Characterizing Network Cohesion

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

February 20, 2023

Local density

Local density, clustering coefficient and group centrality

Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure

Network cohesion

- 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

  $\Rightarrow$ Scale from local (e.g., triads) to global (e.g., giant components)

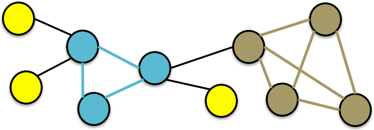
  $\Rightarrow$ Specified explicitly (e.g., cliques) or implicitly (e.g., clusters)

Cohesive subgroups

- 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
  $\Rightarrow$ Familiarity (degree);
  $\Rightarrow$ Reachability (distance);
  $\Rightarrow$ Robustness (connectivity); and
  $\Rightarrow$ 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



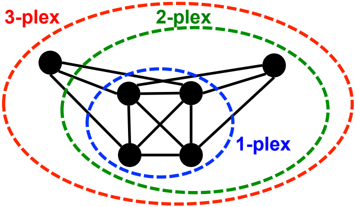
- Sufficient condition for the existence of a size-n clique

$$N_e > \frac{N_v^2}{2}\frac{(n-2)}{(n-1)}, \text{ while sparse graphs have } N_e = O(N_v)$$

- 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



$\Rightarrow$ Degrees are in the induced subgraph G', not in G

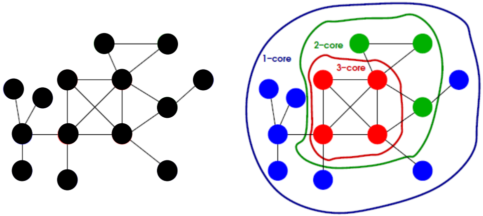
- No vertex is missing more than $k-1$ of its possible $|V'| - 1$ edges

$\Rightarrow$ 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

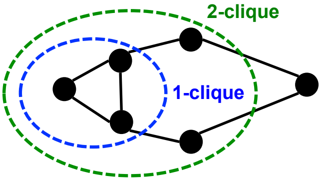


- **Def:** An induced subgraph $G'(V', E')$ is a k-core if $d_v(G') \geq k$ for all $v \in V'$, and G' is maximal
- Hierarchy: larger "coreness" $\Rightarrow$ larger degrees and centrality
- Algorithm: recursively prune all vertices of degree less than k

  $\Rightarrow$ 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

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

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

$\Rightarrow$ k-clubs are k-cliques but the converse is not true, in general

Quantifying local density

- A natural measure of density of a subgraph $G'(V', E')$ is

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

$\Rightarrow$ Quantifies how close is G' to being a clique

- $\text{den}(G')$ is just a rescaling of the average degree $\bar{d}(G')$

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

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

$\Rightarrow$ Use v's egonet G'_v, subgraph induced by v and its neighbors

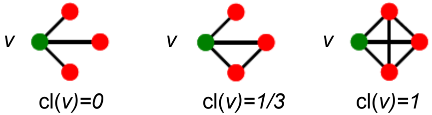
$\Rightarrow$ Density of the overall graph G is $\text{den}(G) = \frac{2N_e}{N_v(N_v-1)}$

Clustering coefficient

- Q: What fraction of v's neighbors are themselves connected?
- **Def:** The clustering coefficient $\text{cl}(v)$ of $v \in V$ is

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

$\Rightarrow |E_v|$ is the number of edges among v's neighbors

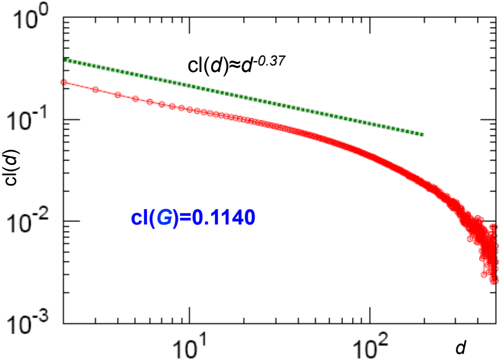


- An indication of the extent to which edges 'cluster'
- The global (average) clustering coefficient is

$$\text{cl}(G) = \frac{1}{N_v} \sum_{v \in V} \text{cl}(v)$$

Example: MSN social network

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



- Average clustering coefficient $\text{cl}(G) = 0.1140$ is large
- Compare with the Erdös-Renyi random graph model

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

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'

$$d_{V'} = |U_{V'}|$$

$\Rightarrow$ 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

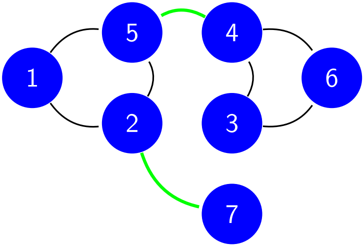
Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure

