Name: Joseph Oluwasanya

Supervisor: Prof. Mark Roantree

Course Code: CA336

Centrality/Community Detection Report

1 Introduction

Dataset name: The Marvel Universe Social Network.

This dataset was found on Kaggle (link below [1]). The dataset is split into 2 CSV files: “nodes” and “edges”. It consists of two node labels, ‘hero’ and ‘comic’. The ‘hero’ label relates to nodes that represent characters in the Marvel Universe. These can be any character including villains. The ‘comic’ label relates to nodes that represent comics released by Marvel Comics or Timely Comics Inc., their predecessor. There are 12651 comics and 6431 heroes with approximately 100,000 relationships in total.

Using this dataset, I aim to analyse the marvel universe network in terms of centrality, finding which characters are most prominent in the comics. To do this, I will use several centrality algorithms (Degree, Closeness, Betweenness), comparing top results considering the meaning of centrality for these different methods.

In the next section, I will discuss the data cleaning and preparation process and the graph structure.

2 Data Preparation

First, I had to clean the data.

node	type
2001 10	comic
2001 8	comic
2001 9	comic
24-HOUR I	hero
3-D MAN/	hero
4-D MAN/	hero

Figure 2.1. Original spreadsheet

The hero nodes have a name attribute that is in the structure HERO_NAME/REAL_NAME in all caps. For example, ‘SPIDER-MAN/PETER-PARKER’. Using excel, I separated the hero and comic node types into separate CSV files. I then separated the hero names from HERO_NAME/REAL_NAME into two separate columns, ‘hero_name’ and ‘real_name’ which are capitalized correctly.

node	type	hero_name	real_name
24-HOUR I	hero	24-Hour Man	Emmanuel
3-D MAN/	hero	3-D Man	Charles Chan
4-D MAN/	hero	4-D Man	Mercurio
8-BALL/	hero	8-Ball	
ABBOTT, J	hero	Abbott, Jack	

Figure 2.2. Cleaned hero spreadsheet example

Many characters are not given real names in the dataset. For example, 8-ball’s real name is ‘Jeff Hagees’ but this isn’t recorded in the dataset. Conversely, some characters were given real names and not their ‘hero names’ but because of the structure of the files, every character has a hero name stored. if they have no ‘hero name’, then their real name is stored as ‘hero name’. For example, Jack Abbot is a supporting character in the

‘Daredevil/Spiderman #1’ January 2001 issue (DD 1 in the dataset) and isn’t given a hero name.

I didn’t make any changes to the comics.csv file but noted the styles for comic names are varying. Many of the Avengers comics are labelled as ‘A [number]’ where A is an abbreviation for ‘Avengers’ and number is the issue number. There the general naming format appears to be ‘ABBR1/ABBR2 [number 1]/[number 2]’ where ABBR1 is an abbreviation of the comic name, ABBR2 is another similar abbreviation and numbers 1 and 2 refer to issue number details. Many comics don’t have an abbr2 and number 2 or may have one and not the other. There were no weights in the graph dataset originally. I considered adding random weights in a range, but this would skew results later so instead I assigned equal weights of 1 to all nodes to allow algorithms considering weight to run.

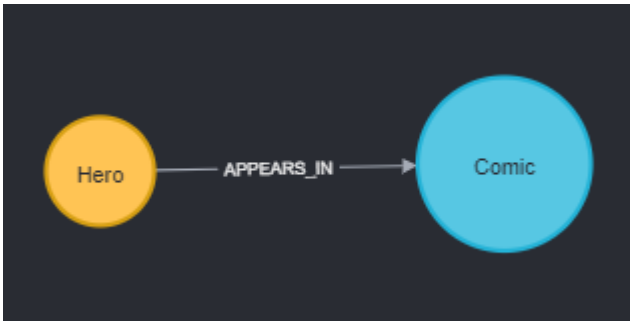


Figure 2.3. Database schema

The ‘hero’ nodes have attributes ‘char_name’ which is ‘hero_name’ from the excel sheet and ‘real_name’. Below is a snapshot of the graph giving an idea of the structure. As shown here, a hero usually appears in multiple comics. Each of those comics also features

many heroes. The graph is too big to view the entire structure but here we can see that ‘Cobra’ and ‘Grey Gargoyle’ both appear in ‘FF 3’ (Fantastic Four issue #3)

