Characterizing Network Cohesion

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

February 27, 2018
Local density

Local density, clustering coefficient and group centrality

Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure
Many network analytic questions pertain to network cohesion

Example

- **Q1:** Do common friends of an actor end up being friends?
- **Q2:** What collections of proteins in a cell work closely together?
- **Q3:** Does Web page structure separate relative to content?
- **Q4:** What portion of the Internet topology constitutes a ‘backbone’?

Definitions of network cohesion depend on the context

⇒ Scale from local (e.g., triads) to global (e.g., giant components)
⇒ Specified explicitly (e.g., cliques) or implicitly (e.g., clusters)
Cohesive subgroups defined by social network analysts as:

‘Actors connected via dense, directed, reciprocated relations’

Allow sharing information, creating solidarity, collective actions

Ex: religious cults, terrorist cells, sport clubs, military platoons, …

Desirable properties of a cohesive subgroup

⇒ Familiarity (degree);
⇒ Reachability (distance);
⇒ Robustness (connectivity); and
⇒ Density (edge density)

Natural to think of cliques, i.e., complete subgraphs of $G$
Local density and cliques

- Large cliques are rare; single missing edge destroys property

- Sufficient condition for the existence of a size-$n$ clique
  \[ N_e > \frac{N_v^2 (n - 2)}{2 (n - 1)}, \]  
  while sparse graphs have $N_e = O(N_v)$

- Complexity of clique-related algorithms varies widely
  - Is $U \subseteq V$ a clique? Is it maximal? $O(N_v + N_e)$ complexity
  - Identifying all triangles in $G$? $O(N_v^3)$ ($O(N_v^{\sqrt{2}})$ for sparse graphs)
  - Does $G$ have a maximal clique of size $\geq n$? NP-complete
Relaxing cliques by familiarity

- Cliques tend to be an overly restrictive notion of cohesiveness. Relax!

- **Def:** An induced subgraph $G'(V', E')$ is a $k$-plex if $d_v(G') \geq |V'| - k$ for all $v \in V'$, and $G'$ is maximal.

  - Degrees are in the induced subgraph $G'$, not in $G$
  - No vertex is missing more than $k - 1$ of its possible $|V'| - 1$ edges
  - A clique is a 1-plex

- **Complex:** problems involving $k$-plexes scale like clique counterparts
The $k$-core decomposition

- Recall the $k$-core decomposition. A dual notion of cohesiveness

  ![Diagram of k-core decomposition]

- **Def:** An induced subgraph $G'(V', E')$ is a $k$-core if $d_v(G') \geq k$ for all $v \in V'$, and $G'$ is maximal

- **Hierarchy:** larger “coreness” $\Rightarrow$ larger degrees and centrality

- **Algorithm:** recursively prune all vertices of degree less than $k$
  $\Rightarrow$ Complexity $O(N_v + N_e)$, very efficient for sparse graphs
Relaxing cliques by reachability

- **Idea:** specify that any two actors are no more than $k$ hops away

- **Def:** An induced subgraph $G'(V', E')$ is a $k$-clique if $d(u, v) \leq k$ for all $u, v \in V'$

$\Rightarrow$ Useful if important social processes occur via intermediaries

$\Rightarrow$ diam($G'$) may exceed $k$, if distances used are in $G$

- Likewise, a $k$-club is a subgraph $G'$ with diam($G'$) $\leq k$

$\Rightarrow$ $k$-clubs are $k$-cliques but the converse is not true, in general
A natural measure of density of a subgraph \( G'(V', E') \) is

\[
\text{den}(G') = \frac{|E'|}{|V'|(|V'| - 1)/2} \in [0, 1]
\]

⇒ Quantifies how close is \( G' \) to being a clique

\[
\bar{d}(G') = \frac{1}{|V'|} \sum_{v \in V'} d_v = \frac{2|E'|}{|V'|} \Rightarrow \text{den}(G') = \frac{\bar{d}(G')}{|V'| - 1}
\]

⇒ Flexibility in choosing \( G' \) to measure local density via \( \text{den}(G') \)

⇒ Use \( v \)'s egonet \( G'_v \), subgraph induced by \( v \) and its neighbors

⇒ Density of the overall graph \( G \) is

\[
\text{den}(G) = \frac{2N_e}{N_v(N_v - 1)}
\]
Q: What fraction of \( v \)'s neighbors are themselves connected?

