Investigating the MARVEL Comic Universe

By Danai Avgerinou , Holly Capell, Shannon McNish, Taylor Pellerin, Kaya Tollas

Motivation

Over the past 10 or so years, Marvel Studios has been churning out more and more installments of *The Avengers*, and related stories like the wildly popular *Black Panther*. With the recent acquisition of *Spider-Man* and *X-Men*, Disney (owner of Marvel Studios) now has the rights to almost all of the relevant characters created by Stan Lee, opening up exciting possibilities for heros to make guest appearances and go on collaborative adventures. The film adaptations that make up the Marvel Cinematic Universe often rely heavily upon the original print material, and so understanding who the most popular characters have been across time could be indicative of who might appear in the soon to be released *Avengers: Infinity War*.

In order to understand the social structure of the Marvel Comic Book Universe, we used a Kaggle-curated data set, *The Marvel Universe Social Network*, provided by csanhueza. The segment of this data that we are interested in consisted of an edge list between characters who have appeared in comics together, in the form of a csv file. With this, we are quickly able to construct a graph object and investigate regions of high density, particularly in the neighborhoods of key players from the Marvel Cinematic Universe (MCU), such as Thor, Captain America, IronMan, Black Widow, and SpiderMan.

For this analysis, we would like to investigate which heroes are most important in the Marvel universe by quantifying centrality and connectivity metrics. We are also interested in identifying and investigating communities within the Marvel universe. Specifically, we are interested in investigating three questions:

- Are there were heroes that are not highly connected themselves, but are important in serving as a bridge, creating a path between other heros and their corners of the universe?
- Are the communities that we expect to see, such as X-Men or Avengers, densely connected as to form detectable communities in the data?
- How does one of the most famous Marvel characters, Captain America, and his Avengers crew interact with the rest of the characters?

Methods

Before we could dive into any analysis, we first had to process the data. The data from Kaggle consisted of an unweighted edge list, where each edge represents whether or not characters existed in the same comic book (each edge represents a different comic book issue). This edge list contained several duplicates, since many characters had multiple adventures together and individual characters may appear across many issues and series of comic books. To generate edge weights, we deduplicated this edge list, counting the number of times that each team appeared and setting the edge weight equal to this value. We also created node weights for each hero equal to the number of times that each character appeared. The network we created

is undirected since each edge represents simply the presence of two heroes in a comic and there is no directed interaction between two connected heros.

After preprocessing the data, we selected appropriate methods for each of our previously stated questions. The following sections detail the methods we used to address each question.

I. Investigating Degree and Centrality Measures

We were able to use the edge and node weights as a baseline metric for prolificity of a hero in the Marvel Comic Universe, as this showed who appeared frequently and who teamed up the most often. However, just because a hero has appeared in comic books with many other characters doesn't necessarily mean that those connections were the most influential. To dig deeper into this and gain a better understanding of the network as a whole, we investigated using other measures of centrality: betweenness and eigenvector centrality. Betweenness centrality quantifies the number of times a hero connects two other heros along the shortest path between them. Eigenvector centrality is a measure of the influence of a hero in the Marvel network. A high eigenvector score means that a hero is connected to many heros who themselves have high scores.

II. Investigating Communities in the Marvel Universe

We next investigated the modularity of the network, in order to detect communities and see if teams such as the Avengers, Fantastic Four, and X-Men could get detected by our analysis. We also wanted to identify as any other communities that may be more prevalent in the comics but did not make it to film. We used community detection on the subset of top 1,000 characters in number of connected edges.

Community detection methods identify densely connected subgraphs of the network . The idea is to maximize modularity by splitting the network into different communities so each community has a high density of edges. Since obtaining exact modularity is NP-hard, we used Louvain modularity for the approximation. This algorithm was used because it achieves similar results as other community detection algorithms in less time.

III. Investigating Ego Networks: Captain America and His Friends

With the communities detected above in mind, we took a deeper dive into the Avengers community, seeing how they interact with each other as well as how the leader, Captain America interacts with the rest of his social network.

For this analysis, we used an ego network to visualize Captain America and the heroes he directly connects to. Ego networks reveal exactly who a particular node interacts with. We also used closeness centrality on this network as a measure of the degree to which each hero is near

all other individuals in the network. The higher the degree of closeness centrality for a node, the closer it is to all other nodes.

Results

I. Investigating Degree and Centrality Measures

We first turned to the edge and node weights for an idea of the importance of characters to the universe, investigating who were the heaviest edges and nodes. First, we can look at Figure 1 for an idea of just how well connected the top 100 most commonly appearing characters were. Tables 1 & 2 contain the top 20 characters as well as the 20 most common co-appearances.

