Community Analysis

CS 579 Online Social Network Analysis

Dr. Cindy Hood
10/28/25

Exam 1

- Most graded and posted
  - Still finalizing section 3

- Preliminary statistics – will post final statistics when calculated
  - Average in the low 70s
- Will post solutions
  - Hold exam questions until after you view the solutions
    - Unless I made an error in calculating your score

- No curving of exam, any curving done for final grade
- But, if I had to assign grades only on exam
  - A >= 85
  - 65 >= B < 85
  - 50 >= C < 65
  - E < 50

Exam Observations

Homework Assignments

- HW #4 – Chicago Community Areas + Census Data
  - You may work in groups up to 4 students (no exceptions) on this hw
  - Due date now 11/3 <-- Another Change

- Please contact TAs with questions

HW #4 Using census data to explore Chicago Community Areas (CAs)

Each student has been assigned a Community Area (1-77) to explore. You will use data from census.gov to do an analysis of your CA using census data from the ACS (block group level) and decennial census (block level). Depending on the number of block groups in your CA, you may consider the broader geographic neighborhood of your CA as well (see instructions below). The deliverables are:

1. A data-based characterization of your CA including the recent history (back to 2010) of your CA

I suggest you use compare a few variables from the decennial census in 2010 and the decennial census in 2020

- Population
- Demographics

This must be done at the block group level for all of the block groups you are including in your study

2. A description of the data (i.e. variables) that you have chosen to use as the basis of your study along with a discussion of why you chose these variables and any preliminary analysis you did to narrow the data. You will choose at least 6 variables that are available at the block group level (i.e. from the ACS).
3. An analysis of similarity of the block groups that comprise your CA. If your CA has less than 60 block groups, you will add the block groups of geographic neighbors to get at least 60 block groups. This analysis should include modeling the block groups as a network.
    a. You will describe the modeling and analysis you did, stating assumptions and justifying decisions.
    b. You will provide data-based arguments including visualizations to support why your CA is a community.
    c. You will provide data-based arguments including visualizations to support why your CA is not a community.
    d. You will provide a proposal for alternative community(ies). If your CA has 60 or more block groups, this will be a proposal for organization of communities within your CA. If your CA has <60 block groups, you will propose an organization of block groups from your CA and neighbors (resulting in analysis of 60 or more). This proposal should include visualizations.
4. If you are working in a team, there is an additional step where you will put all your data together and propose an organization of communities. Each team member must have a unique CA so if your team has duplicates, please email me asap.
5. The above should be compiled into a report that includes citations and transcripts of any AI assistance. You will submit a pdf and code.

Teams will submit one report. Be sure to describe which Cas each team member analyzed

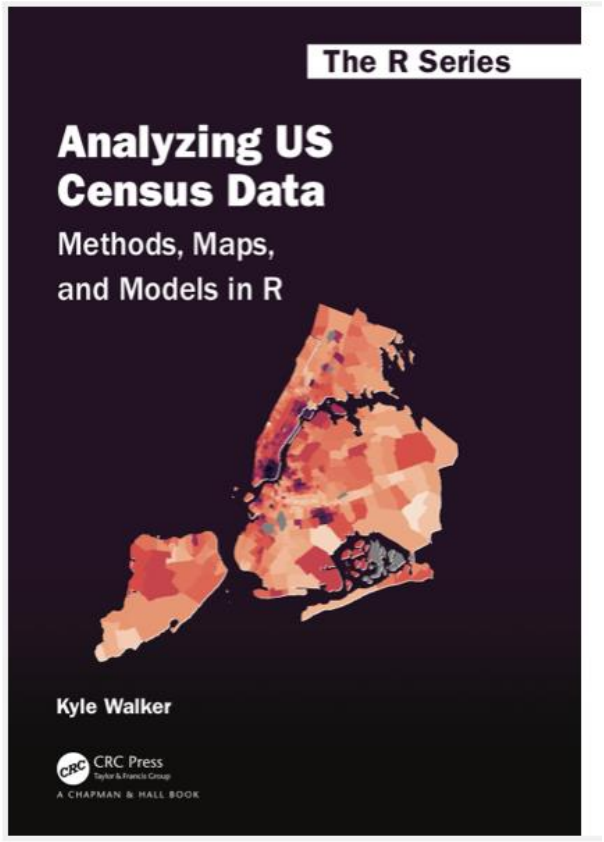
Exams and Final Project Poster Presentation

