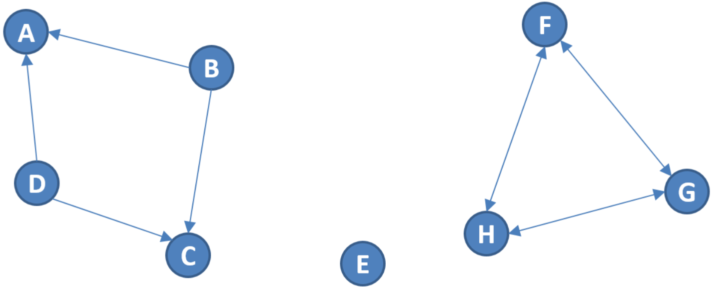
Components

- From a network perspective, a component is considered to the minimum requirement for a cohesive subgroup
- A component consists of a subgroup of individuals, whereby all the individuals are connected to one another by at least one path.
- Components can be weak or strong.

Components

- Weak components
  - Nodes connected to one another regardless the direction of ties
- Strong components
  - Nodes connected to one another through direct or indirect ties
- Isolates
  - Nodes not having ties to any other nodes

Three components



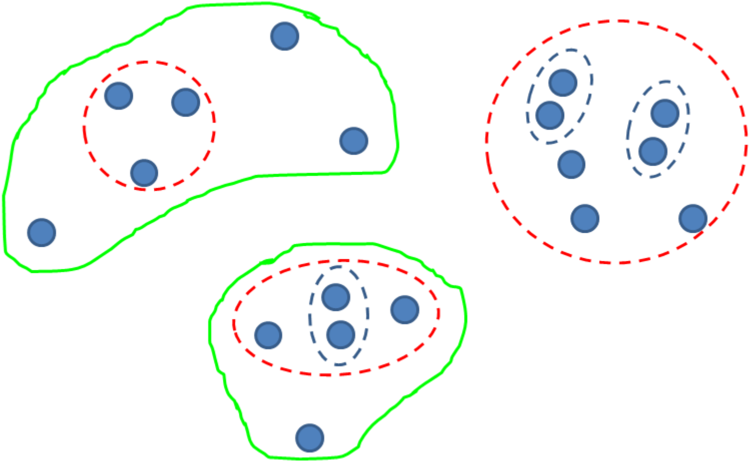
Component identification

- Component identification is important to decide the nature of subgroup analysis
- From the standpoint of cohesion
  - A network consisting of one component is seen as a connected graph
  - A network that is connected - but only when we ignore the direction of edges - is seen as a weakly connected graph
  - A network that is broken into several components is a disconnected graph

The idea of clustering

- Cluster analysis (or clustering) is the task of grouping a set of objects
- Such that
    - objects in the same group (called a cluster) are more similar (in some sense or another) to each other
    - Compared to those in other groups (clusters).

The process of clustering



Hierarchical Clustering example

- Dendrogram: cluster and leaves
- sample of children - their age, sex, height (inches), and weight (pounds)
- Different clusters based on
  - Sex
  - Sex and Age
  - Sex, Age, and Weight
  - Sex, Age, Weight, and Height

Age	Sex	Weight	Height
6	M	35	41
7	M	51	36
6	M	52	43
5	M	58	44
6	M	36	35
7	M	39	42
8	M	35	37
7	M	54	47
6	M	52	48
5	M	60	46
6	M	36	37
7	M	39	47
6	M	46	38
7	F	55	47
8	F	32	38
7	F	53	44
6	F	44	39
7	F	60	42
8	F	57	48
7	F	48	43
6	F	50	47
5	F	34	44
6	F	57	46
7	F	59	35

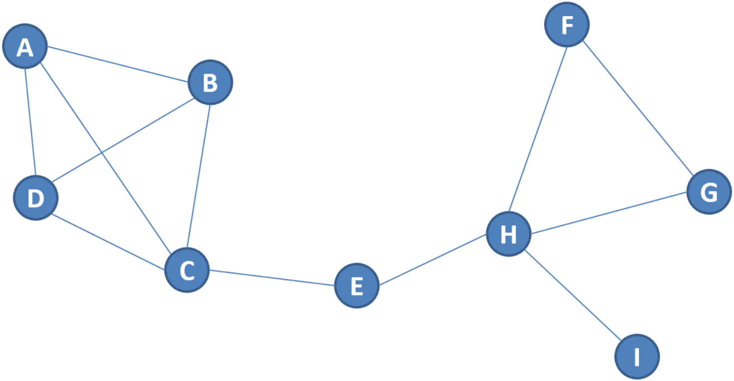
Cliques

- Simplest and oldest subgroup definition
- Early definition by Warner and Lunt (1941): cliques are informal groupings of people in which feelings of intimacy exist, and where the presence of particular group norms and sub-culture exist.
- Distinction of a clique from the larger network structure is based on cohesiveness