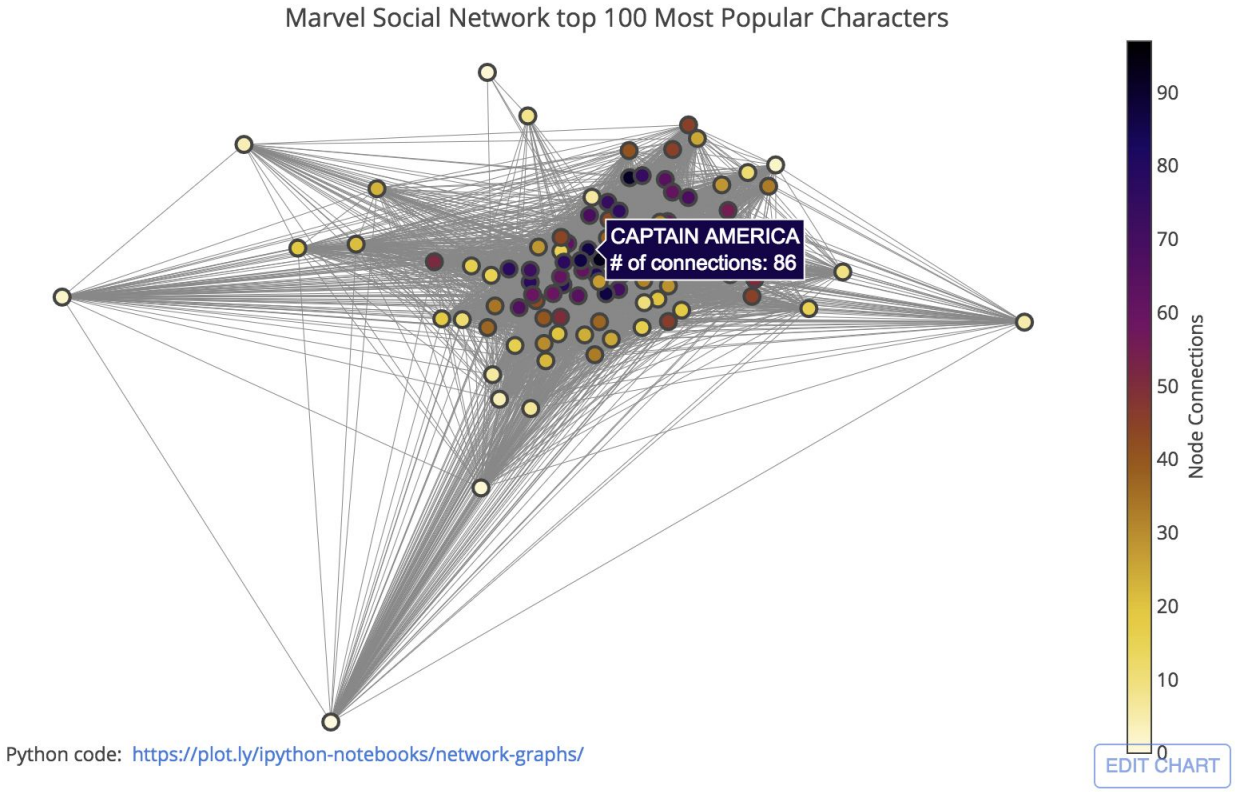


Figure 1: Static plot of the interconnectedness of the top 100 most popular Marvel characters, by number of appearances. Interactive version located here (https://plot.ly/~tjpell/143).

Rank	Hero	Degree		Rank	Hero	Degree
1	CAPTAIN AMERICA	16499		**11**	INVISIBLE WOMAN/SUE	9326
2	SPIDER-MAN/PETER PAR	13717		**12**	BEAST/HENRY &HANK& P	9287

3	IRON MAN/TONY STARK	11817		**13**	CYCLOPS/SCOTT SUMMER	9099
4	THOR/DR. DONALD BLAK	11427		**14**	STORM/ORORO MUNROE S	8795
5	THING/BENJAMIN J. GR	10681		**15**	HAWK	8483
6	WOLVERINE/LOGAN	10353		**16**	WASP/JANET VAN DYNE	8426
7	HUMAN TORCH/JOHNNY S	10237		**17**	COLOSSUS II/PETER RA	7863
8	SCARLET WITCH/WANDA	9911		**18**	PROFESSOR X/CHARLES	7840
9	MR. FANTASTIC/REED R	9775		**19**	HULK/DR. ROBERT BRUC	7515
10	VISION	9696		**20**	ANT-MAN/DR. HENRY J.	7343

Table 1: The 20 most popular Marvel characters, by number of appearances.