- Exam 2 – Dec 2 in class
- Final Project Poster Session – Dec 4 in class
- Online students (sections 2 and 3) will have remote options

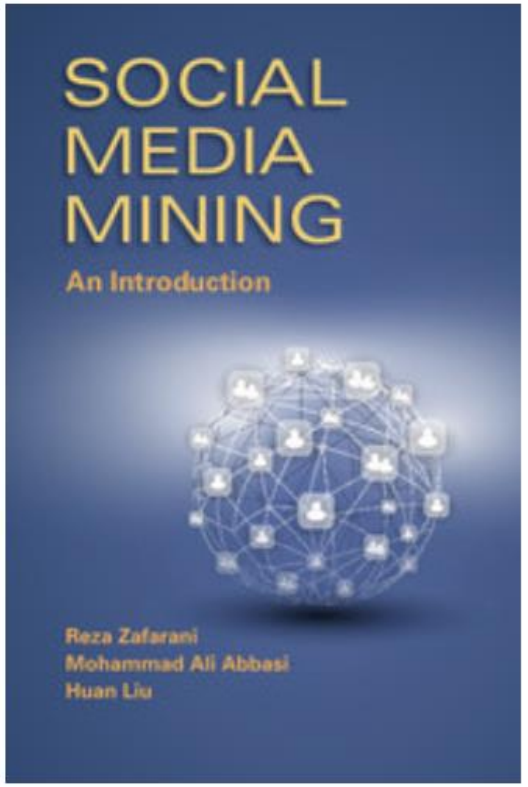
Final Project

- Extension of HW 4
- Teams of up to 4 permitted
- Proposal due 11/4 <- Change
  - Detailed requirements posted soon

