Social Network Analysis Final

Shreyans Kothari

5/4/2022

---

1) Compare the method of finding subgroups based on cliques (make sure to include a brief discussion of hierarchical clustering) vs. based on the Girvan-Newman algorithm:

(a) Explain how each method works.

(b) How are they similar?

(c) How are they different?

(d) What one benefit does each offer?

(e) In each case, how would you determine the "right" number of groups?

(f) In the end, which one would you prefer and why?

Researchers working in the social network sciences often attempt to find groups in networks based on various factors- the attributes of the nodes (explicit), their connection to other nodes (implicit), their position in the network, etc. As you might know, birds of a feather flock together, and subgroups in social network analysis seek to locate those birds that have similar feathers. Finding/discovering these groups allows researchers to make causal claims, decide on routes to administer treatments in experiments, study how certain behaviors spread in a network, etc.

One way to locate groups in a social network is to use the `bottom-up approach` ; work from the smallest units within the network (like pairs or triads) and group them together based on some predetermined factor until the groups can not be stretched any further. Locating `Cliques` is one form of finding subgroups in a social network using a bottom-up approach that is based on the idea of complete mutuality. Cliques are essentially "the maximum number of actors who have all possible ties present among themselves".[1] Cliques tend to produce very small groups because of their highly stringent criterion for gaining membership into a group: there need to be direct ties among all members of the group. Clique analysis highlights many small super-cohesive groups and doesn't provide any information about the overall group structure of the network. Thus, in some cases it is better to use `Hierarchical clustering` . Hierarchical clustering utilizes the co-occurrence matrix of clique overlaps and puts the nodes that are "closer" together, and for the ones that are "far apart" are placed far apart. Other forms of bottom-up approaches include `N-Cliques` which includes nodes who do not directly share a tie with the ego but are connected by some degrees of separation (specified by N), `K-cores` which groups nodes based on a minimum number of other nodes the each node is connected to in a potential group, and `K-plexes` which groups nodes based on the number of ties that a group does not need to share in order to be in that group.

Another way to locate groups in a social network is called the `top-down approach` where we start from the whole network and try to break it apart into components. One type of top-down approach is the `Girvan-Newman algorithm` . The Girvan-Newman algorithm employs graph partitioning or community detection to look for instances where edges between the groups occur at lower density than edges within the groups. The idea is to find very tightly-knit groups with sparse connections in-between them. This algorithm relies on edge betweenness; it looks at the edge that carries the most traffic and cuts it into components, then it looks for the next most traffic-ed edge and cuts that and this goes on until a defined stopping point based on the modularity (a measure of the structure of a graph). The algorithm removes the edges with high betweenness to separate the graph into clearly demarcated groups.

Both clique analysis and Girvan-Newman are similar in the sense that they are both recursive methods that stop after a certain condition is fulfilled. Clique analysis concludes based on k, the size of the clique. It samples a sub-network from a given network and finds a clique / group in that sub-network. In order to find a clique of size k, clique analysis prunes all nodes with degree <= k-1. This process is repeated until the entire network is put into groups.[2] Girvan-Newman, on the other hand, looks at the entire network and calculates the edge betweenness centrality for all nodes. Then, it removes those key edges that carry the most traffic in the network. Then it calculates the edge betweenness centrality for the rest of the nodes and again removes the edges with highest betweenness centrality. This process is repeated based on Q, the modularity which compares the number of internal links found in the group to the number of links that would be found if links were distributed at random with the same number of nodes and where each node keeps its degree, but edges are otherwise randomly attached. The Girvan-Newman algorithm is a type of a modularity optimization algorithm as its stopping point is based on the value of Q. The algorithm repeats calculation of betweenness-centrality and removal of edges until no edges remain.

The biggest differences between the two techniques is that one is a bottom-down approach (clique analysis), whereas the other is a top-down approach (Girvan-Newman algorithm). The bottom-down approach seeks to find out how far a close-knit relationship can be extended in a network. This allows a researcher to analyze how more complex groups and structures evolve/emerge from simple ones. Bottom-down approaches emphasize "how the macro might emerge out of the micro," i.e. they try to investigate how an individual is embedded in a web of overlapping groups in the larger network.[3] The top-down approach seeks to examine whether a tight structure in a small group can be extended outward to bigger groups. This approach starts from the whole network and identify sub-networks that are locally denser than the entirety. In a sense, this approach looks for holes / vulnerabilities / weak spots in the bigger structure of the network and uses them to decompose the network into smaller groups.[4] Another difference between clique analysis and Girvan-Newman algorithm is that clique analysis allows for overlapping groups whereas the Girvan-Newman algorithm accounts for overlapping communities using Edge-Betweenness Modularity.