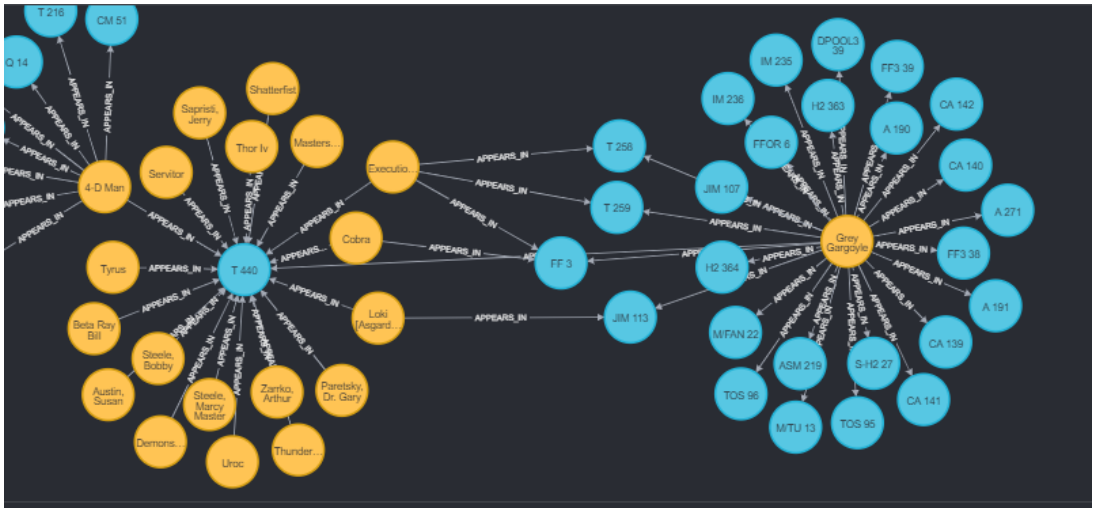


Figure 2.4. Graph snapshot

3 Experiments

3.1 Degree Centrality

Degree Centrality of a node is the number of outgoing or incoming connections. For the context in which I am using this, I will only consider outgoing relationships (Directed) and hence ‘hero’ nodes. In the context of my graph, the degree centrality is the number of comics a hero appears in.

name	n_comics
Spider-Man	1577
Captain America	1334
Iron Man	1150
Thing	963
Thor	956
Human Torch	886
Mr. Fantastic	854
Hulk	835
Wolverine	819

Invisible Woman	762
Scarlet Witch	643
Beast	635
Dr. Strange	631
Watson-Parker, Mary	622
Daredevil	619

Figure 3.1.1, Degree Centrality results, top 15

3.2 Harmonic Centrality

Harmonic Centrality measures a node's inverse distance to all other nodes. Nodes with a high centrality score have the shortest distance from all the other nodes. I will use the normalized score which is calculated as follows:

$$H_{norm}(u) = \frac{\sum_{v=1}^{n-1} \frac{1}{d(u,v)}}{n-1}$$

Where u is the node for which we are calculating centrality, $d(u,v)$ is the distance between node u and node v (any node except node u) and n is the number of nodes in the graph.

