UNIT 3 CLIQUES, CLUSTERS AND COMPONENTS 9 Hrs.

Components and Subgraphs: Sub graphs - Ego Networks, Triads, Cliques, Hierarchical Clustering, Triads, Network Density and conflict. Density: Egocentric and Sociocentric - Digression on Absolute Density – Community structure and Density, Centrality : Local and Global - Centralization and Graph Centres, Cliques and their intersections, Components and Citation Circles - Positions, Sets and Clusters.

Components and Subgraphs

A subgraph is a subset of the nodes of a network, and all of the edges linking these nodes. Any group of nodes can form a subgraph—and further down we will describe several interesting ways to use this.

Component subgraphs (or simply components) are portions of the network that are disconnected from each other. Before the meeting of Romeo and Juliet, the two families were quite separate (save for the conflict ties), and thus could be treated as components.

Many real networks (especially these collected with random sampling) have multiple components. One could argue that this is a sampling error (which is very possible)—but at the same time, it may just mean that the ties between components are outside of the scope of the sampling and may in fact be irrelevant.

Blocks and Cutpoints (Bi-components)

An alternative approach to finding the key "weak" spots in the graph is to ask: if a node were removed, would the structure become divided into un-connected parts? If there are such nodes, they are called "cutpoints." And, one can imagine that such cutpoints may be particularly important actors -- who may act as brokers among otherwise disconnected groups. The divisions into which cut-points divide a graph are called blocks. We can find the maximal non-separable sub-graphs (blocks) of a graph by locating the cutpoints. That is, we try to find the nodes that connects the graph (if there are any). Another name for a block is a "bi-component."

Factions

Imagine a society in which each person was closely tied to all others in their own sub-population (that is, all sub-populations are cliques), and there are no connections at all among sub-populations (that is, each sub-population is a component). Most real populations do not look like this, but the "ideal type" of complete connection within and complete disconnection between sub-groups is a useful reference point for assessing the degree of "factionalization" in a population.If we took all the members of each "faction" in this ideal-typical society, and put their rows and columns together in an adjacency matrix (i.e.

permuted the matrix), we would see a distinctive pattern of "1-blocks" and "0-blocks." All connections among actors within a faction would be present, all connections between actors in different factions would be absent.

Network Measurements

A number of measurable network characteristics were developed to gain a greater insight into networks, with many of them having their roots in social studies on the relationships among social actors. In this section, we will discuss three categories of measurements that have been defined in the social network analysis stream:

1. Network connection, which includes transitivity, multiplexity, homophily, dyads and mutuality, balance and triads, and reciprocity
2. Network distribution, which includes the distance between nodes, degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and density
3. Network segmentation, which includes cohesive subgroups, cliques, clustering coefficient, k-cores, core/periphery, block models, and hierarchical clustering

Network Connection

Network connection (or connectivity) refers to the ability to move from one node to another in a network. It is the ratio between route distance and geodesic distance. Connectivity can be calculated locally (for a part of the network) and globally (for the entire network). Let’s take a look at some of the important metrics of network connection.

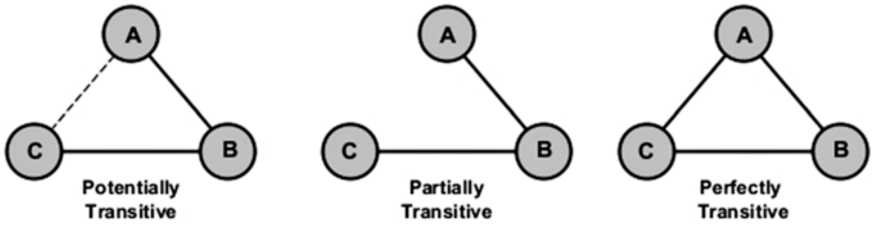


Fig.3.1 Transitivity between nodes

Transitivity

Transitivity is a network property that refers to the extent to which a relation between two nodes is transitive. It is a very important measure in social networks but less important in other types of networks. In social networks, the term transitivity reflects the friend-of-a-friend concept. It is sometimes used as a synonym of whole-network clustering coefficient.

