Analysis of Egocentric Network Data

Tso-Jung Yen

Institute of Statistical Science
Academia Sinica

tjyen@stat.sinica.edu.tw

Summer School 2014

August 1, 2014
Outline

- Introduction
- Data collection
- Basic concepts
  - Centrality measures
  - Local bridge and embeddedness
  - Principle of triadic closure
  - Principle of homophily
Introduction

• What is an egocentric network?
  • From a graph viewpoint:
    • One node called \textit{ego} is at center, surrounded by other nodes called \textit{alter}.
    • By "surrounded" we mean all alters surrounding the ego have links to the ego.
    • \textit{Remark}: We use "node" and "link" instead of "vertex" and "edge" when talking about graph.
Introduction

• What is an egocentric network?

• From a graph viewpoint (contd):

  • Mathematically, an egocentric network is $G = (V, E)$ with

    $$ V = \{\{\text{ego}\}, \{\text{alter 1, alter 2, \cdots, alter n}\}\}, $$

    being the node set and $E = \{E_{\text{ego,alter}}, E_{\text{alter,alter}}\}$

    being the link set, where

    $$ E_{\text{ego,alter}} = \text{all links connecting alters to ego}, $$

    $$ E_{\text{alter,alter}} = \text{all links connecting alters to alters}. $$
Introduction

• Examples of egocentric networks:

• Computer networks: One master and many slaves. Possibly slaves can communicate with each other.

• Friend networks: You are at center, surrounded by your friends.

• Commonly-seen real world examples: Facebook or LINE friend list, Twitter following list and so on.

• In principle, an egocentric network can always be constructed from a complete network, e.g. choosing a focal node and keeping nodes that are connected to the focal node.
Introduction

Figure: Graphical representation of an egocentric network.
Introduction

Figure: Graphical representation of an egocentric network.
Introduction

$|E_{ego}| = 161, \ |E_{alters}| = 1587$

Figure: Graphical representation of an egocentric network.
Introduction

Figure: Graphical representation of an egocentric network.
Data Collection

• Collecting egocentric network data:
  • Name generators (Laumann, 1973; Wellman, 1979):
    • Ask respondents (ego) to provide a list of contacts.
    • Frequently asked questions: Who are your best friends? With whom you exchange specific resources and information most often?
    • Criteria for filtering these contacts: friends, relatives, acquaintances, strangers and so on.
Data Collection

• Collecting egocentric network data:

  • Position generators (Granovetter, 1973; Lin and Dumin, 1986):
    
    • Ask respondents whether they have connections to someone who holds a specified occupation such as the doctor, lawyer and so on.

    • Better than name generators in measuring bridges or weak ties, i.e. the links by which people get access to important resources.
Data Collection

• Collecting egocentric network data (contd):
  • Contact diary (de Sola Pool and Kochen, 1978; Freeman and Thompson, 1989; Fu, 2007):
    • Collects egocentric network data via self-reporting.
    • Collects egocentric network data on a daily basis. Data are in longitudinal format.
Data Collection

- Collecting egocentric network data (contd):
  - Contact diary (contd):
  - Advantages:
    - Highly structured.
    - Cost-effective.
    - Comprehensive and complete.
    - More likely to avoid recalling biases.
  - Disadvantages:
    - Labor-intensive and demanding.
    - Subject to self-selection, social desirability and manipulation.
    - In the long run, no incentive, no participation.
    - May only be suitable for small-scale research.
Data Collection

• ClickDiary:
  • Collects egocentric network data using the contact diary method:
    • Online platform.
    • Asks respondents whom they contact with on a daily basis.
    • Contact type, time, duration, location, emotion change, health information.... and so on.
  • 22 egocentric networks:
    • Collected between May 1, 2014 and June 30, 2014.
    • The largest one contains 358 alters. The smallest one contains 43 alters.
    • In total, the 22 egocentric networks contain 2,634 alters and 34,483 links.
Basic Concepts

• How can we statistically summarize an egocentric network?
  • Properties at the node level:
    • Centrality measures such as degree, betweenness, closeness and so on.
    • Transitivity measures such as the clustering coefficient.
  • Properties at the dyad level:
    • local bridge, embeddedness, and neighborhood overlap.
    • A dyad means a pair of nodes.
  • Properties at the global level or due to context:
    • Transitivity measures such as the clustering coefficient.
    • Homophily measures such as the assortative mixing coefficient.
Basic Concepts

• Centrality measures:
  
  • Let $A_{ij}$ be the link connecting node $i$ to node $j$. Assume the network is symmetric and $A_{ij}$ is binary-valued.