In the context of the Marvel Universe, heroes with high centrality scores can transmit information across the universe the fastest.

name	centrality
Spider-Man	0.38006
Captain America	0.37345
Iron Man	0.36146
Thing	0.35371
Thor	0.35103
Human Torch	0.34998
Mr. Fantastic	0.34918
Wolverine	0.34638
Invisible Woman	0.34361
Hulk	0.34333
Scarlet Witch	0.34007
Beast	0.33966
Vision	0.33707
Dr. Strange	0.33493
Cyclops	0.33482

Figure 3.2.1, Harmonic Centrality results, top 15

Note that before Harmonic Centrality I tried using Closeness but found that some of the results were unreliable due to disconnectedness. For example, the hero ‘Blare’ had a 1:1 relationship with ‘MTU2 1’. Harmonic Centrality deals with disconnected components more cleanly. (i.e., $if\ d(u,v) -> \infty, then\ 1/d(u,v) -> 0$)

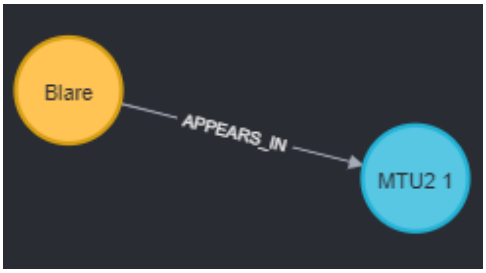


Figure 4.1. Disconnected graph component

3.3 Betweenness Centrality

Betweenness Centrality measures the number of shortest paths a node appears on. Nodes with a high centrality score tend to appear on more shortest paths. The calculation is as follows:

$$B(u) = \sum_{s \neq u \neq t} \frac{p(u)}{p}$$

Where u is the node whose centrality we are calculating, $p(u)$ is the number of shortest paths between nodes s and t that pass through node u, and p is the total number of shortest paths between nodes s and t.

In the context of my graph, betweenness centrality is the influence of a Hero on the spreading of information across the marvel universe or the functioning of groups/systems

within the universe. For example, the systems making up the Spiderman comic series would be affected severely by the absence of Spiderman.

name	centrality
Spider-Man	27258886.24
Captain America	17542921.13
Iron Man	12529323.88
Hulk	10509194.14
Thor	9804819.923
Wolverine	9406839.982
Dr. Strange	9151501.484
Daredevil	8146500.621
Thing	7318849.455
Human Torch	5393891.301
Mr. Fantastic	5033986.001
Beast	4998792.274
Sub-Mariner	4997456.381
Hawk	4534529.087
Fury, Col. Nicholas	4520229.675

Figure 3.3.1. Betweenness Centrality results, top 15

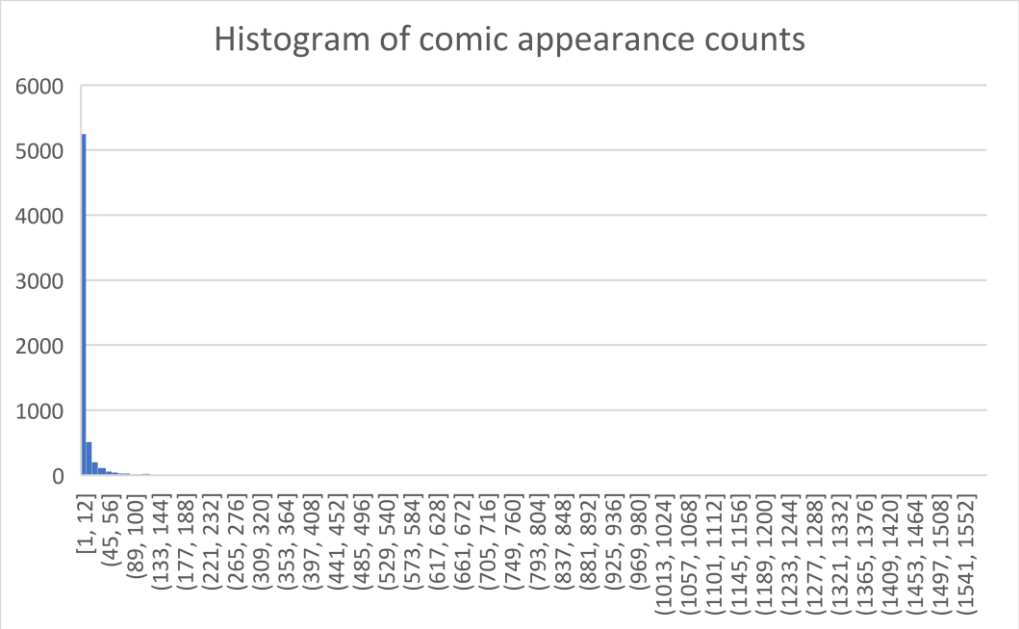
4 Analysis

4.1 Degree Centrality

As mentioned earlier, degree centrality in the context of my graph is simply the number of comics a hero appears in. The top 3 heroes being ‘Spider-man’, ‘Captain America’ and ‘Iron Man‘ would be expected by a subscriber Marvel, whether Movie-only or a Comic reader.