Both clique analysis and Girvan-Newman are very advantageous techniques that allow researchers to explore/identify subgroups in a network. The Girvan-Newman algorithm actively analyzes the entire network to find subgroups and communities. This ensures that all of the information about the nodes and their edges are utilized in a very efficient manner to separate nodes into subgroups. Additionally, this approach allows us to identify constraints in a network under which nodes form groups/networks by operating at a level of group-selection. Since Girvan-Newman decomposes the network based on weak spots in the network, it provides researchers a high-level insight into what dynamics play in the formation of subgroups/sub-networks. Clique analysis is a very beneficial technique in that it allows us to investigate how complex social networks emerge / evolve from very simple ones.[5] By studying the dynamics of how groups emerge in a network, researchers can explore how these conditions can be applied to settings with high-level of conflict to encourage cohesion in that network.

Clique analysis determines the “right” number of groups based on k, the size of the groups- how many nodes need to be connected in order to be in a group. To find groups of size k, the method removes all nodes with degree less than k-1. In the case that we want groups of size k= 3, clique analysis will discard nodes that have either only two or only one node connected to them. Conversely, Girvan-Newman determines the “right” number of groups by calculating edge betweenness centrality and removing those edges with high betweenness centrality score. The process of calculating edge betweenness and removing edges is repeated until no edges remain- until the right number of groups is discovered.

Both clique analysis and Girvan-Newman are very useful techniques that allow us to identify subgroups in a network. However, I prefer Girvan-Newman over Clique analysis because of two reasons:

1. Clique analysis is too stringent and requires all nodes in a groups to be connected to each other- this does not always reflect reality. There exist many social networks where not all nodes are connected to each other- say in a student organization network at a college, or even in a friend network with lots of nodes. Clique Analysis, on the other hand, being a top-down method accounts for each and every node and places even the nodes not connected to other nodes in larger groups in the network. This is more reflective of what most real-life social networks look like.
2. Every network is made up of its individual nodes- each nodes carries a lot of information about that network. Leaving out nodes that have degree size less than k means that this method omits a lot of very useful information about the network. Girvan-Newman does not remove any nodes just because they are not connected to an arbitrary “required” number of other nodes.

---

2)

(a) Explain why is it so hard to separate out what are the effects of influence/contagion/diffusion in social networks vs. what are effects due to self-selection in networks?

(b) Give an example of this tension.

(c) How does longitudinal data help overcome this?

The idea of `influence` in social networks is a very rudimentary concept, according to which, people/nodes who form ties with each other will influence each others subsequent behavior. Contagion and diffusion are almost an extension of influence- they attempt to identify spread of behavior/habit or an attribute in a network in a kind of a wave. `Selection` or self-selection, on the other hand, is the idea that people who share similar behaviors are more likely to form ties than those that are drastically different. This idea of selection is based on homophily in a social network: common traits/attitudes shared by nodes. These contrasting ideas disagree on the basis of relationship formation and changes among those nodes and their relationships in a network; almost akin to the age-old question about what came first, the egg or the chicken. Selection and influence are quite intertwined and often difficult to separate. Up to a certain degree, we can reason that they are correlated- for example two people who like to play the violin become friends because they have a shared interest in the violin- this is selection. As they become better friends, they start playing more and more violin since that is what they have in common- this is influence.