References



https://walker-data.com/census-r/



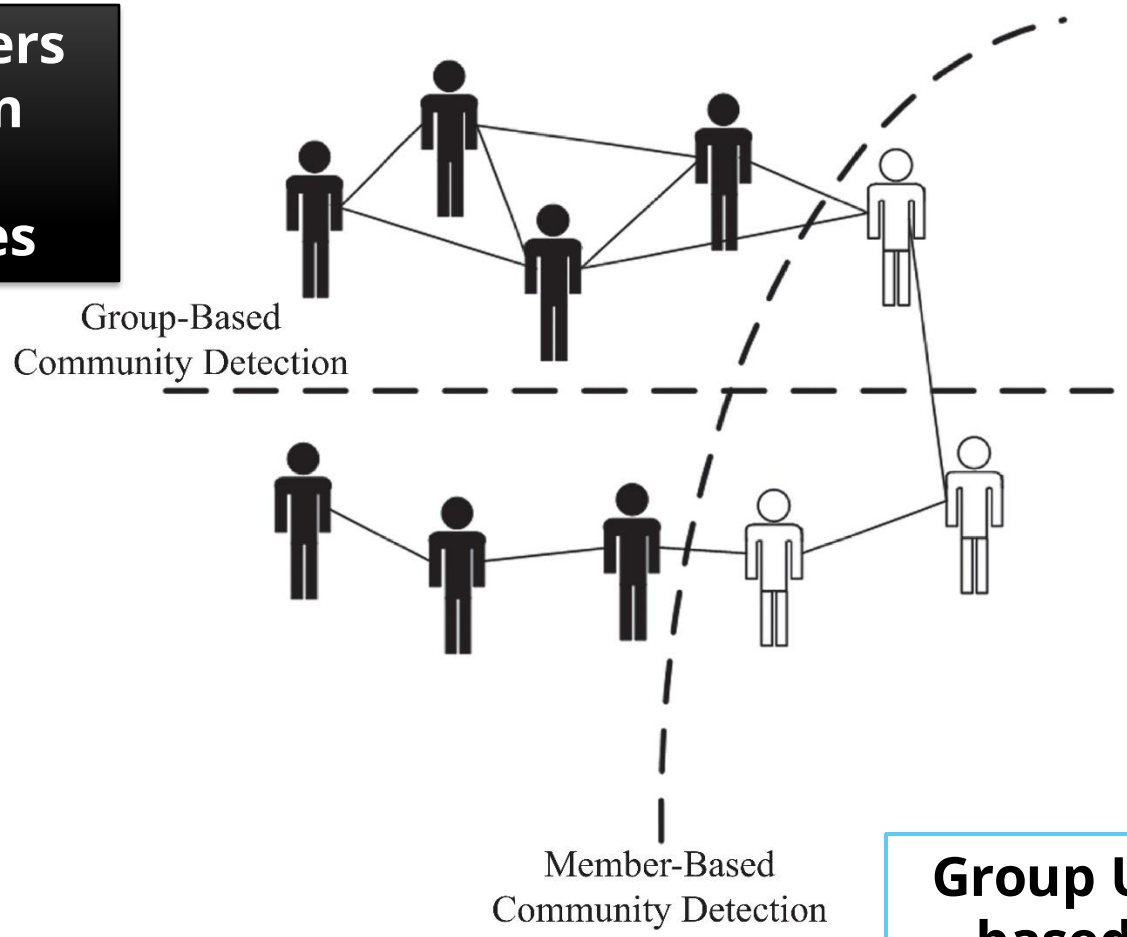
http://www.socialmediamining.info

Some additional resources

- Myatt and Johnson (2014), *Making Sense of Data I*, 2nd Edition, Wiley, ISBN: 978-1-118-40741-7
  - https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781118422007
- Networks, Crowds, and Markets: Reasoning About a Highly Connected World by David Easley and Jon Kleinberg.
  - http://www.cs.cornell.edu/home/kleinber/networks-book/
- Hanneman, Robert A. and Mark Riddle. 2005. Introduction to social network methods. Riverside, CA: University of California, Riverside
  - https://faculty.ucr.edu/~hanneman/nettext/

Community Detection Algorithms

Group Users based on Group attributes



Group Users based on Member attributes

Member-Based Community Detection

Member-Based Community Detection

- Look at node characteristics; and
- Identify nodes with similar characteristics and consider them a community

Node Characteristics

A. Degree

- Nodes with same (or similar) degrees are in one community
- Example: cliques

B. Reachability

- Nodes that are close (small shortest paths) are in one community
- Example: k-cliques, k-clubs, and k-clans

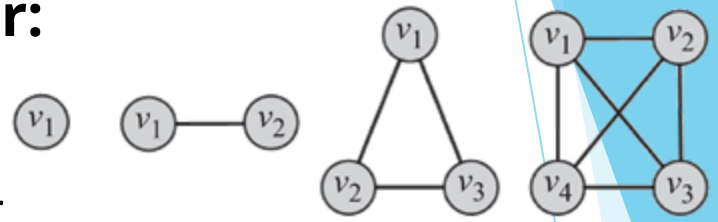
C. Similarity

- Similar nodes are in the same community

A. Node Degree

Most common subgraph searched for:

- **Clique**: a maximum complete subgraph in which all nodes inside the subgraph adjacent to each other



Find communities by searching for

1. **The maximum clique**: **the one** with the largest number of vertices, or
2. **All maximal cliques**: cliques that are not subgraphs of a larger clique; i.e., cannot be further expanded

To overcome this, we can

I. Brute Force
II. Relax cliques
III. Use cliques as the core for larger communities

Both problems are NP-hard

Cliques (Social Science Perspective)

- The idea of a clique is relatively simple. At the most general level, a clique is a sub-set of a network in which the actors are more closely and intensely tied to one another than they are to other members of the network.
- In terms of friendship ties, for example, it is not unusual for people in human groups to form "cliques" on the basis of age, gender, race, ethnicity, religion/ideology, and many other things.
- The smallest "cliques" are composed of two actors: the dyad. But dyads can be "extended" to become more and more inclusive -- forming strong or closely connected regions in graphs.
- A number of approaches to finding groups in graphs can be developed by extending the close-coupling of dyads to larger structures.
- The formal definition of a "clique" as it is used in network analysis is much more narrow and precise than the general notion of a high local density.
- Formally, a clique is the maximum number of actors who have all possible ties present among themselves. A "Maximal complete sub-graph" is such a grouping, expanded to include as many actors as possible.

https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html

I. Brute-Force Method

Can find all the maximal cliques in the graph

For each vertex v_x, we find the maximal clique that contains node v_x

Algorithm 1 Brute-Force Clique Identification

Require: Adjacency Matrix A, Vertex v_x

1: **return** Maximal Clique C containing v_x
2: CliqueStack = $\{\{v_x\}\}$, Processed = $\{\}$;
3: **while** CliqueStack not empty **do**
4: C=pop(CliqueStack); push(Processed,C);
5: v_{last} = Last node added to C;
6: $N(v_{last}) = \{v_i | A_{v_{last},v_i} = 1\}$.
7: **for all** $v_{temp} \in N(v_{last})$ **do**
8: **if** $C \bigcup \{v_{temp}\}$ is a clique **then**
9: push(CliqueStack, $C \bigcup \{v_{temp}\}$);
10: **end if**
11: **end for**
12: **end while**
13: Return the largest clique from Processed