From this, I observed that more popular characters in the comics tend to appear in digital media proceedings (movies, shows). Although, some popular comic characters don’t have their respective shows or movies yet (e.g., Daredevil) while some less frequently appearing characters do (e.g., Loki). I also observed that Marvel has a very small proportion of primate characters that appear in the vast majority of comics. This results in the sharp right skew in the histogram below. With 6432 heroes in total, 5250 appear in 12 comics or less.

Figure 4.1.1. Histogram of Degree Centrality values



4.2 Harmonic Centrality

Harmonic Centrality in my graph's context refers to a character's ability to transmit information across the Marvel universe. The ranking order here is similar to that of Degree Centrality, with most heroes in the top 15 being common to both. This is intuitive, as characters showing up in the most comics will have the shortest paths to most other heroes.

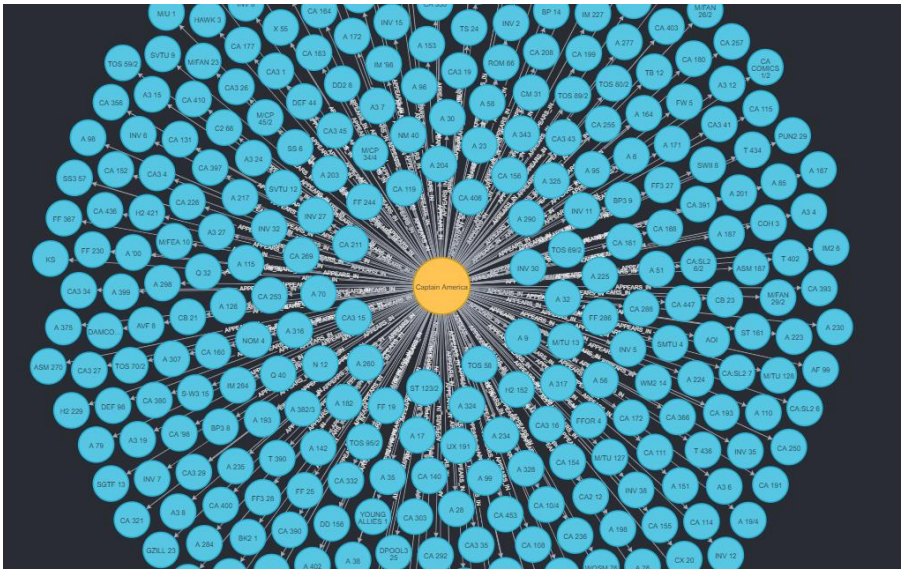


Figure 4.2.1. Captain America and 300 comics he has appeared in

The values of harmonic centrality are distributed unimodally with a lesser right skew than that of Degree centrality. A small number of outlier nodes show very low centrality, these are the disconnected components of the graph. (E.g., 'Blare')