It is difficult (if not impossible) to separate between the two because of a multitude of reasons that go beyond any academic theories of social network analysis, and take root in real-life relationships. It is important to point out here that people in the real-world do not form relationships based on just one common habit/behavior/attribute they share with someone else. Just because you like to run and someone else likes to run does not mean that you both will most definitely connect or form a relationship. A lot more goes into forming connections than just having that one common trait. Often, it is difficult to separate between influence and selection because if we look at two nodes who share an attribute it is impossible for us as researchers to go back in time and see if those two nodes shared that same attribute in the past as well or if that commonness arose from that connection between them. Furthermore, even if we had the ability to do that, we don’t commonly know when those nodes connected in the first place- the temporal aspect brings in a lot of complexity. There’s another aspect to it which can’t quite be measured: it is hard to identify if certain people are more predisposed to like certain things than other people. If one of my friends who likes Jazz introduces me to it, I might start liking Jazz too. This could very well be a case of influence, but who is to say that I am not predisposed to liking Jazz such that I would have found discovered Jazz even if I was not connected to my friend who likes Jazz in the first place. Following along this logic, if I am someone who is predisposed to like Jazz, i.e., I am the kind of person who “would want to like Jazz”, my friendship with the person who likes Jazz could have very well been a result of that common trait we share that made us both predisposed to liking Jazz. Or maybe since I started liking Jazz after I became friends with that one person who likes Jazz, but I am predisposed to liking it, the connection with that person is just “incidental” because I was always going to like Jazz even if I didn’t end up being friends with them.

The concepts of selection and influence become hard to separate because they are quite convoluted; we don’t have the ability to travel back in time or to another universe to see how a person would behave if things had turned out quite differently. Furthermore, our understanding of habits and behaviors, how people form those habits and behaviors, and what changes them is also extremely limited. Instead of looking at the dynamic networks as selection vs influence, we should try to find cases where the two exist together, like in the Crandall et al. (2008) study about Wikipedia editors covered in lecture 12, where selection and influence led editors to connect to each other and to update their editorial practices.

I have already mentioned a few examples that depict this ‘tension’ between self-selection and influence but to mention one other based on an actual study: the “Is Voting Contagious” study by David W. Nickerson (2008)[6] discussed in lecture 12 concluded that influence in the network led entire families to vote in an election based on just the one member of the family who opened the door to hear about the benefits of voting. According to the study, one member’s behavior influenced the rest of the people in that house to change their behavior in favor of voting. However, one could also argue that generally people get married to and start families with those who share similar ideologies about life, politics, religion, etc. In the case of a married couple, if one spouse is more proactive about voting because of their beliefs, I would assume that their spouse would also feel the similar way about voting. Thus, in the study, the effect that was captured could have very well been the case that the person who opened the door- if they were more inclined to vote just off their own accord and not because of the person who knocked on their door to explain the benefits of voting (something that can’t be measured) -had family members who were inclined to vote because of a common trait they both share.

Longitudinal network data comes in handy to study the differences between self-selection and influence because it is looks at the same nodes and their attributes over time: it captures how people’s ties with each other evolve along with their own attributes. We need a network to be in disequilibrium to be able to see these concepts in effect but any network at one given point would be in a sort of an equilibrium, and thus would restrict us from identifying selection and/or influence in the network. Thus, we need to see how people change and evolve over time to understand and separate between influence and selection. In a network with nodes who share an attribute are connected, it is hard for us to conclude that those two nodes formed that connection because of that common trait. It could very well be the case that the two nodes already

had that attribute before they were connected but their connection to each other amplified that one particular attribute- but if our data is only about nodes in a network at one point in time, that would be difficult for us to identify. Longitudinal data allows us to account for the temporal complexities that we discussed above that restrict us from exploring how the networks change over time.

---

3) A researcher wanted to see how much some networks were like "small worlds." She found that the Lazega friendship network (which is n=71 lawyers, with unvalued and directed ties) had an average shortest path length of 1.54 and a clustering coefficient of 0.51. In contrast, she found that the LinkedIn network (which is n=360 million users, with unvalued and binary ties) had an average shortest path length of 3.33 and a clustering coefficient of 0.25. Facebook (with n=1 billion members, with unvalued and binary ties) had an average shortest path length of 4.76 with a clustering coefficient of 0.18.

(a) What is a mean shortest path length?

(b) What is the clustering coefficient?

(c) Based on this information, which one of these networks is the most "small world" and why?–Explain your answer.

[Note that you do not have enough information to calculate Q, so simply make an argument about how you would decide on which is the "smallest" world, based on the information you do have.]

In the context of social networks, a path is a series of steps from one node to another. The length of the path is the number of edges it contains. Shortest path, in this case, refers to the least number of steps it takes to get from one node to another node in a network. The average shortest path is the average distance, for each node, of the steps it takes to get to every other node in the network. If there are unconnected nodes in the network, the geodesic distance between unconnected nodes is treated as a length greater than that of any real distance in the data.