Impractical for large networks:

- For a complete graph of only 100 nodes, the algorithm will generate at least $2^{99} - 1$ different cliques starting from any node in the graph

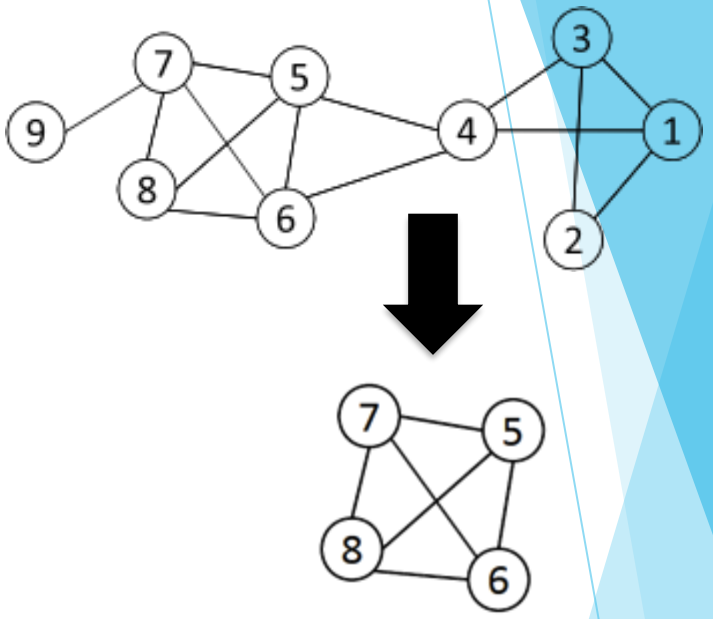
Enhancing the Brute-Force Performance

[Systematic] Pruning can help:

- When searching for cliques of size k or larger
- If the clique is found, each node should have a degree equal to or more than $k-1$
- We can first prune all nodes (and edges connected to them) with degrees less than $k-1$
  - More nodes will have degrees less than $k-1$
  - Prune them recursively
- For large k, many nodes are pruned as social media networks follow a power-law degree distribution

Maximum Clique: Pruning…

Example. to find a clique ≥ 4, remove all nodes with degree $\leq (4-1)-1=2$



- Remove nodes 2 and 9
- Remove nodes 1 and 3
- Remove node 4

Even with pruning, cliques are less desirable

- Cliques are rare
- A clique of 1000 nodes, has 999x1000/2 edges
- A single edge removal destroys the clique
- That is less than 0.0002% of the edges!

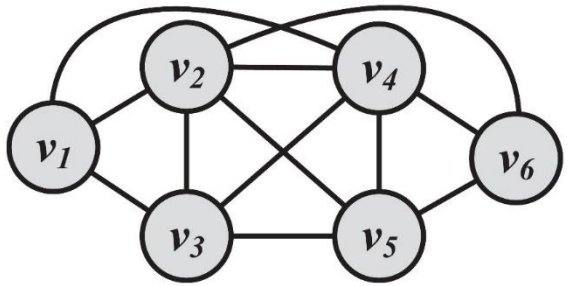
II. Relaxing Cliques

- ***k*-plex**: a set of vertices V in which we have

$$d_v \geq |V| - k, \forall v \in V$$

- d_v is the degree of v in the induced subgraph
  - Number of nodes from V that are connected to v

- Clique of size k is a 1-plex
- Finding the maximum k-plex: **NP-hard**
  - In practice, relatively easier due to smaller search space.



1-plex : $\{v_2, v_3, v_4, v_5\}$

2-plex : $\{v_1, v_2, v_3, v_4, v_5\}, \{v_2, v_3, v_4, v_5, v_6\}$

3-plex : $\{v_1, v_2, v_3, v_4, v_5, v_6\}$

Maximal k-plexes

K-Plex (Social Science Perspective)