Rank	Hero 1	Hero 2	Degree	Rank	Hero 1	Hero 2	Degree
1	MISS AMERICA / MADELIN	PATRIOT / JEFF MACE	1894	**11**	SCARLET WITCH / WANDA	VISION	422
2	HUMAN TORCH / JOHNNY S	THING / BENJAMIN J. GR	744	**12**	ANT-MAN / DR. HENRY J.	WASP / JANET VAN DYNE	406
3	HUMAN TORCH / JOHNNY S	MR. FANTASTIC /REED R	713	**13**	CYCLOPS / SCOTT SUMMER	MARVEL GIRL / JEAN GRE	390
4	MR. FANTASTIC / REED R	THING / BENJAMIN J. GR	708	**14**	STORM / ORORO MUNROE S	WOLVERINE / LOGAN	389
5	INVISIBLE WOMAN / SUE	MR. FANTASTIC / REED R	701	**15**	CAPTAIN AMERICA	THOR / DR. DONALD BLAK	386
6	HUMAN TORCH / JOHNNY S	INVISIBLE WOMAN / SUE	694	**16**	CAPTAIN AMERICA	VISION	385
7	INVISIBLE WOMAN / SUE	THING / BENJAMIN J. GR	668	**17**	CAPTAIN AMERICA	WASP / JANET VAN DYNE	384

8	SPIDER-MAN / PETER PAR	WATSON-PARKER MARY	616	**18**	PARKER MAY	SPIDER-MAN / PETER PAR	380
9	JAMESON J. JONAH	SPIDER-MAN / PETER PAR	526	**19**	CAPTAIN AMERICA	SCARLET WITCH / WANDA	374
10	CAPTAIN AMERICA	IRON MAN / TONY STARK	446	**20**	IRON MAN / TONY STARK	SCARLET WITCH / WANDA	372

Table 2: The 20 most popular Marvel pair-wise teams, by number of appearances.

Looking at the above tables for the 20 heaviest weighted edges (Table 2) and nodes (Table 1), we notice that while Captain America is by-and-large the most commonly appearing hero, he is part of the 10th most common tag team, which is shared with Iron Man, the 3rd most popular hero by appearances. Interestingly enough, Miss America and Patriot are not in the top 15 heroes, even though they make up the most popular pair by far. This indicates that when they appear in comic books, it is most often together.

We next investigated the betweenness of our social network, to gain better insight into the key players in the universe and identify the one-hit-wonders. Table 3 includes this measure of centrality for the top 20 ranked characters, and Figure 2 contains a screenshot of an interactive plot of the top 100 characters by betweenness and their connectivity. Interactivity can be found here.

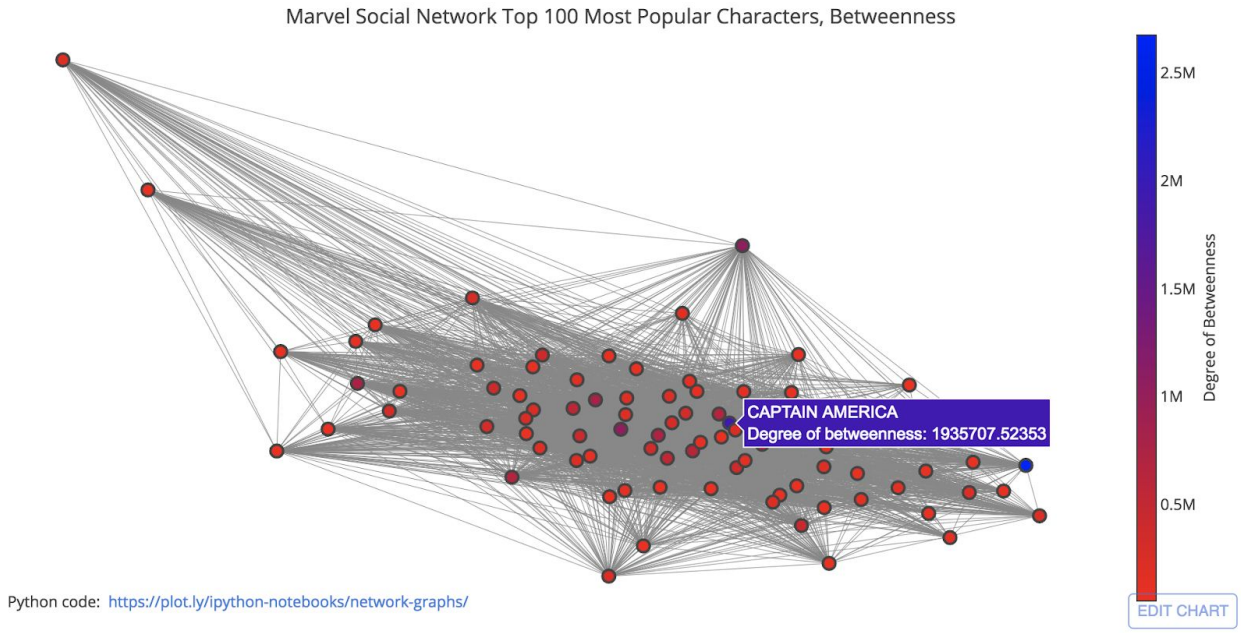


Figure 2: Static plot of the interconnectedness of the top 100 most popular Marvel characters, by measure of betweenness. Interactive version located here (https://plot.ly/~tjpell/155).