Network view of cliques

- Notion of a clique based on complete mutuality.
- Cliques are subgroups of people with reciprocal ties.
- Formal definition: a clique is a maximally complete subgraph of three or more nodes.
- Offers a clear and precise articulation of clique based on network structural features and does not depend on notions of culture, norms, intimacy (Wasserman & Faust, 1994)

What cliques look like



Practical guidelines for analyzing cliques

- Cliques are analyzed with graphs as opposed to digraphs
- Hence you should symmetrize your data before conducting clique analysis
- If you want to conduct clique analysis on digraphs, pay attention to directionality because only reciprocal ties are included
- Digraph has fewer cliques than a graph (simplified digraph)

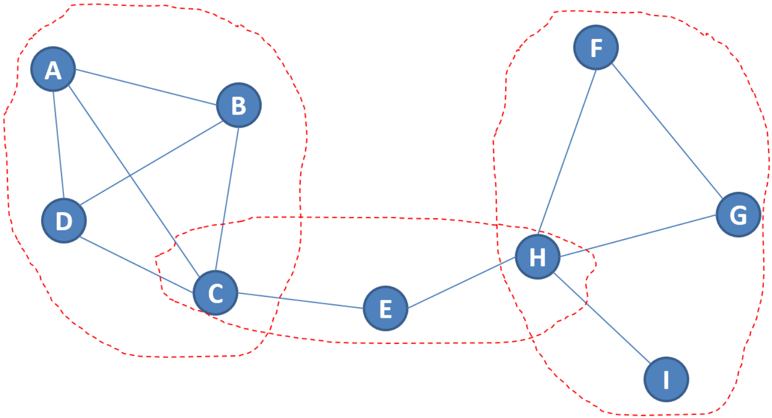
Potential issues in clique analysis

- Clique overlap
  - One or more actor(s) from one clique can be a member of another clique
  - Reflective of "social circles"
  - Bridging roles
- Reduced clarity in interpreting results
  - Works to hide underlying subgroup/clique structure
  - Analyzing clique overlap (Everett and Borgatti, 1998)
  - Alternate approach - Clique percolation

n-Cliques

- An n-clique is a subgroup in which every pair of actors is connected by a path of n or less.
- n refers to the length of the path connecting one actor to another
- Traditional clique analysis would imply that everyone in a clique are friends.
- However in a 2-clique (or a dyad) this definition would be expanded to say "everyone on this clique is a friend of a friend"
- Implication: the rules by which one would "join" a clique are relaxed.

Two-clique



n-Cliques

- Use of n-cliques makes sense when the idea is to understand how information flows through intermediaries.
- This is because the identification of n-cliques is based on paths and not on being part of a subgroup.

Issues in n-Cliques

- An n-clique higher than three may not be particularly useful
- A three-clique implies "a friend of a friend of a friend"
- As n goes beyond three, we start moving away from the sociological meaning of a clique
- If direct interaction is considered important for determining cohesion in a group, the n-clique approach is not advised.

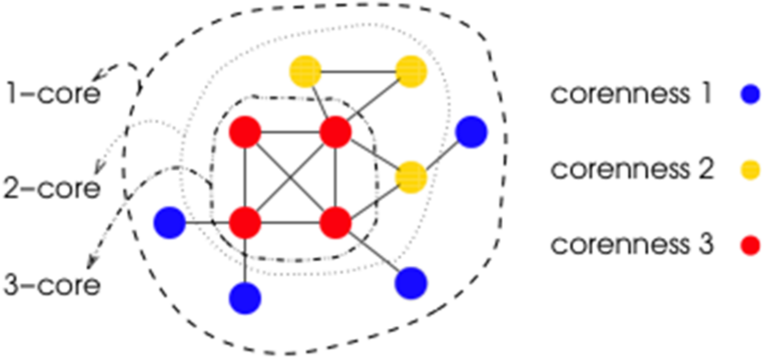
k-Cores

- When direct interaction is considered important for determining cohesion in a group, the k-core method is more applicable.
- The k-core approach is based on direct ties between actors.
- k-cores are not "cohesive" on their own.
- k-cores contain cohesive subgroups.
- They build on the concept of degree centrality.