- An alternative way of relaxing the strong assumptions of the "Maximal Complete Sub-Graph" is to allow that actors may be members of a clique even if they have ties to all but k other members.
- For example, if A has ties with B and C, but not D; while both B and C have ties with D, all four actors could fall in clique under the K-Plex approach. This approach says that a node is a member of a clique of size n if it has direct ties to n-k members of that clique.
- The k-plex approach would seem to have quite a bit in common with the n-clique approach, but k-plex analysis often gives quite a different picture of the sub-structures of a graph.
  - Rather than the large and "stringy" groupings sometimes produced by n-clique analysis, k-plex analysis tends to find relatively large numbers of smaller groupings. This tends to focus attention on overlaps and co-presence (centralization) more than solidarity and reach.

https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html

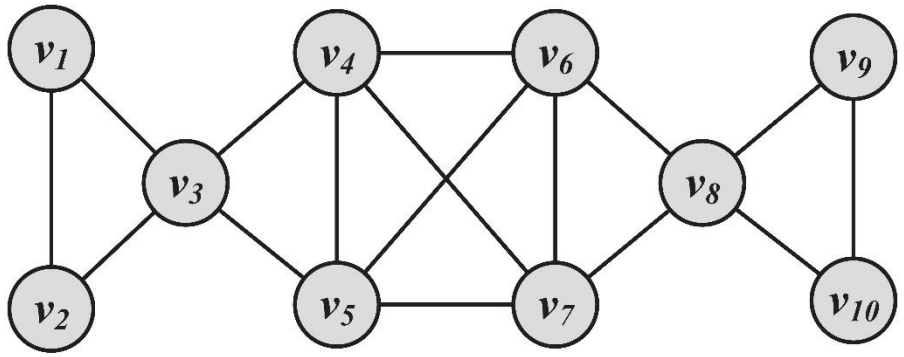
III. Using Cliques as a seed of a Community

Clique Percolation Method (CPM)

- Uses cliques as seeds to find larger communities
- CPM finds overlapping communities

- **Input**
  - A parameter k, and a network
- Procedure
  - Find out all cliques of size k in the given network
  - Construct a clique graph.
    - Two cliques are adjacent if they share $k-1$ nodes
  - Each connected components in the clique graph form a community

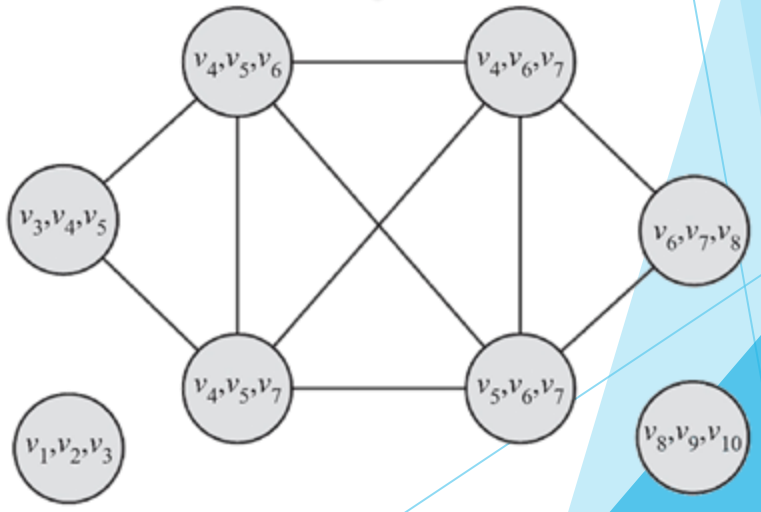
Clique Percolation Method: Example



(a) Graph

Cliques of size 3:

$\{v_1, v_2, v_3\}$, $\{v_3, v_4, v_5\}$,
$\{v_4, v_5, v_6\}$, $\{v_4, v_5, v_7\}$,
$\{v_4, v_6, v_7\}$, $\{v_5, v_6, v_7\}$,
$\{v_6, v_7, v_8\}$, $\{v_8, v_9, v_{10}\}$



(b) CPM Clique Graph

Communities:

$\{v_1, v_2, v_3\}$,
$\{v_8, v_9, v_{10}\}$,
$\{v_3, v_4, v_5, v_6, v_7, v_8\}$

B. Node Reachability

The two extremes

Nodes are assumed to be in the same community

1. If there is a path between them (regardless of the distance) or
2. They are so close as to be immediate neighbors.

How? Find using BFS/DFS

Challenge: most nodes are in one community (giant component)