Rank	Hero	Betweenness		Rank	Hero	Betweenness
1	SPIDER-MAN/PETER PAR	2676603.3		**11**	BEAST/HENRY HANK	660047.4
2	CAPTAIN AMERICA	1935707.5		**12**	HAWK	623125.4
3	IRON MAN/TONY STARK	1067304.5		**13**	MR. FANTASTIC/REED R	565053.9
4	WOLVERINE/LOGAN	1062680.3		**14**	HUMAN TORCH/JOHNNY S	539222
5	DR. STRANGE/STEPHEN	846621.3		**15**	FURY, COL. NICHOLAS	492337.6
6	HAVOK/ALEX SUMMERS	806363.7		**16**	SILVER SURFER/NORRIN	439868.2
7	THOR/DR. DONALD BLAK	770925.3		**17**	SCARLET WITCH/WANDA	415569.1
8	HULK/DR. ROBERT BRUC	743858		**18**	SUB-MARINER/NAMOR MA	396431.1
9	THING/BENJAMIN J. GR	681995.6		**19**	PUNISHER II/FRANK CA	395655.4
10	DAREDEVIL/MATT MURDO	677435.4		**20**	SHE-HULK/JENNIFER WA	388712
Highlighted rows represent heros in the top 20 betweenness scores but not in the top 20 by degree						

Table 3: The 20 most popular Marvel characters, by Betweenness Centrality.

Looking at the above table, there are seven heroes (highlighted in yellow in Table 3) that are in the top 20 most central heroes by betweenness but were not ranked in the top 20 heroes with highest degrees. For these seven characters, the connections that they have are important in linking the network together, however they may not have as many connections as other heroes. For example, Punisher II has the 19th highest betweenness centrality but doesn't even rank in the top 100 in terms of degree.

We next investigated eigenvector centrality as a measure of closeness and importance in this network. Figure 3 contains a screenshot of an interactive plot of the top 100 characters by eigen-centrality and their connectivity. The The top 20 most popular heroes in terms of eigenvector centrality are presented in Table 4 below.

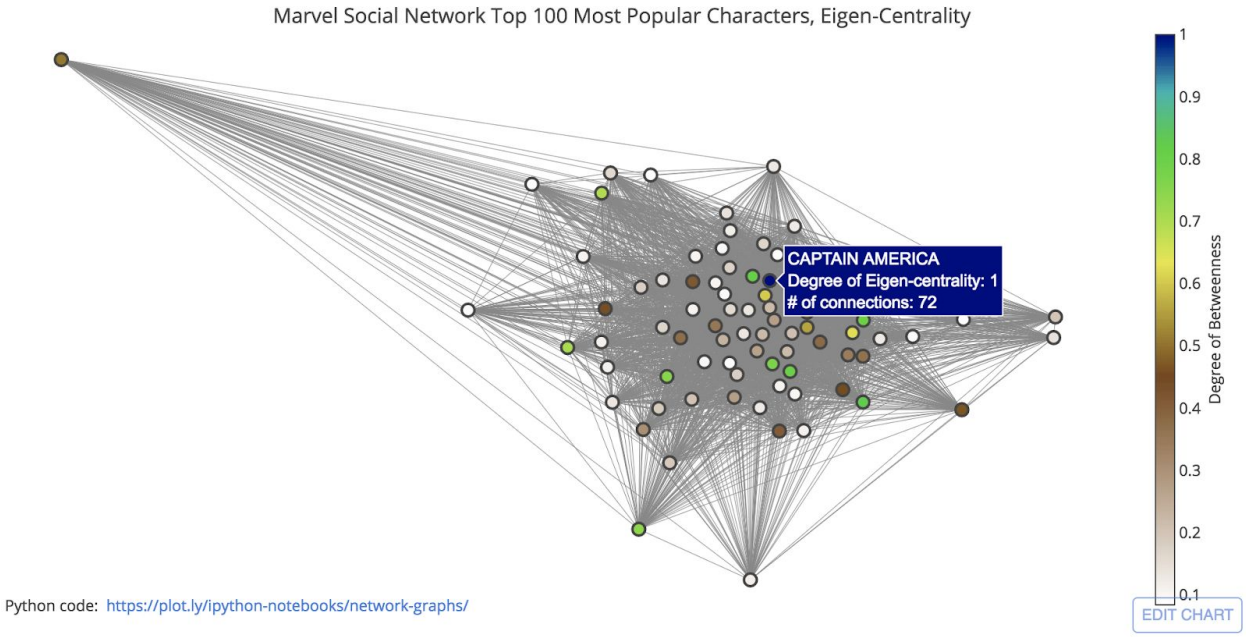


Figure 3: Static plot of the interconnectedness of the top 100 most popular Marvel characters, by measure of eigencentrality. Interactive version located here (https://plot.ly/~tjpell/163).