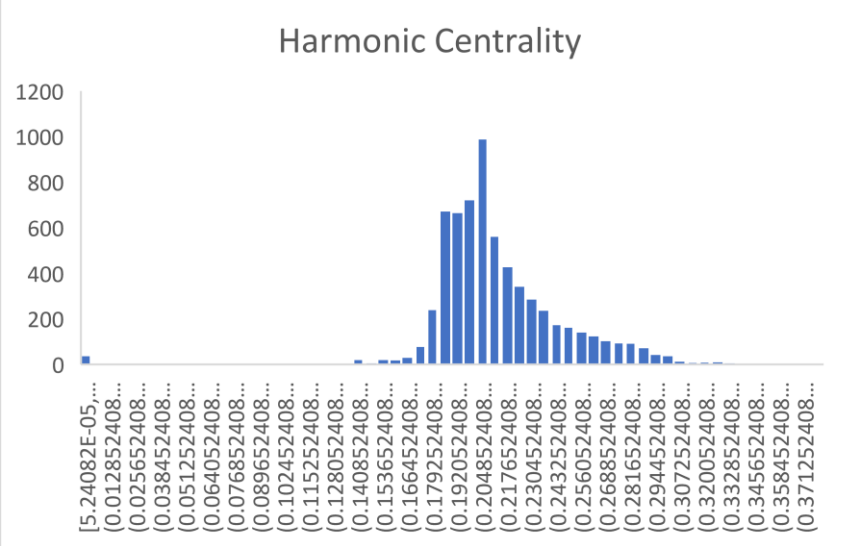
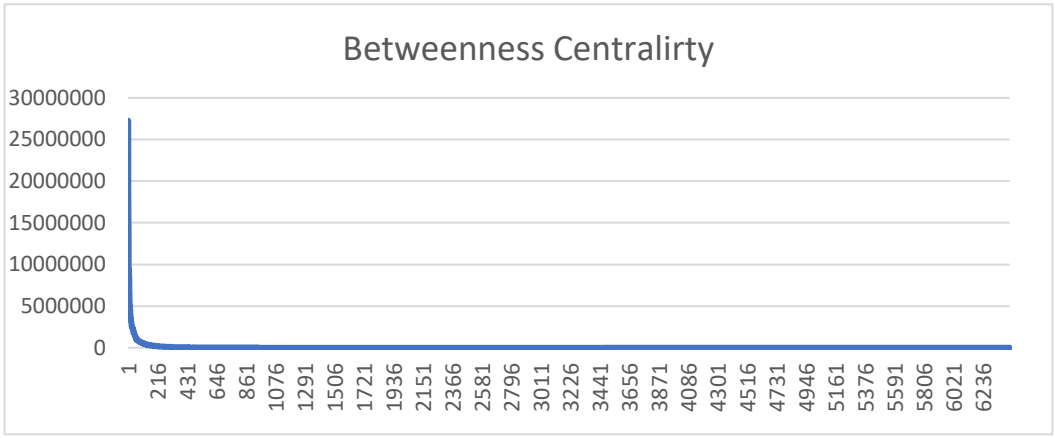


Figure 4.2.1. Histogram of Harmonic Centrality values

4.3 Betweenness Centrality

In the context of my graph, Betweenness Centrality is the influence of a Hero on the spreading of information across the marvel universe or the functioning of groups/systems within the universe. Heroes with high betweenness centrality tend to appear on most of the shortest paths between other nodes, meaning that the structure of the entire Marvel Universe is dependent on these heroes. There's a similar trend with the sharp decline in scores like that of Degree Centrality, implying that the structure of the Marvel Universe relies on a small proportion of the characters.



Several things could happen if one of these pivotal characters were to be removed from the Marvel Universe. Firstly, comics that have these characters as the main character would not exist. In this way, many other characters who appear in their comic series would also not exist. For example, without Spiderman, Mary-Jane's character would be incomplete and probably wouldn't exist. The same goes for Green Goblin, Venom, Vulture, Sandman, and many more.

5 Conclusion

In this research report, I carried out graph analytics on the Marvel Universe network graph using several Centrality algorithms. I investigated the influence on the system which individual characters have, and the distribution of influence across the entire system. By doing this, I was able to extract some insights into the structure of the Marvel Universe. Although this research was based on a fictional system, the approaches I used would open ample opportunity for valuable graph analytics on a similar real system.

References

[1]

C. Sanhueza, “The Marvel Universe Social Network,” *Kaggle.com*, 2017. https://www.kaggle.com/csanhueza/the-marvel-universe-social-network?select=nodes.csv (accessed Nov. 22, 2021).