How? Finding Cliques

Challenge: Cliques are challenging to find and are rarely observed

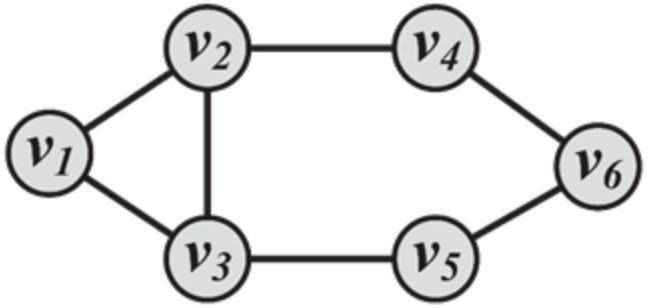
Solution: find communities that are in between cliques and connected components in terms of connectivity and have small shortest paths between their nodes

Special Subgraphs

1. k-**Clique**: a **maximal** subgraph in which the largest shortest path distance between any nodes is less than or equal to k

2. k-**Club**: follows the same definition as a k-clique
   - **Additional Constraint:** nodes on the shortest paths should be part of the subgraph (i.e., diameter)

3. k-**Clan:** a k-clique where for all shortest paths within the subgraph the distance is equal or less than k.
   - All k-clans are k-cliques, but not vice versa.



2-cliques : $\{v_1, v_2, v_3, v_4, v_5\}, \{v_2, v_3, v_4, v_5, v_6\}$

2-clubs : $\{v_2, v_3, v_4, v_5, v_6\}, \{v_1, v_2, v_3, v_4\}, \{v_1, v_2, v_3, v_5\}$

2-clans : $\{v_2, v_3, v_4, v_5, v_6\}$

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.

https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html

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. This is the n-clan approach.

https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html

Summarizing Relaxation of Clique Definition

- The n-clique and n-clan approaches provide an alternative to the stricter "clique" definition, and this more relaxed approach often makes good sense with sociological data.
- In essence, the n-clique approach allows an actor to be a member of a clique even if they do not have ties to all other clique members; just so long as they do have ties to some member, and are no further away than n steps (usually 2) from all members of the clique.
- The n-clan approach is a relatively minor modification on the n-clique approach that requires that the all the ties among actors occur through other members of the group.
- If one is uncomfortable with regarding the friend of a clique member as also being a member of the clique (the n-clique approach), one might consider an alternative way of relaxing the strict assumptions of the clique definition -- the K-plex approach.

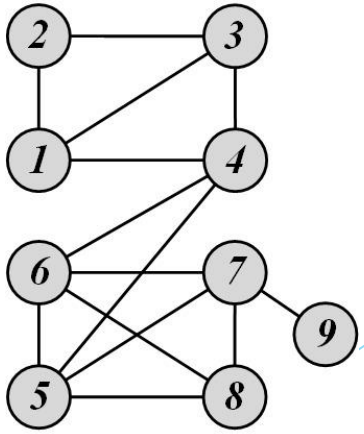
https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html

C. Node Similarity

- Similar (or most similar) nodes are assumed to be in the same community.
  - A classical clustering algorithm (e.g., k-means) is applied to node similarities to find communities.
- Node similarity can be defined
  - Using the similarity of node neighborhoods (**Structural Equivalence**) – Ch. 3
  - Similarity of social circles (**Regular Equivalence**) – Ch. 3

Structural equivalence: two nodes are structurally equivalent iff. they are connecting to the same set of actors

Nodes 1 and 3 are structurally equivalent, So are nodes 5 and 7.



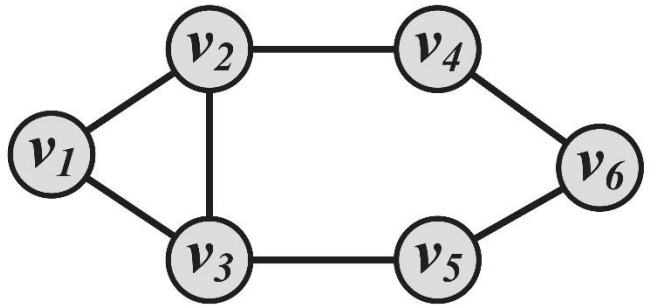
Node Similarity (Structural Equivalence)

Jaccard Similarity

$$\sigma_{\text{Jaccard}}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$$

Cosine similarity

$$\sigma_{\text{Cosine}}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}$$



$$\sigma_{\text{Jaccard}}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{|\{v_1, v_3, v_4, v_6\}|} = 0.25$$

$$\sigma_{\text{Cosine}}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cap \{v_3, v_6\}|}{\sqrt{|\{v_1, v_3, v_4\}||\{v_3, v_6\}|}} = 0.40$$