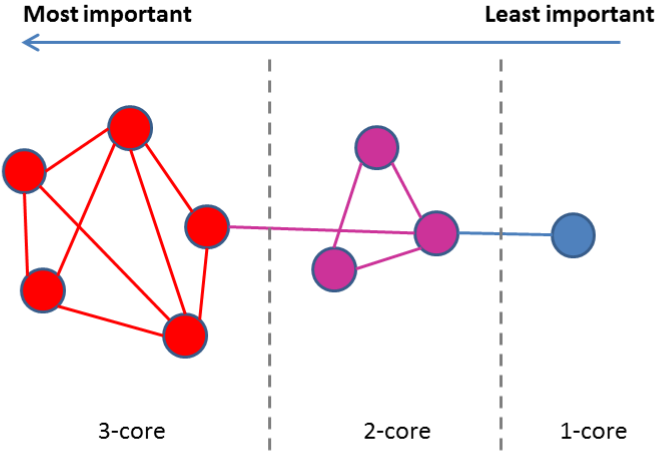
Determining k-cores

- An actor is part of a k-core if they have at least a degree centrality of k in that subgroup.
- An actor is part of a k-core if they are part of k members of that subgroup
  - As k$\downarrow$, sugroup sizes $\uparrow$ (it becomes easier to draw boundaries around the subgroups)
  - As k$\uparrow$, it becomes harder for an actor to join a group
- Implication: k-cores are nested
  - If you are member of a four-core, you are also member of a three-core, and a two-core

Visualizing k-cores



K-cores example



Lambda Sets

- Cohesive subgroups can also be identified by contrasting ties within the group to those outside the group
- The idea is that if a subgroup is cohesive, it should be difficult to difficult to break apart by removing one or more of its ties
- If two actors are well connected to one another, then ideally there is more than one path connecting them.

Understanding Lambda Sets

- Two actors, i and j, are considered to be part of the same lambda set if the number of paths connecting i and j is larger than the number of paths connecting either i or j to an outside actor, say k.
- As the λ values increase, this restricts how many actors can join the group
- Lambda sets maximize line connectivity within the group and minimize line connectivity between any nodes within the group and nodes elsewhere on the graph. Line connectivity, λ(i,j) is the minimum number of connections that must be removed from the graph to leave no path between the two nodes

Girvan Newman algorithm

- Used to break down a complex network into clusters or communities
- Based on the concept of edge-betweenness or betweenness-centrality
- Betweenness score for an edge is the number of times that edge lies on a geodesic path between a pair of disconnected actors
- Based on edge removal

Essential idea

- Identify which edges lies between cohesive segments of a network
- Then remove those edges
- This results in a fragmentation process
- The edges that remain are those contained with a cohesive subgroup
- Repeat the process till the desired number of cohesive subgroups is found
- The extreme case in this iteration is that only isolates remain.
- Therefore, it makes sense to identify how many subgroups are needed.

Components

- Primarily observing whether actors are all connected by either weak or strong ties.
- You are looking for weak or strong components
- This is typically used as the starting point for our analysis of cohesion

Cliques

- When you are interested primarily in the mutuality of ties
- Also, when it is important that all actors in a subgroup have ties to each other
- A natural extension of this analysis is to look for clique overlaps, affiliation networks (2-mode networks), and valued data (not covered in this class)

k-cores

- If direct ties in the cohesive subgroup are important but the requirements of cliques are too stringent
- Here you define subgroup membership as being dependent on an actor being tied to a specified proportion of other subgroup members versus all subgroup members
- This is preferable if multiple redundant paths between group members is an important subgroup property.

n-cliques

- n-cliques are a more relaxed version of cliques
- If we are looking at how information can quickly reach others, this approach is preferable
- This is also preferable if our aim is to study the role of intermediaries in information flow in the group.

Lambda Sets

- If there is reason to believe that comparing the number of within subgroup ties to the number of ties outside the subgroup is important, we will employ lambda sets

Girvan-Newman algorithm

- When faced with large networks it may not be possible to make an initial determination of subgroups
- This is a useful alternative to clique analysis especially if clique analysis is not particularly helpful in identifying subgroups.