Homophily

Homophily is the tendency of individuals to connect with others who share the same attitudes and beliefs. The tendency of individuals to associate with similar others based on gender, education, race, or other socioeconomic characteristics is very common in social communities. Coordination and cooperation are typically more successful between people who show some similarity to each other such that individuals in homophilic relationships are likely to hear about new ideas or ask for help from each other.Homophily in the context of online social networking can be understood from the similarity of users who are using the network in terms of age, educational back- ground, region, or profession. In the sense of corporate networks, homophily is translated as the similarity of professional or academic qualifications.

Balance and Triads

A triad is a network structure consisting of three actors and three dyads. Given a complete graph of three actors (triad), we can identify four different types of rela- tionships, depending on the number of negative relationships between nodes: (a) a friend of my friend is my friend, (b) an enemy of my enemy is my friend, (c) a friend of my friend is my enemy, and (d) an enemy of my enemy is my enemy.Triads with an odd number of “+” edges are bal- anced, while triads with an even number of “-” edges are unbalanced. Imbalanced graph configurations usually create stress for individuals located on them.

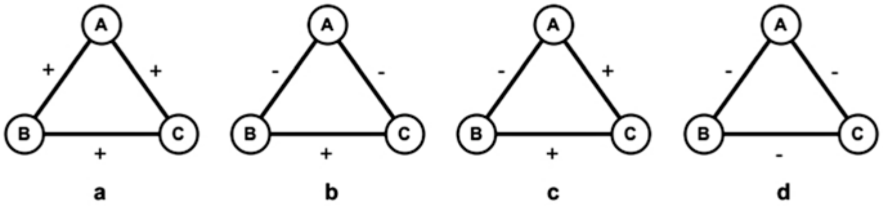


Fig. 3.2 Graph with three nodes in four states

The above graph is a type of signed graphs which have been studied since the 1950s. They are a special case of valued graphs in which ties are allowed to have one of two opposing values to convey the positive or negative sentiment. Examples of signed graphs include friend/foe, trust/distrust or like/dislike, esteem/disesteem, praise/blame, influence/negative influence, etc. They are very common in sociology and psychology but less common in fields such as physics and chemistry.

- In figure a, all the three actors have positive feelings, and there is no place for conflict among them. The configuration is coherent and lacks inner tensions between members.
- Figure b is also stable since two actors (B and C) share the same negative feeling towards actor A, but they like each other.
- Figure c is unstable because actors A and B have a negative feeling towards each other, while both have a positive feeling towards actor C which has to divide its loyalty between the other two actors.
- Figure d is also unstable and will eventually break down, as it has an odd number of negative signs.

In Fig. 3.2, b, types of balanced subgraphs are shown, whereas in Fig. 3.2 c, d, types of unbalanced graphs are presented. An obvious way to avoid unbalances in subgraphs is by sign shifting, which includes changing signs such that enmities (negative signs) become friendships (positive signs) or vice versa. Within real net- works, stable configurations appear far more often than unstable configurations.It should be noted here that a negative sign between two nodes does not mean the lack of tie between these two nodes. While a negative sign between two nodes is a clear mark of an inimical relationship, the absence of a tie between these nodes suggests the absence of interaction or communication between them.

Reciprocity

Reciprocity is a measure of the tendency towards building mutually directed connections between two actors. It refers to the number of reciprocated tie for a specific actor in a network. For example, if *u* connects to *v*, then *v* connects to *u* and vice versa. In real life scenarios, it is important to know whether received help is also given or whether given help is translated as help by the receiver.For a given node *v*, reciprocity is the ratio between the number of nodes which have both incoming and outgoing connections from/to *v*, to the number of nodes which only have incoming connections from *v*. For an entire network, reciprocity is calculated as the fraction of edges that are reciprocated. Average reciprocity is calculated by averaging reciprocity values of all nodes in the network.

Network Distribution

Measurements of network distribution are related to how nodes and edges are distributed in a network.

Distance Between Two Nodes

Distance is a network metric that allows the calculation of the number of edges between any pair of nodes in a network. Measuring distances between nodes in graphs is critical for many implementations like graph clustering and outlier detection. Sometimes, the distance measure is used to see if the two nodes are similar or not. Any commonly used shortest path calculation algorithm (e.g., Dijkstra) can be used to provide all shortest paths in a network with their lengths.We can use the distance measure to calculate node eccentricity, which is the maximum distances from a given node to all other nodes in a network. It is also possible to calculate network diameter, which is the highest eccentricity of its nodes and thus represents the maximum distance between nodes.In most social networks, the shortest path is computed based on the cost of tran- sition from one node to another such that the longer the path value, the greater the cost.Within a community, there might be many edges between nodes, but between communities, there are fewer edges.

