**About this Assignment:**

Centrality measures can be used to predict (positive or negative) outcomes for a node.

Your task in this week’s assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes.  As an additional constraint, there should be at least one categorical variable available for each node (such as “Male” or “Female”; “Republican”, “Democrat,” or “Undecided”, etc.)

In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

For this week’s assignment, you are not required to actually load or analyze the data.  Please see also Project 1 below.

You may work in a small group on the assignment.   You should post your document to GitHub by end of day on Sunday.

# Table of Contents

[Table of Contents 1](#_Toc42946151)

[Summary: 2](#_Toc42946152)

[Datasets 2](#_Toc42946153)

[Sources: 2](#_Toc42946154)

[Categorical Feature 3](#_Toc42946155)

[High Level Plan 3](#_Toc42946156)

[Data Loading & Analysis 3](#_Toc42946157)

[Degree Centrality 3](#_Toc42946158)

[Hypothetical Outcome 3](#_Toc42946159)

# Summary:

The KWJ team decided to focus on the Marvel Comics dataset available from Kaggle. None of us are hard-core Marvel enthusiasts, but we’ve seen several of their blockbuster movies over the past 10 years. And, “back in the day” when comic books were still sold in stores we’d occassionally browse through a Spiderman or Captain America comic to read their latest adventures.

We also have access to subject matter experts (SME’s) if the need arises -- one team member’s 16 year old son has spent a large portion of his life studying Marvel. And, Quora is a hotbed of Marvel enthusiasts that could be referenced as SME’s.

It is also our expectation, subject to change, that the work we complete for this assignment will serve as a starting point for Project 1 for this course.

# Datasets

## Sources:

|  |  |  |  |
| --- | --- | --- | --- |
| *#* | *Dataset* | *Description* | *Challenges and Potential Issues* |
| 1 | Kaggle’s Marvel Universe Social Network [link to site](https://www.kaggle.com/csanhueza/the-marvel-universe-social-network) | This is our primary dataset consisting of three files – one of nodes, one of edges, and one called the “hero network”. | This dataset does not provide any categorical variables, so we’ll have to use our other datasets to obtain this. |
| 2 | Kaggle’s Marvel Superheroes  [link to site](https://www.kaggle.com/dannielr/marvel-superheroes) | This dataset consists of 8 csv files with a wealth of categorical information about each character. | One of the challenges is that there is not a consistent key between the 8 csv’s.  Another challenge is how to match any categorical variables back to our primary Social Network dataset. |
| 3 | Marvel Fandom  [link to site](https://marvel.fandom.com/wiki/Category:Characters_by_Gender) | Our primary interest in this site is it provides the gender of each character. | Retrieval method would have to be web scraping and no team members have experience in this.  Matching back to Social Network dataset would have to be by character name and same names are probably written differently |
| 4 | Five-Thirty-Eight’s Comic Characters  [link to site](https://github.com/fivethirtyeight/data/tree/master/comic-characters)  [download link](https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/marvel-wikia-data.csv) | This dataset was used for a story called “Comic Books are Still Made by Men…”.  It appears to have been derived from the Marvel Fandom site, but more research is needed. | Similar challenges to what has previously been mentioned for the Marvel Superheroes dataset. |

## Categorical Feature

Once we solve the challenges noted above with matching the supporting datasets to our primary Marvel Social Networking dataset, we’ll have the following categorical features to consider using: Alignment, Gender, Haircolor, Eyecolor, and Race. (see marvel\_characters\_info file).

Our going in assumption is that we’ll use Gender as our categorical feature, but once we explore the dataset and estimate effort & timeline we may substitute another feature, or add additional features.

# High Level Plan

## Data Loading & Analysis

Here is the approach we’ll take to loading and analyzing our data…

1. *Initial focus* is on the Social Network (dataset #1 above).
   1. Load the node and edge files
   2. Perform an analysis similar to what we’ve done in the first two assignments of this course.

1. *Analyze datasets 2 – 4* and decide how to best match the categorical data to the social network nodes.
2. *Decide* if we stay with using Gender or switch to some other categorical feature.
3. *Validate* if our hypothesis (see “Hypothetical Outcomes” below) can be tested via our data.
4. *Apply degree centrality* to arrive at the edges count of each male and female character’s node.
5. *Run a t-test or z-test* to assess if the difference is statistically significant

## Degree Centrality

Here is a brief refresh on what degree centrality is excerpted from [ScienceDirect’s site](https://www.sciencedirect.com/topics/computer-science/degree-centrality):

*Degree centrality* is one of the easiest to calculate. The degree centrality of a node is simply its degree—the number of edges it has. The higher the degree, the more central the node is. This can be an effective measure, since many nodes with high degrees also have high centrality by other measures.

*Degree centrality* is a simple count of the total number of connections linked to a vertex. It can be thought of as a kind of popularity measure, but a crude one that does not recognize a difference between quantity and quality. Degree centrality does not differentiate between a link to the president of the United States and a link to a high school dropout. Degree is the measure of the total number of edges connected to a particular vertex.

## Hypothetical Outcome

1. Male characters are more popular than female characters
2. Others….