Social Network Analysis 2022

Class 9 – Summary & Feedback

What We’ve Done So Far ~ Basics

- How to create a network in R:
  - From a manual command, a list of edges, or data from a social media API like Twitter.

- How to ”prune” a network:
  - Removing isolates, identifying and removing small connected components.
  - Removing or hiding unimportant node, edges, or labels.

- How to calculate basic statistics about a network:
  - Network size, diameter, mean distance, edge density.

- Similarly, how to calculate statistics which tell us about each node’s role in the network:
  - Indegree / Outdegree, and various measures of centrality

What We've Done So Far ~ Communities

- Community Detection
  - How to use various algorithms (Fast-Greedy, Infomap, etc.) to find the densely-connected communities that exist within your overall graph.
  - How to "trim" those communities – focusing on the most important ones, or combining smaller communities logically to allow for easier analysis.
  - How to separate out communities from the overall graph, so you can analyse them separately – finding their most important / central nodes, etc.

- Community Identification
  - Using a combination of network analysis (centrality algorithms etc.) and other techniques like text analysis to interpret communities and uncover their identities or unifying properties.

What We've Done So Far ~ Visualisation

- Network Layout
    - Using a variety of algorithms (Fruchterman-Reingold, Kamada-Kawai, Graphopt etc.) and parameters to find the optimal layout for viewing your network.
    - Saving that layout in an R object, so you can reliably re-create it as you work on other aspects of the visualisation.
- Colours, Sizes and Transparencies
    - Setting the colours, sizes, transparency levels and other aspects of your nodes and edges to create an easy to read visualisation that highlights key aspects of your data.
- Selective Plotting
    - Plotting only the most important aspects of a graph – key nodes, edges and labels – in order to focus viewers' attention on the key points of your research.

Today…

- Today we’ll look at one final tool that you may wish to use as you work with networks – a free piece of software called Gephi, which is primarily used by people working with biological networks, but has many powerful features that can be useful to us as social scientists as well.

- This class will also be an opportunity for feedback and any questions you may have about your final research project for the module – I’ve left the class schedule quite open so there should be plenty of time for anything you’d like to ask or discuss.

General Feedback

Final Assignment

How much data is enough data?

- This question depends very much on what kind of data you’re working with, and what you hope to find out.

- If you’re doing a small-scale study of the behaviour of a handful of entities (people, companies, Twitter accounts, etc.), then a relatively small number of network edges can still be enough – a few hundred, or possibly as few as several dozen.

- If you’re working with something like a retweet network on a popular topic, I’d suggest getting as much data as you can, within reason. The structure of your network will be more clear, and the communities more clearly defined, when you have more data to work with.

A note on retweet network data…

- We collected 18,000 tweets using the keyword “vaccine”, and were able to construct a fairly interesting network from that. However – recall that this data only spanned a **four hour period**.
- Consequently, it was strongly focused on North America (as it was very late at night in Europe when I collected the data), and cannot be considered to be representative of the debate overall!

- Students who want to examine similarly popular and contentious keywords – some of you have proposed things like “refugees”, “Ukraine”, “Putin”, or “LGBT” – need to be aware of this limitation. If possible, try to collect more data (using the rtweet “retryonratelimit” option) to help overcome this problem and get a more representative data set.

A note on hashtags

- Several of you are proposing to gather data based on #hashtags.

- This is a common approach in social science, but to be honest… I think it might be an idea that's mostly coming from older social scientists who don't actually use social media very much.
- The decision to use a #hashtag instead of just writing out a keyword normally is quite unusual, and may be suggestive of a very specific perspective – in other words, searching for "#vaccine" instead of just "vaccine" might introduce a major bias to your data, because people using the hashtag are more likely to see themselves as campaigners or influencers on this issue.

- Note that a search for "vaccine" would also return every instance of "#vaccine", as these are sub-string searches!

Data Collection

- Several of you have decided to supplement your data collection by adding extra hashtags or keywords to your searches.
- This is fine – it's often a a totally reasonable approach that can help you to identify a broader community and discourse around your chosen topic.
- However: Be careful to make sure that you're choosing keywords that actually reflect discussion of the same topic.
  - If your keywords are too different, the data for them may not overlap – so you end up with a lot of connected components – i.e. "subnetworks" in your data that are totally disconnected from each other.

Connected Components

- It's worth revising again the difference between connected components and communities.
- Connected components are *separate networks* that exist in your data – there are no connections between them.
  - They usually indicate that you've gathered data for multiple unconnected topics, or just some data that's peripheral to the *main* conversation about a topic.
- Communities are *clusters of tightly-connected users within a single network*. They are connected to the other communities – they're not separate networks entirely.
  - They indicate "closeness" between a group of nodes; for example, a group of characters in a book who interact with each other a lot, or a group of people on Twitter who mostly agree with each other.

Detecting Connected Components

- While there are many different ways to detect communities (which will give you quite different results), connected components are a strictly defined concept. Any algorithm to detect them will give the same results.
- The only setting you can use in igraph is "**min.vertices**" – but this doesn't actually change how the function detects connected components. It just filters the list of components that's shown to you.
    - In other words, min.vertices is just a setting to help you see the components in a graph by hiding really tiny ones.
    - By setting min.vertices to 2, for example, you'd hide all the isolates – individual disconnected nodes.

Best Practices with Connected Components

- Generally speaking, when we're analysing a network we want to ignore any small components and focus exclusively on the largest component – called the **giant component** – since this is where the actual community we're interested in can be found.
- In a situation where there's one giant component (80% - 90% of your nodes), you can simply extract that component and ignore everything else.
- However, you may find that your network contains two or more similarly sized large components. **This usually indicates a problem in your data collection / specification.** It means you gathered data related to topics or communities that aren't connected to each other at all.
  - You won't be able to analyse this data in a single network – you'll need to extract each of the large components and analyse them separately. You should also try to understand and explain why they were disconnected from each other; why didn't your nodes have any connections?

Assignment 9

- Your assignment for tonight is to prepare a presentation of a research plan outlining your final project for this module.
- This plan should include:
  - Your research question (and a brief introduction of your topic).
  - Your hypothesis (this can be quite broadly phrased)
  - Details of the data you're going to use and your progress in acquiring it
  - Your intended research methodology
- This research plan should be presented in tomorrow's class.
  - You can prepare some slides for your presentation if you wish.
  - Students who cannot present in class for any reason may submit a research plan over email instead.