Bottom line

- This analysis is not normative
- Compare the results of different analyses
- Look for persistent patterns in data that are repetitively captured by different approaches
- If there are differences, they may point to issues that could shed more light on your data.

Doing a cohesive subgroup analysis
Borgatti

- Step 1: If data are binary go to step 2 or use MDS or hierarchical clustering
- Step 2: Establish the components (identify both strong and weak components). If you are able to do this, you are done. If not, proceed to step 3

Doing a cohesive subgroup analysis (cont)

- Step 3: If partition is required and you know the approx no of groups, go to step 6. Otherwise find all the cliques. If there are no or very few cliques, try this
  - If min clique size > 4, reduce it - do not go below 3
  - If data were dichotomized reduce cut off
- If all these fail go to step 5; if too many cliques found (in step 4) try reverse of options above. If you are ok with the cliques, stop. Else proceed to step 4

Doing a cohesive subgroup analysis (cont)

- Step 4: Analyze pattern of overlap - idea is to identify major groupings of cliques and actors (ousiders and possibly leaders and boundary spanners). This should be good enough to deduce the approximate number of subgroups - and then do a factions analysis (not covered). If we need to partition data go to step 6 else stop.
- Step 5: Apply the Girvan-Newman algorithm. If results satisfactory, stop else go to step 6.
- Step 6: Faction analysis (not covered)

Doing a cohesive subgroup analysis (cont)

- This study (Roethlisberger and Dickson, 1939) provides binary network data in a workplace.
- Six 14x14 matrices
  - RDGAM, participation in horseplay;
  - RDCON, participation in arguments about open windows;
  - RDPOS, friendship;
  - RDNEG, antagonistic (negative) behavior;
  - RDHLP, helping others with work; and
  - RDJOB, the number of times workers traded job assignments.

BWR setting

The setting is 1931–1932 at an industrial plant in Chicago, the Western Electric Company's Hawthorne Plant, where a group of employees were observed continuously over a 6-month period while they wired, soldered, and inspected switchboards. The original study of these workers was designed to explore findings that had come to light in previous studies of the Hawthorne Plant. The investigators had noticed that "social groups in shop departments were capable of exercising very strong control over the work behavior of their individual members" (Roethlisberger and Dickson, 1939: 379). In some groups, the wage incentive systems appeared to be defeated by informal practices that restricted output and maintained a norm on what constituted a "day's work." To elucidate the formation of these norms, a small number of workers were selected, separated from the larger shop department of which they were a part, and relocated to a room — the Bank Wiring Observation Room — where they could be studied. The idea was that, over a 6-month observation period, the contours of the social structure of the work group would stabilize and the key social control processes would appear by which work performance norms are formed and maintained.

Ref: Friedkin, N. E. (2001) Norm formation in social influence networks, Social networks, 23, 167-189

BWR description - 1/2

The BWR had three work groups, each comprised of three wiremen connecting banks of terminals for telephone equipment and a solderman who soldered the wired connections. Two inspectors examined the constructed banks for quality control. Each wireman worked at two fixed adjacent benches and moved between them when a bank was completed. The soldering locations were also fixed. Wiremen and soldermen interacted in creating the equipment and both interacted with inspectors examining soldered banks. In response to these design features, sentiments towards others in the room developed forming part of the group's external system.

BWR description - 2/2

The organization forbade workers from helping each other and trading jobs, yet both occurred in the BWR as part of the internal system. In these instances, proscribed activities and interactions occurred as a consequence of sentiments that included boredom with tasks and a desire to work with specific others. Non-work related chatter while tasks were completed, playing games during work time and breaks, as well conflicts over windows (being open or closed) were parts of the internal system. Friendships and animosities also formed as a part this internal system. The internal and external systems are coupled (Homans, 1950, p. 91).

Homans G. (1950). The human group. New York: Harcourt-Brace.

Zachary Karate Club

Zachary observed 34 members of a karate club over a period of 2 years. During the course of the study, a disagreement developed between the administrator of the club and the club's instructor, which ultimately resulted in the instructor's leaving and starting a new club, taking about a half of the original club's members with him. Zachary constructed a network of friendships between members of the club, using a variety of measures to estimate the strength of ties between individuals.