Degree Centrality

In degree centrality metric, the importance of a node is determined by how many nodes it is connected to. It is a measurement of the number of direct links to other actors in the network. This means that the larger the number of adjacent nodes, the more important the node since it is independent of other actors that reach great parts of the network. It is a local measure since its value is computed based on the number of links an actor has to the other actors directly adjacent to it. Actors in social net- works with a high degree of centrality serve as hubs and as major channels of information.In social networks, for example, node degree distribution follows a power law distribution, which means that very few nodes have an extremely large number of connections. Naturally, those high-degree nodes have more impact in the network than other nodes and thus are considered more important. A node *i*'s degree central- ity $d(i)$ can be formulated as

$$d(i) = \sum_{j} m_{ij}$$

where $m_{ij} = 1$ if there is a link between nodes i and j and $m_{ij} = 0$ if there is no such link. For directed networks, it is important to differentiate between the in-degree centrality and the out-degree centrality.

Identifying individuals with the highest-degree centrality is essential in network analysis because having many ties means having multiple ways to fulfill the require- ments of satisfying needs, becoming less dependent on other individuals, and hav- ing better access to network resources. Persons with the highest-degree centrality are often third parties and deal makers and able to benefit from this brokerage. For directed networks, in-degree is often used as a proxy for popularity .The figure shows that node A and node B are at exceptional structural positions. All communications lines must go through them. This gives us a conclu- sion that both nodes, A and B, are powerful merely because of their excellent posi- tions. However, such a finding is largely based on the nature of links and the nature of embedded relationships.

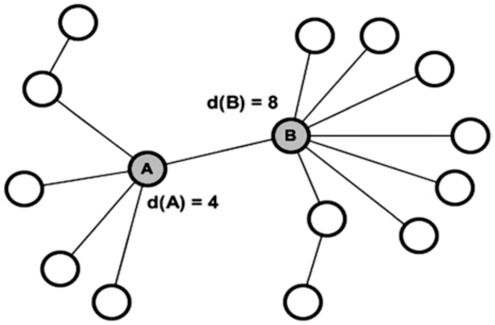


Fig. 3.3 Degree centrality of nodes

Closeness Centrality

Closeness centrality can be defined as how close, to a particular actor, other actors are. It is the sum of the geodesic distances of a node to all other nodes in the net- work. It computes the length of paths from one actor to other actors in the network.

That actor can be important if it is relatively close to the remaining set of actors in the network. The mathematical representation of closeness centrality, $C(i)$, is given as follows:

$$C(i) = \sum_{j} d_{ij}$$

where d_{ij} is the geodesic distance from node i to node j (number of links in the shortest path from node i to node j).

Closeness centrality is important to understand information dissemination in net- works in the way that the distance between one particular node and others has an effect on how this node can receive from or send information (e.g., gossip) to other nodes. In social networks, this ability is limited by what is called "horizon of observ- ability" which states that individuals have almost no sight into what is going on after two steps.Because closeness centrality is based on the distance between network nodes, it can be considered the inverse of centrality because large values refer to lower cen- trality, whereas small values refer to high centrality. Computationally, the value of *C(i)* is a number between 0 and 1, where higher numbers mean greater closeness (lower average distance) whereas lower numbers mean insignificant closeness (higher average distance) .In the figure, the nodes in gray are the most central regarding closeness because they can reach the rest of nodes in the network easily and equally.

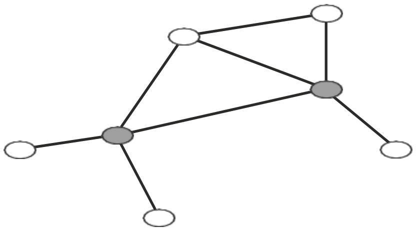


Fig3.4 Closeness centrality of nodes

They have the ability to reach all other nodes in the fastest amount of time. The other nodes lack these privileged positions.Because closeness centrality is based on shortest path calculations, its usefulness when applied to large networks can be brought into question in the way that closeness produces little variation in the results, which makes differentiating between nodes more difficult.In information networks, closeness reveals how long it takes for a bit of information to flow from one node to others in the network. High-scoring nodes usually have shorter paths to the rest of nodes in the network.