  • **Degree centrality of node $i$:**

    $$D_i = \sum_{j=1}^{n} A_{ij}.$$ 

  • A measure on importance of a node in a network.

  • Such importance includes ability of accessing to information, prestige, popularity and so on.

  • In an egocentric network, ego should have the maximum degree.
Basic Concepts

Figure: Box plot of degree values for the 22 egocentric networks.
Basic Concepts

Figure: Degree distributions for the 22 egocentric networks.
Basic Concepts

• Centrality measures (contd):

  • Let $b_{jk}$ be the number of shortest paths between nodes $j$ and $k$, and $b_{jk,i}$ be the number of those shortest paths that contain node $i$.

  • **Betweenness centrality of node $i$ (Freeman, 1977):**

    $$B_i = \sum_{j=1}^{n} \sum_{k=j+1}^{n} \frac{b_{jk,i}}{b_{jk}}.$$

  • A measure on a node’s ability to control over information flow in a network.

  • Not about ”how well-connected a node is”, but about ”how much a node falls between other nodes”.

  • A node may have low degree but high betweenness.
Figure: Box plot of betweenness values for the 22 egocentric networks (zero valued data excluded).
Basic Concepts

Figure: Betweenness distributions for the 22 egocentric networks.
Basic Concepts

- Centrality measures (contd):
  - Let $h_{ij}$ be the shortest path between nodes $i$ and $j$.
  - **Closeness centrality of node $i$:**
    \[
    H_i = \frac{n - 1}{\sum_{j=1}^{n} h_{ij}}.
    \]
  - More meaningful definition (Eq. 7.30, Newman, 2010):
    \[
    H_i^* = \frac{1}{n - 1} \sum_{j \neq i} \frac{1}{h_{ij}},
    \]
    $H_i^*$ avoids the situation when $h_{ij} = \infty$, and gives more weight to nodes close to $i$. 

Basic Concepts

• Comparisons between sociocentric and egocentric networks:
  • Qualitative comparisons:
    • Well-defined ego. One ego one egocentric network.
    • Ego is connected to all other nodes in the egocentric network.
    • Egocentric networks can be constructed from the sociocentric network by: (1) selecting focal nodes; (2) keeping links connected to the focal nodes; (3) deleting links that are not connected to the focal nodes.
Basic Concepts

• Comparisons between sociocentric and egocentric networks (contd):

  • Quantitative comparisons (Marsden, *Social Networks*, 2002):
    
    • Closeness centrality is not suitable for describing egocentric networks.
    
    • In an egocentric network, closeness centrality of ego is equal to 1.
    
    • The maximum shortest path between any nodes in an egocentric network is equal to 2.
Basic Concepts

• Comparisons between sociocentric and egocentric networks:
  • Quantitative comparisons (contd):
    • In an egocentric network, betweenness centrality of ego may exaggerate the extent to which the ego lies on the shortest path between two arbitrary nodes in the ego centric network.
    • By adopting Mizruchi et al.’s criteria (Mizruchi et al., 1986), the ego with high hub centrality and low bridge centrality will have a higher value in betweenness centrality in an egocentric network than in a sociocentric network.
    • In general, findings suggest there is a high correlation between betweenness of an egocentric network and a sociocentric network.
Basic Concepts

• Local bridge and embeddedness:

  • **Local bridge**: A link between nodes \( i \) and \( j \) is called a local bridge if \( i \) and \( j \) are not connected to any common node.

