CPSC 572/672: Final Project Report

Each of these sections must be present in your final report. I present them in the same way I suggest you structure your report, but you may structure your report at your own discretion given everything I stipulate below is present. Read this document very carefully, and refer back to it often as you prepare your project:

Project title
>Make it concise, specific, and descriptive

Team
>Include everyone's name and UCID, and whether you are taking CPSC 572 or CPSC 672.

Project summary (~200-300 words)
>This should be structured like an abstract to a paper, and aimed at someone who has a scientific education, but is not in the field (of network science or of your specific project).
>Make sure to include brief descriptions of:
>> - The domain and the problem you are trying to solve / question you are trying to answer
>> - The network
>> - Your main results
>> - Implications / future outlook

Research questions
>2-4 clearly stated research questions. These should strike a balance between being 'big picture' goals, yet achievable through the approaches you will take. Briefly lay out your intended approach to address each question.

Introduction (~500 words)
>Appropriately frame your project in the existing literature. This section must be properly referenced.

Dataset description
- Raw data – where you got it (reference source), how you got it, what it looks like
- Data cleaning and data wrangling (if appropriate)
- Network construction
- Clear definitions of nodes and edges, and what type of network you have
- Clear definition of any metadata (e.g. edge weights and what they represent)

Basic statistics
- Number of nodes and edges
- Number of connected components
- Degree distribution (including fit)
- Clustering coefficient
- Path length

Network visualization

You may choose to show multiple visualizations. You must include one that covers the entire network in one figure. Be careful and selective – one thoughtful figure is better than 10 variations with little attention to detail. This said, for those of you with large networks especially you may want to display subsets as well the whole network visualization.

Results

Here is where you should systematically address each of your research questions. As appropriate, utilize ideas from class such as motifs, centrality measures, community structure, and/or spreading. You are **not** expected to try everything (in fact, that would reflect badly on you as it lacks clear direction). You **are** expected to try at least one sophisticated approach and ideally more. Make good use of visualizations to communicate your results.

Note: This section will be highly variable between teams. Get in touch with me if you are unsure if you have enough or too much content.

Note: The team going above and beyond may define their own appropriate metric, growth model, or other additions or extensions to the material we learned in class.

Comparison to suitable null model

You must create at least one ensemble of randomized networks as a benchmark for your network. This may be for example an ER network, a degree preserved randomization, or a geometric randomization. I recommend an ensemble of 1000 networks. Calculate clustering and path length (average and standard deviation), and compare to your network. If you are not using a degree preserving randomization, show the randomized degree distribution as well. Where appropriate, repeat your approaches from the Results section on this ensemble of networks and compare to your network. Note: I have written this in its own section to draw your attention to its importance. However, the narrative will flow best if you weave it into Basic statistics and Results. You can also describe your generation procedure in Methods.

Discussion (~300 words)
- Discuss your results and where appropriate frame them in existing literature.
- What did you learn about your network?
- Did you successfully answer your research questions? If not, why not?
- Are there any limitations inherent to your data or your approach?
- What would future work want to consider?

Methods

Here is where you can provide detail on calculations, metrics, and approaches you utilized in Basic Statistics and Results. List and reference any software and/or libraries you used, and blocks of code you edited. (It is ok to use code from online sources, in fact I encourage not "reinventing the wheel". Just be explicit about it and reference it.)

Code

You can organize your code however you prefer. I suggest Github – both for its functionality and for the opportunity to build your portfolio as you prepare for your future careers. I suggest you work with Python notebooks, though you may use whatever language you prefer. When you submit your report, include a link to a repository (Github, Google Drive…). Make sure your code is well commented. You will not be graded on your code as such, but it will inform me of: (i) the complexity of your data and approach; (ii) what you wrote yourself and what you leveraged from elsewhere (you **must** properly cite **any** code you did not write yourself). In other words, it will help me give you credit for the work you put in behind the scenes. Some datasets are more challenging than others to work with, and I will recognize that. A final point to bear in mind is that most journals now require code to be made available to enhance reproducibility – it's just good science.

General comments:

- This is a formal piece of scientific writing. You will be assessed on the quality of your writing, so take care with it and seek help if you need to. The University has excellent resources, including workshops through the Student Success Centre.

- You already covered some components in your project proposal. You can copy these components as is if you are satisfied, or you can edit them based on feedback you received and what you have learned since writing them the first time.

- Any word counts provided are for guidance only and to set expectations for the amount of content. Do not worry if you are a little over or under. (Do worry if you are a lot over or under.) You will also notice that I do not provide word counts for all sections – this is deliberate. E.g. the dataset description is highly individual, it just must be thorough and contain each of the points listed here.

- References may be in any format, as long as they are consistent and complete. I prefer numbered styles as lengthy textual references can disrupt the flow of the report.

- Everything you present must be interpreted. It is not enough, for example, to say "average clustering is 0.34". You need to tell us what this means – is it higher or lower than expected? What does it say about the network?

- Do not include things for the sake of it – it will show you don't understand your project. Pick approaches and methods that make sense on your network and shed light on your research questions.

- Bear in mind that teams of three and teams of graduate students will be expected to produce more in the way of results than teams of two and teams of undergraduate students. There will also be higher expectations for writing quality.

- You have been provided with an extra document to discuss everyone's contributions. In extreme cases, I am prepared to award different project grades for different team members. Please don't be an extreme case, it is unpleasant for everyone involved. Additionally, recall that your contributions are worth 5% of your grade. We have many opportunities throughout the term to observe your work, though this part of the grade is primarily decided by this peer assessment that each team member will complete individually.

- Seek advice early and often. I have tried to make this as systematic as possible for you, but at the end of the day it is a research project. There will be many directions you can take, and many possible ways of presenting your work. We are here to support your success.

- For the exceptional project and motivated team, it may be possible to discuss publication of your findings.

Deadline: Wed. 9th Apr. 11:59pm Calgary time