Betweenness Centrality

Betweenness centrality can be described as how important an actor is, as a link between different networks. It represents the number of times an actor needs to pass via a given actor to reach another actor. Nodes with high betweenness centrality control the flow of information because they form critical bridges between other actors or groups of actors. Betweenness centrality of node *i* is calculated as follows:

$$b(i) = \sum_{j,k} \frac{g_{jik}}{g_{jk}}$$

where g_{jk} is the number of shortest paths from node (*j*) to node *k* (*j* and $k \neq i$) and g_{jik} is the number of shortest paths from node (*j*) to node *k* passing through the node (*i*).

Eigenvector Centrality

Eigenvector centrality measurement describes the centrality of a person with regard to the global structure of the network. It assigns relative scores to all nodes in the network based on the concept that connections to nodes with high scoring contribute more to the score of the node in question than co*n*nections to nodes with low scoring.It measures the extent to which a node is connected to well-connected nodes. It is computed by taking the principal eigenvector of the adjacency matrix. Calculating centrality in the way that eigenvector measure proposes differs from the way that degree measure applies to calculate centrality which is based on simply adding up the number of links of each node.

Density

Density is defined as the degree to which network nodes are connected one to another. It can be used as a measure of how close a network is to complete. In the case of a complete graph (a graph in which all possible edges are present), density is equal to one. In real life, a dense group of objects has many connections among its entities (i.e., has a high density), while a sparse group has few of them (i.e., has a low density).

Formally, the density *D*(*G*) of graph *G* is defined as the fraction of edges in *G* to the number of all possible edges. Density values range between zero and one

[0, 1].

Proportion of ties between alters comparedto number possible ties

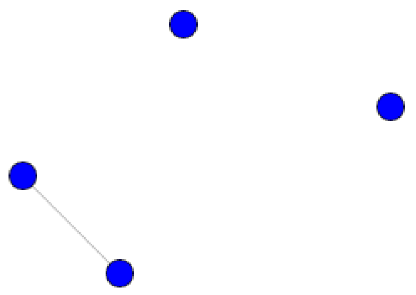


Fig.3.5 Ties

Total ties=1
No:of possible ties=6 (N*(N-1)/2)//N is the number of nodes
Density=1/6

Cohesive Subgroups

Cohesive groups are communities in which the nodes (members) are connected to others in the same group more frequent than they are to those who are outside of the group, allowing all of the members of the group to reach each other. Within such a highly cohesive group, members tend to have strong homogenous beliefs. Connections between community members can be formed either through personal contacts (i.e., direct) or joint group membership (i.e., indirect). As such, the more tightly the individuals are tied into a community, the more they are affected by group standards.

Cliques

A clique is a graph (or subgraph) in which every node is connected to every other node. Socially translated, a clique is a social grouping in which all individuals know each other (i.e., there is an edge between each pair of nodes). A triangle is an exam- ple of a clique of size three since it has three nodes and all the nodes are connected.A maximal clique is a clique that is not a subset of any other clique in the graph. A clique with size greater than or equal to that of every other clique in the graph is called a maximum clique.

Relaxation of Strict Cliques

- Distance (length of paths)

  - N-clique, n-clan, n-club
- Density (number of ties)
  - K-plex, ls-set, lambda set, k-core, component

N-cliques

The strict clique definition (maximal fully-connected sub-graph) may be too strong for many purposes. It insists that every member or a sub-group have a direct tie with each and every other member. You can probably think of cases of "cliques" where at least some members are not so tightly or closely connected. There are two major ways that the "clique" definition has been "relaxed" to try to make it more helpful and general.

One alternative is to define an actor as a member of a clique if they are connected to every other member of the group at a distance greater than one. Usually, the path distance two is used. This corresponds to being "a friend of a friend." This approach to defining sub-structures is called N-clique, where N stands for the length of the path allowed to make a connection to all other members.