Rank	Hero	Eigencentrality		Rank	Hero	Eigencentrality
1	CAPTAIN AMERICA	1		**11**	HAWK	0.6255454
2	THING/BENJAMIN J. GR	0.8266349		**12**	ANT-MAN/DR. HENRY J.	0.6047572
3	HUMAN TORCH/JOHNNY S	0.8121594		**13**	BEAST/HENRY &HANK& P	0.5575718
4	MR. FANTASTIC/REED R	0.7980751		**14**	WONDER MAN/SIMON WIL	0.5073018
5	IRON MAN/TONY STARK	0.7951503		**15**	CYCLOPS/SCOTT SUMMER	0.4824879
6	INVISIBLE WOMAN/SUE	0.7681152		**16**	WOLVERINE/LOGAN	0.4673382
7	SCARLET WITCH/WANDA	0.7519413		**17**	SPIDER-MAN/PETER PAR	0.4576505
8	VISION	0.7407603		**18**	STORM/ORORO MUNROE S	0.4360988
9	THOR/DR. DONALD BLAK	0.7051696		**19**	PROFESSOR X/CHARLES	0.4142344
10	WASP/JANET VAN DYNE	0.6964646		**20**	SHE-HULK/JENNIFER WA	0.4059961
Highlighted rows represent heros in the top 20 eigenvector scores but not in the top 20 nodes by number of degrees						

Table 4: The 20 most popular Marvel characters, by Eigenentrality.

There are two heroes (Wonder Man and She-Hulk) that appear on the list of top 20 most eigencentric heroes (as shown in Table 4) but were not in the top 20 heroes ranked by degree. These heroes are considered highly influential by eigenvector centrality (despite not being in the top 20 in terms of degree) because the heroes they do connect to are also influential. It is interesting to note that though these characters appear to be highly influential in the print comics universe (as illustrated by their eigenvector centrality), however , Wonder Man and She-Hulk have not yet made it to the cinematic universe.

II. Investigating Communities in the Marvel Universe

After performing community detection as described in the Methods section, our optimal number of communities for the Marvel network was six. The communities are visualized in Figure 4 below and examples of heroes in each community are provided in Table 5.

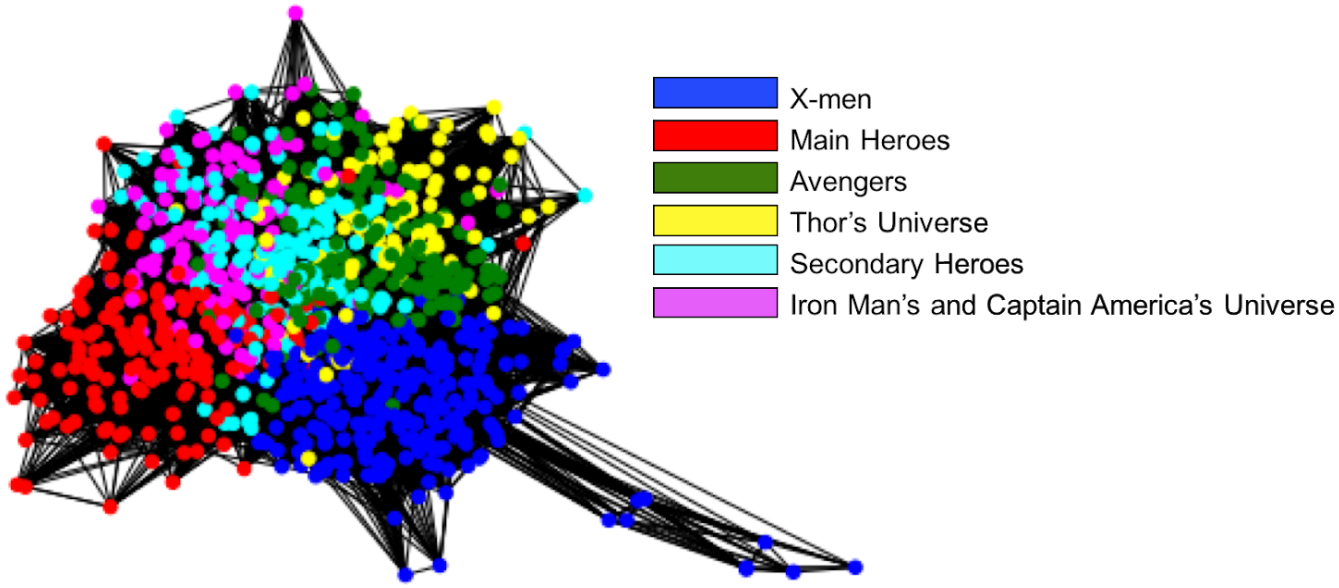


Figure 4: Visualization of communities in the Marvel universe

Hero Community	Team Members
X-Men	Wolverine, Storm, Cyclops, Professor X, Deadpool, Rogue, Beast
Main Heroes	Daredevil, Spiderman, Mr. Fantastic
Avengers	Scarlet Witch, Black Widow, Hulk, She-Hulk, Antman, Vision
Thor's Universe	Thor, Zeus, Loki, Odin, Pluto, Hela, Hercules