The clustering coefficient determines how many cohesive subgroups exist within networks. It is calculated as the average of each nodes' neighborhood density; the weighted clustering coefficient is weighted by the size of the ego network. It essentially captures the percentage of triads in a network that are closed. It ranges between 0 and 1; when the network is dense (many nodes are connected to each other), the clustering coefficient is high and closer to 1. In the case of a sparse network, the clustering coefficient is low and closer to 0.

The concept of small world comes from the idea that a stranger you randomly run into has something in common with you, be it a mutual friend/family, background, etc. This invokes a feeling of a "small world"; because of how big the world actually is, you would not expect to find a lot of commonality with a complete stranger. A social network is referred to as a small-world network if any two random nodes in the network can reach each other through only a few other nodes. [7] According to Watts & Strogatz (1998), there are two main conditions for a graph to be called a small-world network:

1. It needs to have high clustering, as measured by the correlation coefficient
2. It needs to have low average path length, as measured by the Avg. shortest path length

Based on this definition, the idea of a small-world network can be quantified using the small world quotient (Q), which is the ratio of the correlation coefficient and the average path length.

Looking at the Lazega Network, the LinkedIn Network and the Facebook Network:

Lazega Network:

- n = 71
- Avg. path length = 1.54
- Clustering coefficient = 0.51

LinkedIn Network:

- n = 360 million
- Avg. path length = 3.33
- Clustering coefficient = 0.25

Facebook Network:

- n = 1 billion
- Avg. path length = 4.76
- Clustering coefficient = 0.18

Out of these three networks, the Lazega network has the smallest Avg. path length (1.54) and the highest Clustering coefficient (0.51), which would make the Lazega Network the most small-world. LinkedIn has the next smallest Avg. path length (3.33) and the next highest Clustering coefficient (0.25), which would make the LinkedIn network the next most small-world. Finally, Facebook has the biggest Avg. path length of the three (4.76) and the smallest Clustering coefficient (0.18), which would make the Facebook network the least small-world of the three. This would have been the case if we didn't have one additional piece of information, the number of nodes in the network.

The Facebook network has about 1 billion nodes, which is almost 2.778 times bigger than the LinkedIn network and about 14084507 times bigger than the Lazega Network. Despite the sheer size of the Facebook network, the Average path length is still pretty low at 4.76 and the Clustering coefficient is relatively high at 0.18. Both the Average path length and clustering coefficient are pretty susceptible to the size of the network, accounting for which would lead to the conclusion that the Facebook network is the most small-world of the three.

---

1. Hanneman, R. (n.d.). Introduction to social network methods. Introduction to SOICAL network methods: Chapter 11: Cliques and sub-groups. Retrieved May 6, 2022, from https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#(https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#):~:text=The%20formal%20definition%20of%20a,possible%20ties%20present%20among%2 ↩
2. Clusters and communities . Rensselaer Polytechnic Institute. (n.d.). Retrieved May 6, 2022, from https://www.cs.rpi.edu/~slotag/classes/FA16/slides/lec07-comm.pdf (https://www.cs.rpi.edu/~slotag/classes/FA16/slides/lec07-comm.pdf) ↩

3. Hanneman, R. (n.d.). Introduction to social network methods. Cliques and sub-groups. Retrieved May 6, 2022, from https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#Bottom (https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#Bottom) ↩
4. Hanneman, R. (n.d.). Introduction to social network methods. Cliques and sub-groups. Retrieved May 6, 2022, from https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#topdown (https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#topdown)↩
5. Hanneman, R. (n.d.). Introduction to social network methods. Cliques and sub-groups. Retrieved May 6, 2022, from https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#Bottom (https://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html#Bottom)↩
6. Nickerson, D. W. (2008, February 1). Is voting contagious? evidence from two field experiments: American Political Science Review. Cambridge Core. Retrieved May 6, 2022, from https://www.cambridge.org/core/journals/american-political-science-review/article/abs/is-voting-contagious-evidence-from-two-field-experiments/8C2E64552D946C87FD062DD2CCD9054E (https://www.cambridge.org/core/journals/american-political-science-review/article/abs/is-voting-contagious-evidence-from-two-field-experiments/8C2E64552D946C87FD062DD2CCD9054E)↩
7. Kleinberg, J. (n.d.). Small-world phenomena and the dynamics of information. Cornell University Department of Computer Science. Retrieved May 6, 2022, from https://www.cs.cornell.edu/home/kleinber/nips14.pdf (https://www.cs.cornell.edu/home/kleinber/nips14.pdf)↩