N-Clans

The N-clique approach tends to find long and stringy groupings rather than the tight and discrete ones of the maximal approach. In some cases, N-cliques can be found that have a property that is probably undesirable for many purposes: it is possible for members of N-cliques to be connected by actors who are not, themselves, members of the clique. For most sociological applications, this is quite troublesome.

To overcome this problem, some analysts have suggested restricting N-cliques by insisting that the total span or path distance between any two members of an N-clique also satisfy a condition. The additional restriction has the effect of forcing all ties among members of an n-clique to occur by way of other members of the n-clique.

Ego Net

The ego-net approach to social network analysis, which takes discrete individual actors and their contacts as its starting point, is one of the most widely used approaches.

- Ego network (personal network)
- Ego: focal node/respondent
- Alter: actors ego has ties with
- Ties between alters

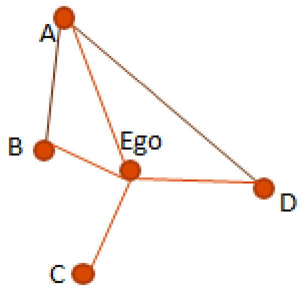


Fig 3.6 Ego network

Why use ego network data?

- From ego's perspective, personal network is important for:
  - Social support
  - Access to resources
  - Influence/normative pressure
- From a more global perspective, ego network data are useful for:
  - Studying mixing patterns between groups
  - Potential for diffusion
    - Disease propagation
    - Adoption of innovation: new product or health practice
- Lots can be had from ego network data!
  - Composition of individual's local social world
    - Demographic characteristics of alters
    - Shared health behaviors
  - Structural features
    - Size
    - Density
  - Nature of the ties
    - Frequency, duration, closeness
    - Specific exchanges

When to use Ego Network Analysis

- If your research question is about phenomena of or affecting individual entities across different settings (networks) use the ego-centric approach
  - Individual people, organizations, nations, etc.

- If your research question is about different patterns of interaction within defined groups (networks), use the socio-centric approach

  - E.g., who are the key players in a group? How do ideas diffuse through a group?

Which Theories are Ego-centric?

- **Most theories under the rubric of social capital are ego-centric**
- Topological
  - Structural holes / Brokerage
  - Embeddedness
- **Compositional**
  - Size
  - Alter attributes

Steps to a SNA study

1. **Identify the population**
   - Sampling, gaining access
2. **Determine the data sources**
   - Surveys, interviews, observations, archival
3. **Collect the data**
   - Instrument design

Step 1.Identify the Population

Sampling Criteria

- Determined by research question
  - High tech entrepreneurs
  - Alumni of defunct organizations
  - Basketball coaches
  - First time mothers returning to the workforce
  - Baseball Hall of Fame inductees
  - Contingent workers
  - People with invisible stigmatized identities

Step 1.Identify the Population

Gaining Access

- Same concerns as other research
  - It depends on the sensitivity of the questions that you are asking
  - Length of interview can be daunting
    - Depends on the number of alters

Step 2: Determine Data Sources

- Surveys
- Interviews
- Observations
- Archival data

Step 3: Collect the Data

- What data should you collect?
  - What questions need to be answered?
- How to format your data collection instrument (e.g., a survey, spreadsheet, database, etc.)?

Data Collection in an Ego-centric Study

1. Attributes about Ego
2. Name generator
   - Obtain a list of alters
3. Name interpreter
   - Assess ego's relationships with generated list of alters?
4. Alter Attributes
   - Collect data on the list of alters
5. Alter – Alter Relationships
   - Determine whether the listed alters are connected

Attributes about Ego

- Typical variables for case based analysis
  - Age
  - Gender
  - Education
  - Profession
  - SES
  - Etc.

Sample Name Generators

Questions that will elicit the names of alters

- From time to time, most people discuss important personal matters with other people. Looking back over the last six months who are the people with whom you discussed an important personal matter? Please just telI me their first namesor initials.

Consider the people with whom you like to spend your free time.

- Over the last six months, who are the one or two people you have been with the most often for informal social activities such as going out to lunch, dinner, drinks, films, visiting one another's homes, and so on?