Network connectivity and robustness

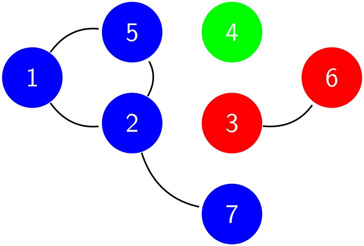
- Connectivity relevant when taking a larger, global perspective
  - Q: Does a given graph G separate into different subgraphs?
  - If it does not, a 'less robust' network is closer to splitting
- **Def:** Graph is connected if $\exists$ walks joining each vertex pair



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

Connected components

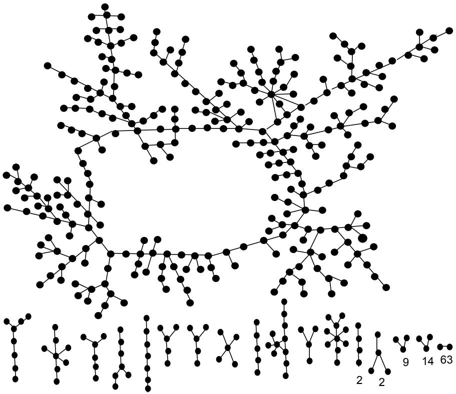
- A component is a maximally-connected subgraph



- In figure $\Rightarrow$ Components are $\{1, 2, 5, 7\}$, $\{3, 6\}$ and $\{4\}$
  $\Rightarrow$ Subgraph $\{3, 4, 6\}$ not connected, $\{1, 2, 5\}$ not maximal
- Disconnected graphs have 2 or more components
  - $\Rightarrow$ Number of components $=$ Multiplicity of eigenvalue 0 for $\mathbf{L}$
  - $\Rightarrow$ 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

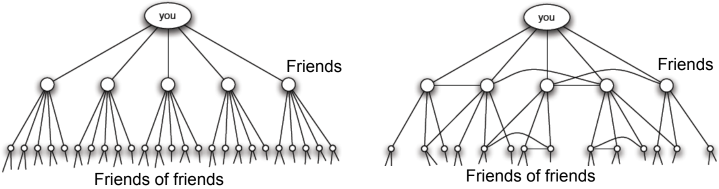
Average path length and small world

- Giant components tend to exhibit the small world property
- Small refers to the average path length

$$\bar{\ell} = \binom{N_v}{2}^{-1} \sum_{u \neq v \in V} d(u,v) = O(\log N_v)$$

Ex: facilitates spread of gossip, diseases, search for WWW content

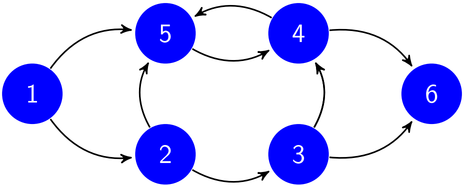
- Not too surprising that the property holds. Informal argument:



- If $d_v = d$, after h_* hops have $d^{h_*} \approx N_v \Rightarrow \bar{\ell} \approx h_* = O(\log N_v)$

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

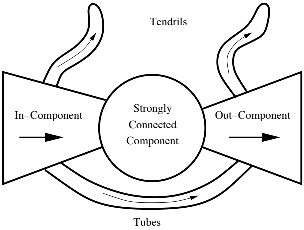


- Above graph is weakly connected but not strongly connected

  $\Rightarrow$ Strong connectivity obviously implies weak connectivity

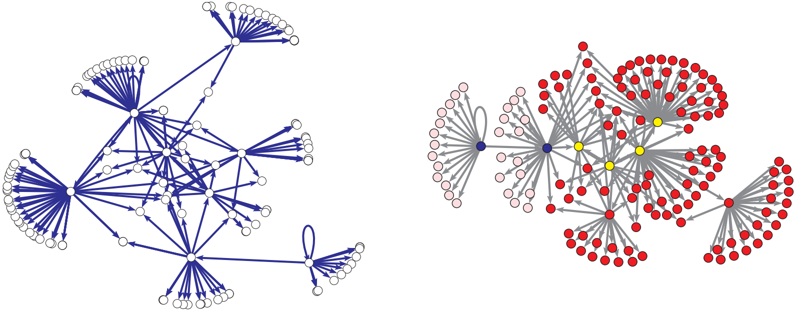
Bowtie structure of directed graphs

- 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 $|\mathrm{SCC}| \approx |\mathrm{IC}| \approx |\mathrm{OC}| \approx 56M$

Assortativity mixing

Local density, clustering coefficient and group centrality

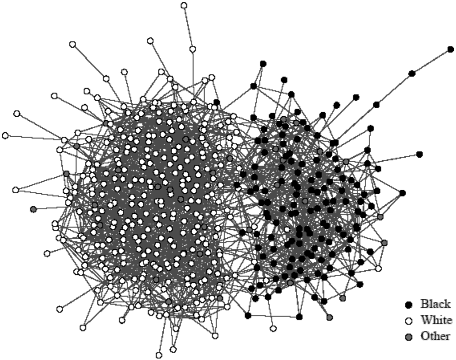
Network connectivity

Assortativity mixing

Case study: Analysis of an epileptic seizure

Assortative mixing

- People have a stronger tendency to associate with equals
  - ⇒ Tendency is called homophily or assortative mixing