**Def:** The clustering coefficient \( \text{cl}(v) \) of \( v \in V \) is

\[
\text{cl}(v) = \frac{2|E_v|}{d_v(d_v - 1)} \in [0, 1]
\]

\( |E_v| \) is the number of edges among \( v \)'s neighbors

An indication of the extent to which edges ‘cluster’

The global (average) clustering coefficient is

\[
\text{cl}(G) = \frac{1}{N_v} \sum_{v \in V} \text{cl}(v)
\]
Example: MSN social network

- MSN social network: $N_v \approx 180M$, $N_e \approx 1.3B$ [Leskovec et al’06]

Average clustering coefficient $cl(G) = 0.1140$ is large

Compare with the Erdös-Renyi random graph model

$$cl(G_{n,p}) = \Pr \text{[Edge closes triangle]} = p = \frac{\bar{d}}{n - 1} \to 0$$
Extending centrality to vertex groups

- Capture the importance of node subgroups [Everett et al’99]
- Q1: Are engineers more popular than accountants in an organization?
- Q2: How do we select board members with most business influence?
- Group centrality measures to generalize vertex centrality
- Ex: Consider subgraph $G'(V', E')$ induced by node subset $V'$
  - Let $U_{V'} \subset V \setminus V'$ with edges to members of $V'$
- Group degree centrality of node subset $V'$
  \[ d_{V'} = |U_{V'}| \]
  \[ \Rightarrow \text{Number of non-group nodes connected to } G' \]
Group centrality measures

- **Def:** Distance from \( v \in V \) to a group of nodes \( V' \subset V \) is
  \[
d_*(v, V') = \min_{u \in V'} d(u, v)
  \]

- **Group closeness centrality** of node subset \( V' \)
  \[
c_{Cl}(V') = \frac{1}{\sum_{u \in V \setminus V'} d_*(u, V')}
  \]

- **Group betweenness centrality** of node subset \( V' \)
  \[
c_{Be}(V') = \sum_{s \neq t \in V \setminus V'} \frac{\sigma(s, t|V')}{\sigma(s, t)}
  \]

- \( \sigma(s, t) \) is the total number of \( s - t \) shortest paths \((s, t \in V \setminus V')\)
- \( \sigma(s, t|V') \) is the number of \( s - t \) shortest paths through \( v \in V' \)
Connectivity

Local density, clustering coefficient and group centrality

Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure
Network connectivity and robustness

- Connectivity relevant when taking a larger, global perspective
  - Q: Does a given graph $G$ separate into different subgraphs?
  - If it does not, a ‘less robust’ network is closer to splitting

- **Def:** Graph is connected if $\exists$ walks joining each vertex pair

$\Rightarrow$ If bridge edges are removed, the graph becomes disconnected
Connected components

- **A component** is a maximally-connected subgraph

![Graph with nodes 1, 2, 3, 4, 5, 6, 7]

- In figure ⇒ Components are \{1, 2, 5, 7\}, \{3, 6\} and \{4\}
  ⇒ Subgraph \{3, 4, 6\} not connected, \{1, 2, 5\} not maximal

- Disconnected graphs have 2 or more components
  ⇒ Number of components = Multiplicity of eigenvalue 0 for \(L\)
  ⇒ Largest component often called **giant component**

- Check for connectivity, identify components with DFS, BFS: \(O(N_v)\)
Giant connected components

- Large real-world networks typically exhibit one giant component
- **Ex:** romantic relationships in a US high school [Bearman et al’04]

Q: Why do we expect to find a single giant component?
- A: Well, it only takes one edge to merge two giant components
Giant components tend to exhibit the \textit{small world} property

Small refers to the \textit{average path length}

\[
\bar{\ell} = \left( \frac{N_v}{2} \right)^{-1} \sum_{u \neq v \in V} d(u, v) = O(\log N_v)
\]

\textbf{Ex:} facilitates spread of gossip, diseases, search for WWW content

Not too surprising that the property holds. Informal argument:

If \( d_v = d \), after \( h_\ast \) hops have \( d_{h_\ast} \approx N_v \) \( \Rightarrow \bar{\ell} \approx h_\ast = O(\log N_v) \)
Connectivity of directed graphs

- Connectivity is more subtle with directed graphs. Two notions

- **Def:** Digraph is **strongly connected** if for every pair \( u, v \in V \), \( u \) is reachable from \( v \) (via a directed walk) and vice versa

- **Def:** Digraph is **weakly connected** if connected after disregarding arc directions, i.e., the underlying undirected graph is connected

- Above graph is weakly connected but not strongly connected
  \[ \Rightarrow \text{Strong connectivity obviously implies weak connectivity} \]
First described for the Web graph in [Broder et al’00]

Core element is the strongly-connected component (SCC)

- In-component (IC): vertices reaching SCC, but not vice-versa
- Out-component (OC): vertices reached by SCC, but not vice-versa
- Tubes: vertices in between the IC and OC, not in SCC
- Tendrils: vertices that cannot be reached by, or reach the SCC

In general, the digraph may be disconnected with a giant SCC
Example: AIDS blog network