Sample Name Interpreter

- Questions that deal with ego's relationshipwith [or perception of] each alter
  - How close are you with <alter>?
  - How frequently do you interact with <alter>?
  - How long have you known <alter>?
- All of these questions will be asked for eachalter named in the previous section

Sample Alter Attribute Questions As far as you know, what is <alter>' s highest

- As far as you know, what is <alter>' s highest level of education?
  - Age, occupation, race, gender, nationality, salary, drug use habits, etc
- Some approaches do not distinguish betweenname interpreters and alter attribute

Sample Alter-Alter Relationship Questions

- Think about the relationship between <alter1> and <alter2>. Would you say that they are strangers, just friends, or especially close?
- Note: this question is asked for each unique alter-alter pair. E.g., if there are 20 alters, there are 190 alter-alter relationship questions!
  - Typically, we only ask one alter-alter relationshipquestion

Why Ego-Centric Analysis

- **Asks different questions than whole networkanalysis.**
- In fact, many of the various approaches to "Social Capital" lend themselves particularly to the analysis of Ego-Centric or Personal networks

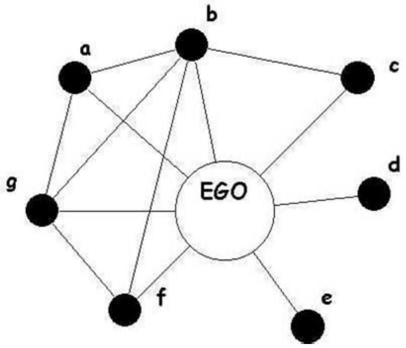
Kinds of Analyses

- In Ego-Centric Network analyses we are typically looking to use network-derived measures as variables in more traditionalcase-based analyses
  - E.g., instead of just age, education, and family SES to predict earning potential, we might also include heterogeneity of network or brokerage statistics

Many different kinds of network measures, thesimplest is degree (size
Data Analysis of Ego Networks

1. **Size**
   - How many contacts does Ego have?
2. **Composition**
   - What types of resources does ego have access to? (e.g., quality)
   - Does ego interact with others like him/herself? (e.g., homophily)
   - Are ego's alters all alike? (e.g., homogeneity?)
3. **Structure**
   - Does ego connect otherwise unconnected alters? (e.g., brokerage,density, etc)
   - Does ego have ties with non-redundant alters (e.g., effective size,efficiency, constraint)

Size



Degree = 7

Composition: Content

- The attributes (resources) of others to whom I amconnected affect my success or opportunities
  - Access to resources or information

– Probability of exposure to/experience with

Composition: Similarity Between Ego & Alter

- **Homophily**
  - We may posit that a relationship exists between some phenomenon and whether or not ego and alters in a network share an attribute
    - Selection
      - Teens who smoke tend to choose friends who also smoke
    - Influence
      - Overtime, having a network dominated by people with particular views may lead to one taking on those views

Composition: Homophily

- A CFO who surrounds herself with all finance people

- A Politician who surrounds himself with allmembers of the same political party

Composition: Dissimilarity Between Ego & Alter

Heterophily

– We may posit that a relationship exists between some phenomenon and a difference between ego and alters along some attribute

- Mentoring tends to be heterophilous with age

Composition: Homophily/Heterophily

Krackhardt and Stern's E-I index

$$\frac{E - I}{E + I}$$

• E is number of ties to members in different groups (external), I is number of ties to members of same group (internal).Varies between -1 (homophily) and +1 (heterophily)

Composition: Heterogeneity

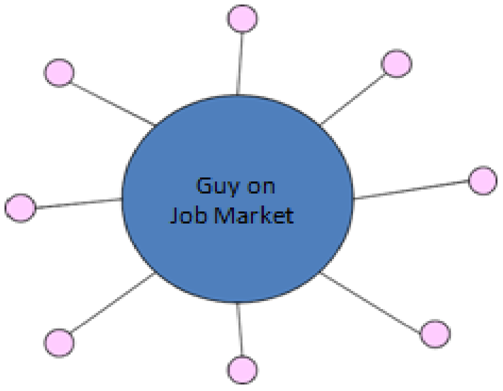
- Similar to homophily, but distinct in that it looks not atsimilarity to ego, but just among the alters

- Diversity on some attribute may be provide access todifferent information, opinions, opportunities, etc.
  - My views about social welfare may be affected by the diversity in SES present in my personal network (irrespective of or in addition to my own SES)

Structural Analyses

- Burt's work is particularly and explicitly ego-network based in calculation
  - My opportunities are affected by the connections that exist (or are absent) between those to whomI am connected