  • **Span of a local bridge** \((i, j)\):
    • The length of the shortest path between \( i \) and \( j \) if \((i, j)\) is deleted.

  • If span of \((i, j)\) is large:
    • Imply that the local bridge may span into many different groups, and therefore bringing different information and opportunities.
Basic Concepts

- Local bridge and embeddedness (contd):
  - **Structure holes:**
    - A structural hole is the empty space between two unconnected networks that share non-redundant information.
  - Advantages of nodes that bridge structure holes:
    - Can access to information that other nodes may not be able to access.
    - May also be more creative as it can access to multiple ideas from different groups.
    - Can serve as a gate-keeping, allowing or blocking other nodes’ access to the group it belongs to.
  - Disadvantages of nodes that bridge structure holes:
    - Less embedded within a single group, and therefore is less protected by the presence of mutual network neighbors.
Basic Concepts

• Local bridge and embeddedness (contd):
  
  • Let \( \text{Ne}(i) \) denote the neighborhood of \( i \).
  
  • **Embeddedness of dyad \((i, j)\):**

    \[
    O_{ij} = |\text{Ne}(i) \cap \text{Ne}(j)|.
    \]

  • **Neighborhood overlap between nodes \( i \) and \( j \):**

    \[
    O^*_{ij} = \frac{O_{ij}}{|\text{Ne}(i) \cup \text{Ne}(j)|}.
    \]

  • Neighborhood overlap can be seen as a normalized version of embeddedness.

  • A local bridge always has the embeddedness equal to zero.
Basic Concepts

• Local bridge and embeddedness (contd):
  
  • For an egocentric network with size $n$:
    
    • **Ego neighborhood overlap score**: 
      
      $$Z_{ego} = \frac{1}{n} \sum_j O_{ego,j}^*.$$ 
    
    • **Alter neighborhood overlap score**: 
      
      $$Z_{alters} = \frac{2}{n(n-1)} \sum_{i,j:i,j\text{ are alters}} \frac{|Ne(i) \cap Ne(j)| - 1}{|Ne(i) \cup Ne(j)| - 1}.$$ 
    
    • Ego has been ignored when calculating $Z_{alters}$. 

Basic Concepts

Figure: Box plot of neighborhood overlap for linked and non-linked dyads.
Basic Concepts

Figure: Neighborhood overlap densities for linked and non-linked dyads.
Figure: Scatter plot of alter neighborhood overlap score versus ego neighborhood overlap score for the 22 egocentric networks.
Basic Concepts

Figure: Scatter plot of alter neighborhood overlap score versus ego neighborhood overlap score for the 22 egocentric networks.
Figure: Scatter plot of alter neighborhood overlap score versus ego neighborhood overlap score for the 22 egocentric networks.
Basic Concepts

- Local bridge and embeddedness (contd):
  - **Weak ties**: Links corresponding to acquaintances (while strong ties are the links corresponding to friends).
  - In an egocentric network, whether a link is a strong tie or a weak tie is dependent on ego’s perception. It is defined by ego itself.
  - The span of a local bridge measures how far one can reach with the local bridge. The span of a local bridge therefore may be used to measure the strength of weak ties.
Basic Concepts

• Principle of triadic closure (Simmel, 1903; 1908):
  • Transitive relation: If $i$ has a link to $j$ and $j$ has a link to $k$, then $i$ also has a link to $k$.
  • Transitivity refers to the probability that a transitive relation holds for any pairs of nodes in a network.
  • Perfect transitivity means that all nodes know each other, e.g. a clique (complete subgraph).

• **Principle of triadic closure:** Two have friends in common will become friends some time later.
Basic Concepts

- Principle of triadic closure (contd):

  - **Local clustering coefficient of node** $i$:

    $$ C_i = \frac{\text{number of pairs of neighbors of } i \text{ are connected}}{\text{number of pairs of neighbors of } i}. $$

  - It is the probability that a pairs of $i$’s friends are friends themselves.