- Network of citations among 146 blogs related to AIDS
  - Small SCC with 4 vertices and IC with 2 vertices
  - OC dominates with 112 vertices, and few tendrils (28 vertices)

- For the WWW, Broder et al. found $|\text{SCC}| \approx |\text{IC}| \approx |\text{OC}| \approx 56M$
Local density, clustering coefficient and group centrality

Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure
Assortative mixing

- People have a stronger tendency to associate with equals
  \[\Rightarrow\] Tendency is called **homophily** or **assortative mixing**

- **Ex:** high-school students by race, bloggers by political party, . . .
  \[\Rightarrow\] Can have **disassortative mixing** e.g., romantic relationships
Quantifying assortative mixing

- Suppose that vertex characteristics are categorical, e.g., male/female

- Let $f_{ij}$ be the fraction of edges joining vertices of categories $C_i$, $C_j$
  \[ f_{i+} = \sum_j f_{ij} \text{ is the } i\text{-th marginal row (column) sum} \]

- Define the assortativity coefficient [Newman’03]
  \[ r_a = \frac{\sum_i f_{ii} - \sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}} \]
  \[ f_{i+} f_{+i} \text{ is the expected fraction of edges joining nodes in } C_i \]
  \[ \Rightarrow \text{ Random edges preserving degree distribution yields } r_a = 0 \]

- Perfectly assortative mixing yields $r_a^{\text{max}} = 1$, while the minimum is
  \[ r_a^{\text{min}} = -\frac{\sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}} < -1 \]
Example: Abilene network

- Abilene network for US universities and research labs
  - ‘Core’ nodes, as well as e.g., ‘Connector’ nodes and ‘Exchange points’

- Hierarchical structure, suggestive of disassortative mixing
Disassortative mixing in Abilene

- Tabulated counts of inter-category edges in Abilene

<table>
<thead>
<tr>
<th></th>
<th>Core</th>
<th>Exchange</th>
<th>Peer</th>
<th>Conn.</th>
<th>Part.</th>
<th>Conn./Part.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>14</td>
<td>6</td>
<td>5</td>
<td>17</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Exchange</td>
<td>6</td>
<td>1</td>
<td>46</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Peer</td>
<td>5</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Conn.</td>
<td>17</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>203</td>
<td>0</td>
</tr>
<tr>
<td>Part.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>203</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>Conn./Part.</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>34</td>
<td>34</td>
<td>0</td>
</tr>
</tbody>
</table>

- Fractions $f_{ij}$ obtained by scaling table entries by the total of 675

- Assortativity coefficient $r_a = -0.3162$, close to $r_{a\text{min}} = -0.3461$

  ⇒ Strongly supports our suspicion of disassortative mixing
Case study

Local density, clustering coefficient and group centrality

Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure
Network analysis and epilepsy

- **Epilepsy** is the world’s most common serious brain disorder
  - ⇒ **Seizures**, i.e., recurrent abnormal neuronal activity

- **Ex:** Network-oriented analysis of epileptic seizure data in humans


- Leverage few summaries of network characteristics we learnt so far
Measurement

- Electrode grid (8x8) implanted in the cortical surface of the brain
  - Also implanted two strips of 6 electrodes (deeper, not shown)

- Electrocorticogram (ECoG) data; voltages indicative of brain activity

- Two 10 sec. intervals of interest for comparison:
  - Preictal period: prior to the epileptic seizure
  - Ictal period: immediately after start of seizure

- Top time-series is smoothed, averaged over 8 seizure signals
Network graph construction

- Network $\rightarrow$ represent couplings among brain regions
  - Graphs for the preictal and ictal periods, for 8 seizures
- Vertices: correspond to the 76 electrodes (cortical brain regions)
- Edges: threshold correlations between pairwise 10 sec. time series

- Brain is in two very different states before and during seizure
  - Thinning of edges, coupling reduction at seizure onset
  - Closest to the strips, where seizure was suspected to emanate

![Network representations of cortical-level coupling between brain regions about each electrode, during preictal (left) and ictal (right) periods.](image)
Summaries of network characteristics

- Evaluated degree, closeness, betweenness centrality; clustering coeff.
  ⇒ Show preictal and ictal periods, as well as their difference

- Identifies spatially localized brain regions that may facilitate seizures
  ⇒ May serve to more precisely guide surgical intervention
Glossary

- Network cohesion
- Cohesive subgroups
- Familiarity
- Reachability
- Robustness
- Local density
- Cliques
- $k$-plex and $k$-core
- $k$-clique and $k$-club
- Egonet

- Clustering coefficient
- Bridge edges
- Giant connected component
- Small world phenomenon
- Average path length
- Bowtie structure
- Strongly-connected component
- (Dis) assortative mixing
- Homophily
- Brain networks