Structural Holes



- Basic idea: Lack of ties among alters may benefit ego
- Benefits
  - Autonomy
  - Control
  - Information

Fig 3.7 Structural holes

Burt's Measures of Structural Holes

- Effective size
- Efficiency
- Constraint

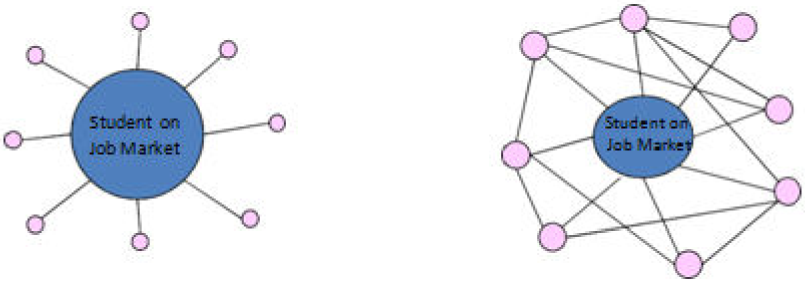
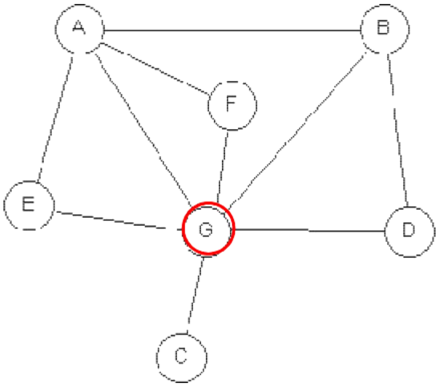
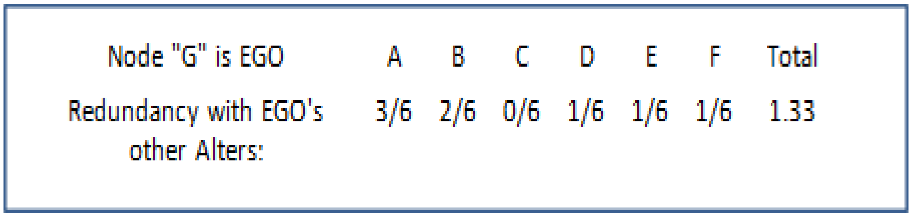


Fig 3.8 Structural holes

Effective Size





Node "G" is EGO	A	B	C	D	E	F	Total
Redundancy with EGO's other Alters:	3/6	2/6	0/6	1/6	1/6	1/6	1.33

Fig 3.9 Effective Size

Effective Size of G = Number of G's Alters – Sum of Redundancy of G's alters

= 6 – 1.33 = 4.67

Efficiency

Efficiency = (Effective Size) / (Actual Size)

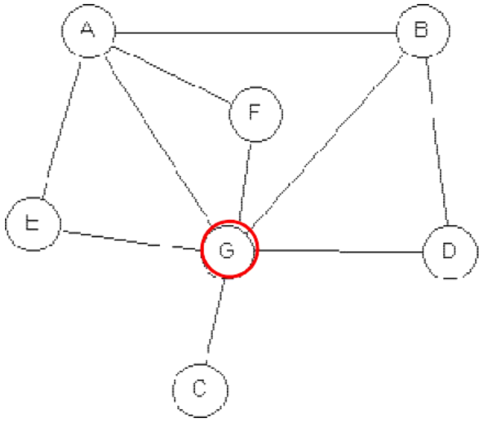


Fig 3.10 Efficiency

Actual Size = 6

Effective Size of G = 4.67

Efficiency = 4.67/6 = ~0.78

Constraint: The Basic Idea

- Constraint is a summary measure that taps the extent to which ego's connections are to others whoare connected to one another.
- If ego's boyfriend bowls with her brother and father every Wednesday night, she may be constrained in terms of distancing herself from him, even if they break up.
- There's a normative bias in much of the literature that less constraint is good

Ego-Centric Network Analysis

- When conducted across many, independent egos, presents different problems
- Many Social Network Analysis tools ill suited to the nature of such analyses
  - Really designed for “whole network” analysis
- Ego Network analyses require either:
  - joining into one large, sparse, blocked network, or
  - repetition of analysis of individual networks