- Ex: high-school students by race, bloggers by political party, ...
  - ⇒ 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

  $\Rightarrow$ $f_{i+} = \sum_j f_{ij}$ (f_{+i}) is the i-th marginal row (column) sum
- Define the assortativity coefficient [Newman'03]

$$r_a = \frac{\sum_i f_{ii} - \sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}}$$

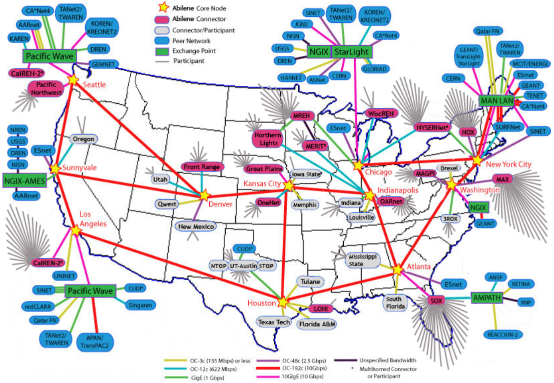
  $\Rightarrow$ $f_{i+} f_{+i}$ is the expected fraction of edges joining nodes in C_i

  $\Rightarrow$ Random edges preserving degree distribution yields $r_a = 0$
- Perfectly assortative mixing yields $r_a^{\max} = 1$, while the minimum is

$$r_a^{\min} = -\frac{\sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}} > -1$$

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

	Core	Exchange	Peer	Conn.	Part.	Conn./Part.
Core	14	6	5	17	0	16
Exchange	6	1	46	2	0	0
Peer	5	46	0	0	0	1
Conn.	17	2	0	0	203	0
Part.	0	0	0	203	0	34
Conn./Part.	16	0	1	34	34	0

- Fractions f_{ij} obtained by scaling table entries by the total of 675
- Assortativity coefficient $r_a = -0.3162$, close to $r_a^{\min} = -0.3461$

  $\Rightarrow$ 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

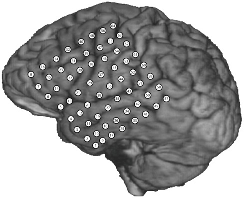
- Epilepsy is the world's most common serious brain disorder

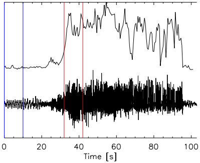
  $\Rightarrow$ Seizures, i.e., recurrent abnormal neuronal activity

- Ex: Network-oriented analysis of epileptic seizure data in humans
- M. A. Kramer et al, "Emergent network topology at seizure onset in humans," *Epilepsy Res.*, vol. 79, pp. 173-186, 2008
- 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

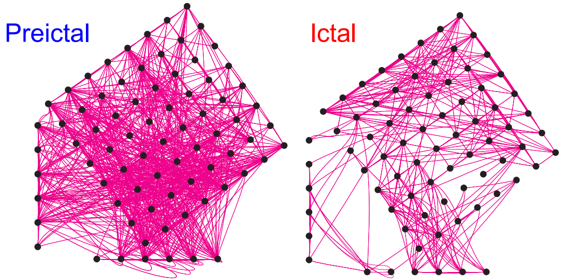




- Two 10 sec. intervals of interest for comparison:
  - ⇒ Preictal period: prior to the epileptic seizure
  - ⇒ Ictal period: immediately after start of seizure
- Top time-series is smoothed, averaged over 8 seizure signals

Network graph construction

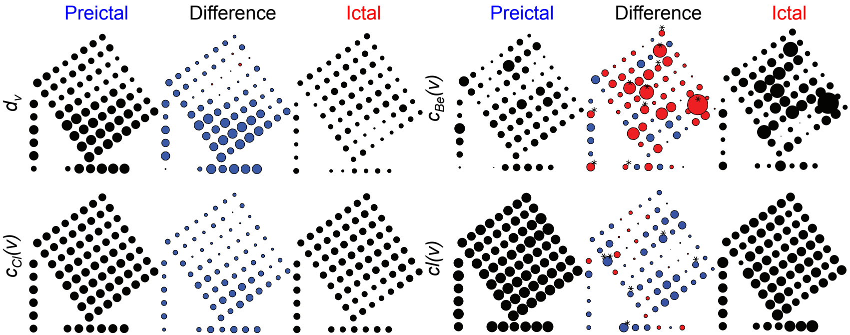
- Network $\rightarrow$ represent couplings among brain regions
  - $\Rightarrow$ Graphs for the preictal and ictal periods, for 8 seizures
- Vertices: correspond to the 76 electrodes (cortical brain regions)
- Edges: threshold correlations between pairwise 10 sec. time series



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

Summaries of network characteristics

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



- Identifies spatially localized brain regions that may facilitate seizures
  - $\Rightarrow$ 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