  - It can be used to measure the structure holes surrounding $i$, e.g. $1 - C_i$ represents $i$’s control over information flow between all pairs of nodes in $i$’s neighborhood.
Figure: Box plot of clustering coefficient values for the 22 egocentric networks.
Basic Concepts

Figure: Clustering coefficient densities for the 22 egocentric networks.
Basic Concepts

• Principle of triadic closure (contd):

  • (global) clustering coefficient:

$$ C = \frac{\text{number of closed paths of length two}}{\text{number of paths of length two}} $$

$$ = \frac{\text{number of triangles} \times 6}{\text{number of paths of length two}} $$

$$ = \frac{\text{number of triangles} \times 3}{\text{number of connected triples}}. $$

• It is the probability that two nodes connected to a common node are themselves connected.

• Social networks tend to have high values in the clustering coefficient, as compared with technological and biological networks (Section 7.9 of Newman, 2010).

• In social networks, two people will have a higher chance of being friends to each other if they have a common friend.
Basic Concepts

Figure: Scatter plot of the clustering coefficient versus edge density for the 22 egocentric networks.
Basic Concepts

Figure: Scatter plot of the clustering coefficient versus edge density for the 22 egocentric networks.
Basic Concepts

Figure: Scatter plot of the clustering coefficient versus edge density for the 22 egocentric networks.
Basic Concepts

- Principle of triadic closure (contd):
  - Three mechanisms behind triadic closure:
    - Higher opportunities of meeting each other.
    - Higher trust to each other.
    - Incentive to meet each other.
Basic Concepts

• Principle of triadic closure (contd):
  
  • **Social capital:**

    • According to Portes (1998), social capital is one’s ability to secure benefits via its position in social networks or other social structure.

    • Coleman (1988) argued that triadic closure and embedded links may enforce norms and reputation, protecting the integrity of social and economic transactions, and therefore are a form of social capital.

    • Burt (2000) argued that a tension between closure and brokerage is a form of social capital.
Basic Concepts

- Principle of homophily:
  
  **Principle of Homophily (Lazarsfeld and Merton, 1954; McPherson et al., 2001):** Phenomenon that in a network links tend to appear between nodes of the same type.
  
  Homophily in social networks: e.g. age, race, gender, class, education, belief, and so on.
  
  
  Homophily is "extrinsic" and "contextual" to network itself, by "extrinsic" we mean to measure homophily we need extra information, e.g. attribute of nodes.
Basic Concepts

- Principle of homophily (contd):
  - Measuring homophily:
    - Let \( m = 2^{-1} \sum_{i,j} A_{ij} \), \( k_j = \sum_i A_{ij} \), and
      \( \delta(c_i, c_j) = 1 \) if \( c_i = c_j \), and \( \delta(c_i, c_j) = 0 \) otherwise.
  
  - Modularity (Newman, 2002; 2003):
    \[
    Q = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j).
    \]

- Assortativity coefficient:
  \[
  Q^* = \frac{Q}{Q_{\text{max}}},
  \]
  where
  \[
  Q_{\text{max}} = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j).
  \]
Basic Concepts

- Principle of homophily (contd):
  - Measuring homophily (contd):
    - **Assortative mixing by scalar characteristics**
      (Section 7.13.2 of Newman, 2010):

\[
Q^* = \frac{Q}{Q_{\text{max}}},
\]

where

\[
Q = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j,
\]

\[
Q_{\text{max}} = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) x_i^2.
\]
Basic Concepts

- Principle of homophily (contd):
  - Interpretation of the assortativity coefficient:
    - Weighted Pearson correlation coefficient, having a value between 1 and −1.
    - When equal to 1: perfect assortative mixing. Links only appear between nodes of the same type.
    - When equal to −1, perfect disassortative mixing. Links only appear between nodes of different types.
    - When equal to 0, link appearances have nothing to do with node type.
Figure: Scatter plot of sex assortativity versus degree assortativity for the 22 egocentric networks.
Basic Concepts

Figure: Scatter plot of sex assortativity versus degree assortativity for the 22 egocentric networks.
Figure: Scatter plot of sex assortativity versus degree assortativity for the 22 egocentric networks.