Secondary Heroes	Tigra, Cabe Bethany, Tyger Lord, Erwin Morley
Iron Man and Captain America's Universe	Iron Man, Captain America, Patriot

Table 5: Examples of heroes in each community

Before performing this analysis, we expected to observe X-Men and Avengers as two mostly disparate communities. We confirmed this hypothesis for X-Men, however the Avengers were split between a few communities. Notably, Iron Man and Captain America are in their own community separate from many of the other Avengers (Black Widow, Scarlet Witch, Hulk, etc.). Similarly, Spiderman is in yet another community, with other leading heroes such as Daredevil and Mr. Fantastic.

Another interesting finding was that not all Fantastic 4 members appear in the same community. The Human Torch and Mr Fantastic are in the Main Heroes community, while the Thing and Invisible Woman are part of the Avengers community. This was surprising, but may be a result of the fact that the Fantastic 4 are allies of the Avengers.

We found that the Avengers and Thor's Universe communities have a lot of overlap. This makes sense because Thor is an Avenger and frequently appears with other Avengers in comics. Furthermore, many of the villains that the Avengers fight are from the Thor Universe.

It is also interesting to note that Secondary Heroes appear to be centralized in the network. We hypothesize that this is because these characters appear in a lot of different comics with heroes from other communities, and perhaps their individual comics are not as prolific as those of heroes in the tighter communities (e.g., X-Men) .

III. Investigating Ego Networks: Captain America and His Friends

Next we investigated the network structures of Captain America and his friends. First, we created a network of just Avengers to assess their level of interconnectivity. Node size represents degree and edge weights are the number of co-occurrences heroes share. As shown in the figure, Captain America, She-Hulk, Black Widow, and Dr. Strange appear on the outer edges of the network while other Avengers are clustered in the center. It appears that while Captain America and Dr. Strange are both strongly connected to most Avengers, they are not strongly connected to one another. Likewise for Black Widow and She-Hulk.

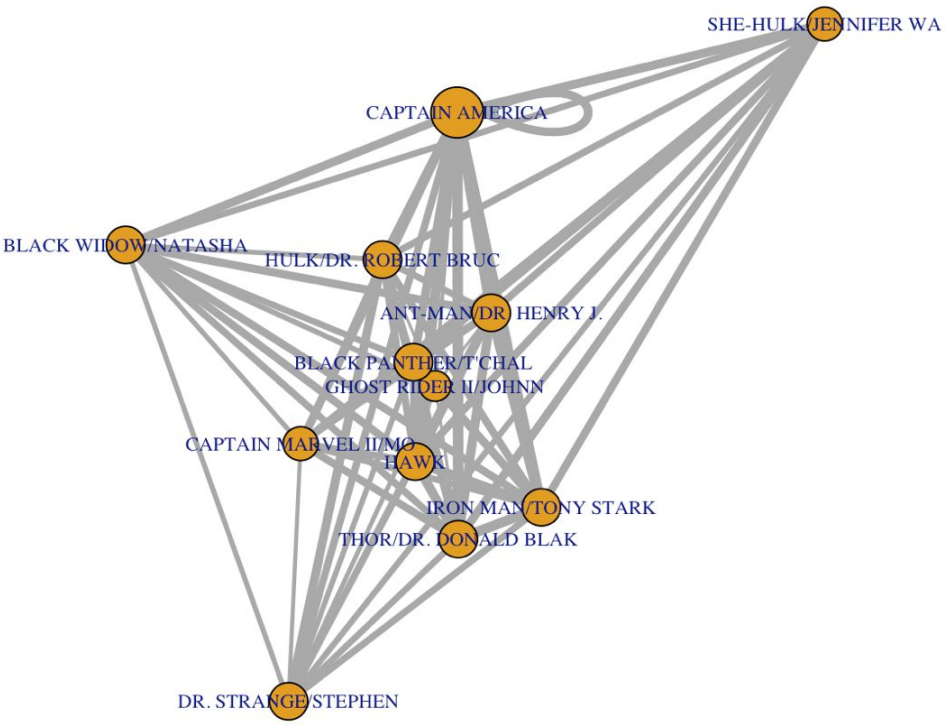


Figure 5: The Avengers Network

We also investigated the betweenness centrality of the Avengers network (results in the table below). We were surprised to see that the most between-hero in the (selected) Avengers is actually Ghost Rider!