- Can be tedious if there’s no facility for batching them

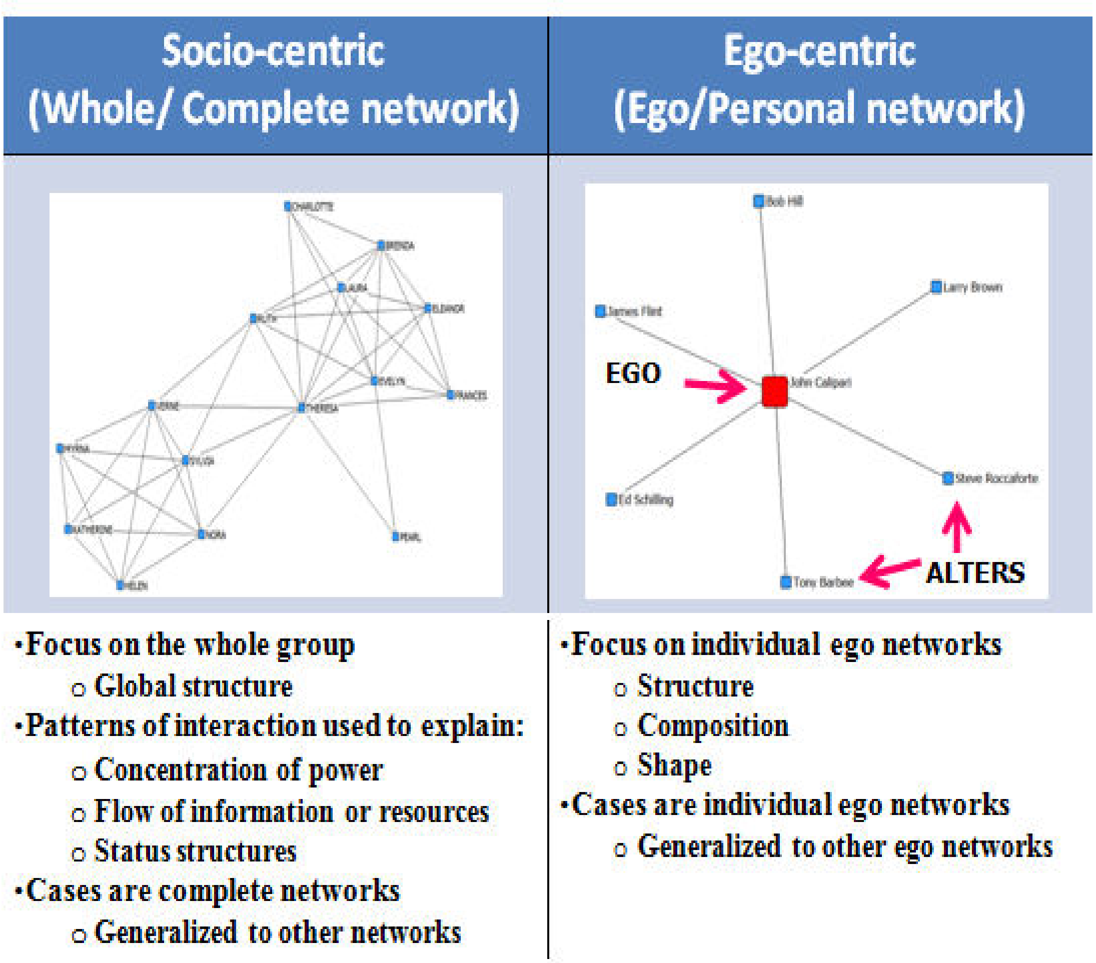
Local and Global Measures

Local: personal network size

- Number of alters (social support) predicting health outcomes
- Number of drug partners predicting future risky behavior

Global: degree distribution by aggregating over all cases

- Distribution of ties per person



Socio-centric (Whole/ Complete network)	Ego-centric (Ego/Personal network)
•Focus on the whole group ○ Global structure •Patterns of interaction used to explain: ○ Concentration of power ○ Flow of information or resources ○ Status structures •Cases are complete networks ○ Generalized to other networks	•Focus on individual ego networks ○ Structure ○ Composition ○ Shape •Cases are individual ego networks ○ Generalized to other ego networks