Avenger	**Betweenness Centrality**
Ghost Rider	45.6666667
Black Panther	15.0000000
The Hulk	3.3333333
Captain Marvel	0.3333333
Iron Man	0.0000000
Black Widow	0.0000000
Captain America	0.0000000
Dr. Strange	0.0000000
Hawkeye	0.0000000

Thor	0.0000000
She-Hulk	0.0000000

Table 6: Avengers and their betweenness centrality

We also induced Captain America's ego network from the entire Marvel network, only including popular heroes (who were in the top 20th percentile of node degree). We were interested in seeing whether Avengers would show up in this network and how important or central they would be.

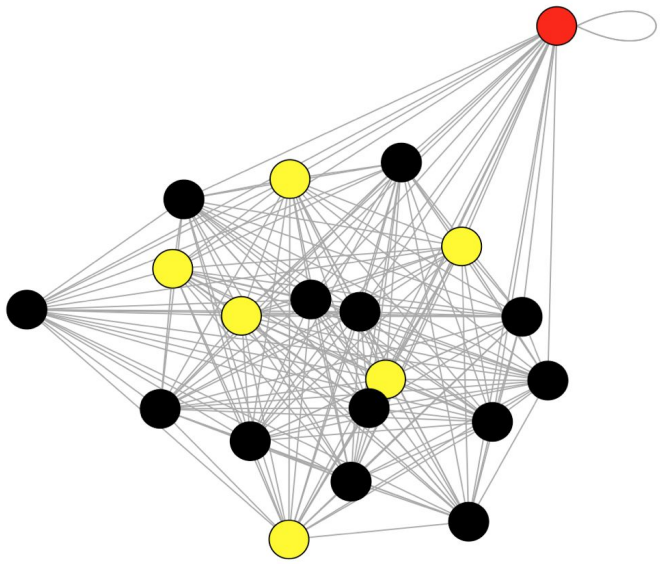


Figure 6: Captain America (red) and his ego network (Avengers in yellow)

We also investigated the closeness centrality of this network, finding that some Avengers were the closest (e.g. She-Hulk), but some were farthest (e.g. Black Widow).

Hero	**Closeness centrality**
Captain America	0.0004549591
Vision	0.0006333122
Scarlet Witch	0.0006793478
She-Hulk	0.0006930007

Table 7: Closest nodes in the Captain America ego network

Hero	Closeness centrality
Wolverine	0.0012594458
Jarvis Edwin	0.0011286682
Black Widow	0.0010706638
Hercules	0.0009560229

Table 8: Farthest nodes in the Captain America ego network

Conclusion

We uncovered several interesting findings during our analyses. We found that some heroes ranked high in terms of number of connections and importance of connections whereas other heroes had important connections but may not have as many connections.

We also found that there is some community structure in the Marvel Universe. Some of these communities were expected (i.e., members of the X-men formed a distinct community), whereas other results were surprising (i.e., some heroes who commonly appear together in comic books were found in different communities). Also in our investigation into Captain America's networks, we found that Avengers are not necessarily the most popular of Captain America's friends.

We also learned a lot about our data in the course of our analyses. The dataset that we selected did not provide as much opportunity for analysis as we initially thought. Since our network was undirected, we were unable to calculate certain measures of centrality such as in-degree and out-degree centrality. We were also unable to quantify other descriptive measures such as reciprocity. Additionally, the dataset we used contained limited features that could be used to distinguish between nodes (i.e., features that could be used for color and or/size), making more complex analyses challenging.

Network Analysis Question 2

Danai Avgerinou, Holly Capell, Shannon McNish, Taylor Pellerin, Kaya Tollas

(a) Here, the nodes represent superheroes and edges represent them coexisting in a comic book together.

(b) There are 167,100 weighted edges and 6,421 nodes in this data set

(c) Yes, the node weights represent the number of comic books that each character appeared in. The edge weights represent the number of times that each character appeared in a comic book together.

(d) This network is undirected, since sharing space in a comic book is mutual.

(e) Using a sparse representation, we only need to store the edges and their weights, and which heroes are involved. Each edge in this edge list takes 2 heros and one weight. There are 167,100 edges and so we only need to store 501,300 observations.

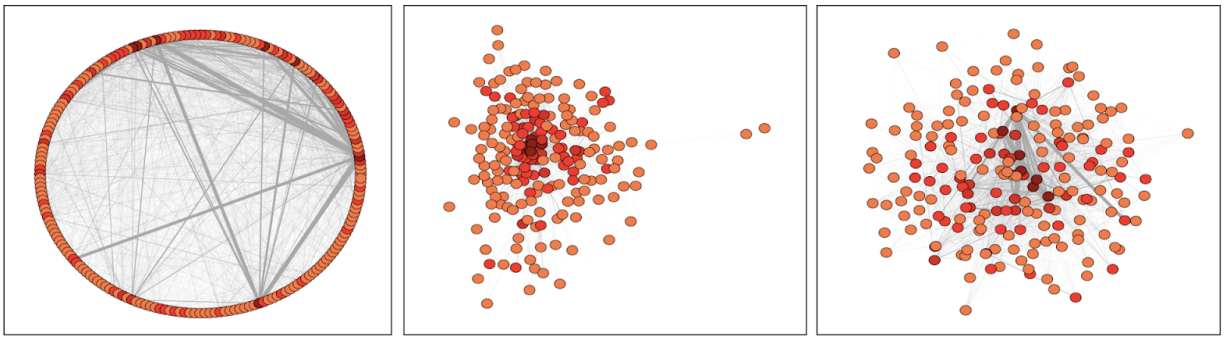
In a weighted adjacency matrix, we would have to store (1/2)(n^2) - n weights where n = 6,206, the number of heros. This is because we have an undirected graph, so we only need the upper triangular quadrant of the matrix. We can also ignore the main diagonal, since we do not care if a character has self references. This comes out to 20,640,312 data points. If we also store the node weights here, we can put these on the “main diagonal” of the adjacency matrix, leaving us with a total of (1/2)(n^2), or 20,646,738 entries to keep track of.

(f) We would like to investigate which heroes are most important in the Marvel universe by quantifying centrality and connectivity metrics. We are also interested in identifying and investigating communities within the Marvel universe. Specifically, we would like to answer the following questions:

- Are there were heroes that are not highly connected themselves, but are important in serving as a bridge, creating a path between other heros and their corners of the universe?
- Are the communities that we expect to see, such as X-Men or Avengers, densely connected as to form detectable communities in the data?
- How does one of the most famous Marvel characters, Captain America, interact with the rest of the characters?

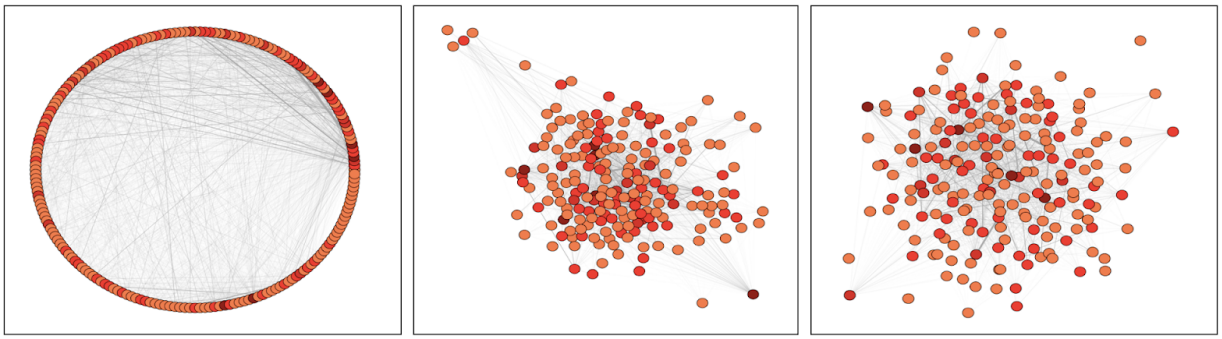
(g) The three graphs below are created based on three random samples of 200 nodes. The coloring scale is based on the number of connections a node has: the darker the color, the more connections it has. Another interesting property is that in the Kamad-Kawai and Fructerman Reingold layout, the nodes with the most connections are centralized and the rest are spread around them.

Subgraph 1



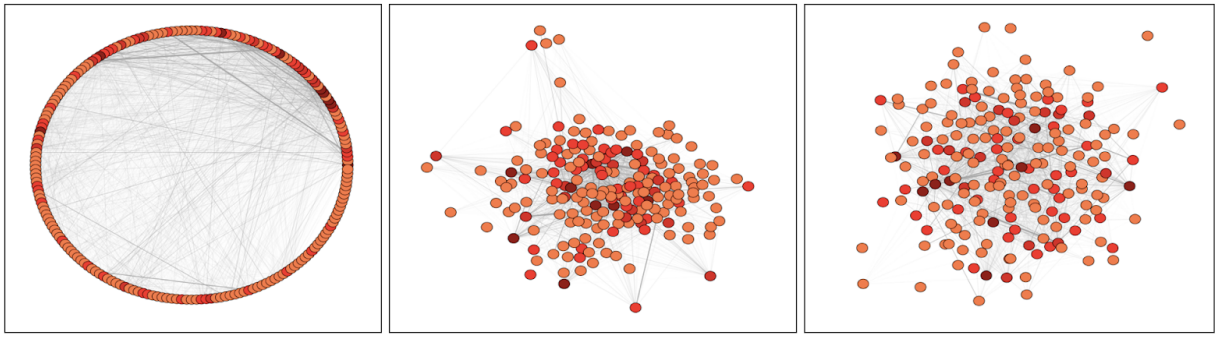
Circle | Fructerman Reingold | Kamada-Kawai

Subgraph 2



Circle | Fructerman Reingold | Kamada-Kawai

Subgraph 3



Circle | Fructerman Reingold | Kamada-Kawai

(h) See markdown report below