Unmasking Marvel Superheroes: A Social Network Analysis of Character Relationships and Importance

1 Introduction

Marvel Comics, founded in 1939 as Timely Publications and later adopting the name Marvel Comics in the 1960s, is a leading global comic book publisher with a profound impact on popular culture. It's renowned for creating iconic superheroes like Spider-Man, Iron Man, Captain America, Thor, Hulk, Black Widow, and the X-Men, transcending the comic book medium to become cultural icons. In this Social Network Analysis (SNA) project, we explore the Marvel Universe, a vast interconnected superhero realm. Marvel's most significant contribution is the creation of the Marvel Universe, a shared fictional world where characters from various series interact, enriching the storytelling experience. The network under examination adopts the structural form of a graph representation within the expansive Marvel superhero universe. Within this graph, superheroes are symbolized as nodes, and the interconnections among them, established through their co-appearances in comics, are delineated as weighted edges. The assignment of edge weights is predicated upon the frequency of co-appearances, thus affording a quantitative assessment of the strength inherent in the relationships between superheroes. Some measurements (e.g. centrality, community detection approaches, cliques detection, robustness testing) are achieved in order to determine the most important character, the most important community, the groups, and the robustness of the network.

2 Problem and Motivation

The Marvel Cinematic Universe (Marvel Universe) boasts an extensive ensemble of iconic superheroes, interconnected storylines, and a rich narrative tapestry. The objective of this project is to employ Social Network Analysis (SNA) to delve into the intricate web of relationships among Marvel superheroes, thereby identifying the most significant characters, exploring character communities, and assessing the network's robustness. Indeed, the Marvel Universe features numerous superheroes, and understanding their interactions and relationships is essential to comprehend the underlying dynamics of the World presented. In addition, investigating the formation of character communities within the Universe enables us to identify tightly-knit groups of superheroes who frequently collaborate or appear together. This analysis aims to explore the alliances, friendships, and partnerships that shape the superhero landscape. Understanding character communities provides valuable insights into superhero team-ups. Finally, the last purpose of this project is about robustness. Indeed, assessing the robustness of the superhero network involves subjecting it to various disruption scenarios. By removing characters based on different criteria, we aim to understand how vulnerable the network is to character removals. The robustness analysis enables us to determine the critical superheroes whose absence might have the most significant impact on the interconnected Marvel superhero network.

3 Datasets

The dataset employed in this Social Network Analysis (SNA) project represents a meticulously sourced and thoughtfully curated collection of data centered around Marvel superheroes and their intricate interconnected relationships within the expansive Marvel Universe. It is important to provide a comprehensive overview of this dataset, as it forms the foundation upon which the entire analysis is built. The dataset was obtained from Kaggle, a renowned platform for data science and machine learning resources, known for its commitment to data quality and reliability.
Comprising three distinct files, this dataset has been carefully structured to facilitate a detailed exploration of Marvel superheroes' social network dynamics:

1. **nodes.csv**: The first file, named "nodes.csv," serves as the fundamental building block of the dataset. It is within this file that essential information about the nodes in the network, encompassing both superheroes and comic titles, is stored. The "node" column of this file lists the names of the various entities, offering a comprehensive roster that includes the Marvel Universe's iconic characters, as well as the titles of comics. The "type" column further distinguishes these nodes by categorizing them as either heroes or comics.

2. **edges.csv**: The second file, known as "edges.csv," delves into the establishment of relationships within the Marvel Universe. This file plays a pivotal role in the analysis, as it defines the connections and interactions between Marvel heroes and the comics in which they appear. Within this file, the "hero" column records the names of superheroes, serving as a reference to the roster established in the "nodes.csv" file. Simultaneously, the "comic" column specifies the exact comic titles in which these heroes make their appearances.

3. **hero-edge.csv**: The third and final file, named "hero-edge.csv," further refines the dataset's focus on character interactions within the Marvel Universe. In this file, the dataset provides in-depth information about the network of heroes who co-appear in the same comics. Each row in this file represents a relationship between two superheroes who share appearances within a specific comic.

The project's initial phase focuses on essential dataset preprocessing, a critical step involving systematic operations to refine and structure raw data for subsequent analysis.
The process commences with the amalgamation of data files representing diverse Marvel superhero universe sources into a comprehensive dataframe. This dataframe, created through Pandas library utilization, pairs each superhero with their respective appearance count. This consolidation simplifies data management and analysis.
To maintain dataset consistency and uniformity, an important data refinement step is executed. Superhero names are truncated to a maximum of 20 characters, streamlining unwieldy and excessively lengthy names.
Post truncation, a meticulous sorting operation takes place. Superheroes are systematically arranged in descending order based on their appearance frequencies. This sorting procedure establishes a hierarchy within the dataset, facilitating the identification of superheroes with the highest appearance frequencies. This level of detail ensures the dataset is thoroughly refined and prepared for comprehensive analysis and exploration.

4 Validity and Reliability

Validity in the context of analyzing Marvel superheroes revolves around the idea of how accurately and meaningfully we represent their roles, significance, and interactions within the Marvel Universe. A valid analysis should provide insights that align with our intuition and knowledge of these characters. For instance, when examining the centrality of Captain America, a valid analysis should not only capture the number of connections but also the quality and importance of these connections. It should recognize that Captain America's collaborations with characters like Iron Man or Thor carry different weight and significance compared to brief interactions with lesser-known heroes. For this reason the edges between two nodes are weighted depending on the number of appearances together in the overall comics scenario.
However, reliability is equally paramount in ensuring that our findings aren't mere artifacts of chance or specific data conditions. Reliability refers to the consistency of results over multiple measurements or analyses. In the Marvel superhero network, if Captain America consistently emerges as the main character across different subsets of data, it suggests that our analysis is reliable. This reliability signifies that the observed centrality isn't a one-time fluke but a robust characteristic of Captain America's position within the Marvel Universe.
From the thus-prepared dataset, four distinct subsets were extracted, each catering to different tiers of hero prominence. These subsets were constructed by isolating specific groups of superheroes based on their frequencies of appearances. The first subset encompassed the top 25 heroes in terms of appearance frequency, the second included the top 50 heroes, the third comprised the top 100 heroes, and the final subset incorporated the top 250 heroes.
Following the extraction of the four datasets, a critical step was the conversion of these datasets into graph structures. This transformation was carried out using the Python library NetworkX, a powerful tool for creating and analyzing complex networks. These graphs served as a fundamental framework for visualizing and quantifying the intricate relationships among superheroes within the Marvel universe.
In this graphical representation, each node within the graph corresponds to an individual superhero, effectively creating a comprehensive network of interconnected characters. To establish the connections between superheroes, edges were employed, with each edge symbolizing instances where two superheroes had coappeared in at least one comic book. Furthermore, the weight assigned to each edge was contingent on the frequency of these coappearances. In essence, the more often two superheroes were found together in comics, the heavier the weight of the corresponding edge.
To enhance the visual clarity of this graph representation, a thoughtful design choice was made. Edges with higher weights, signifying a greater frequency of coappearances, were rendered as thicker lines in the graphical depiction. This visual differentiation serves as an immediate indicator of the strength and significance of the relationships between superheroes. By highlighting the thicker edges, the graphical representation effectively draws attention to the most substantial and frequently occurring superhero pairings.

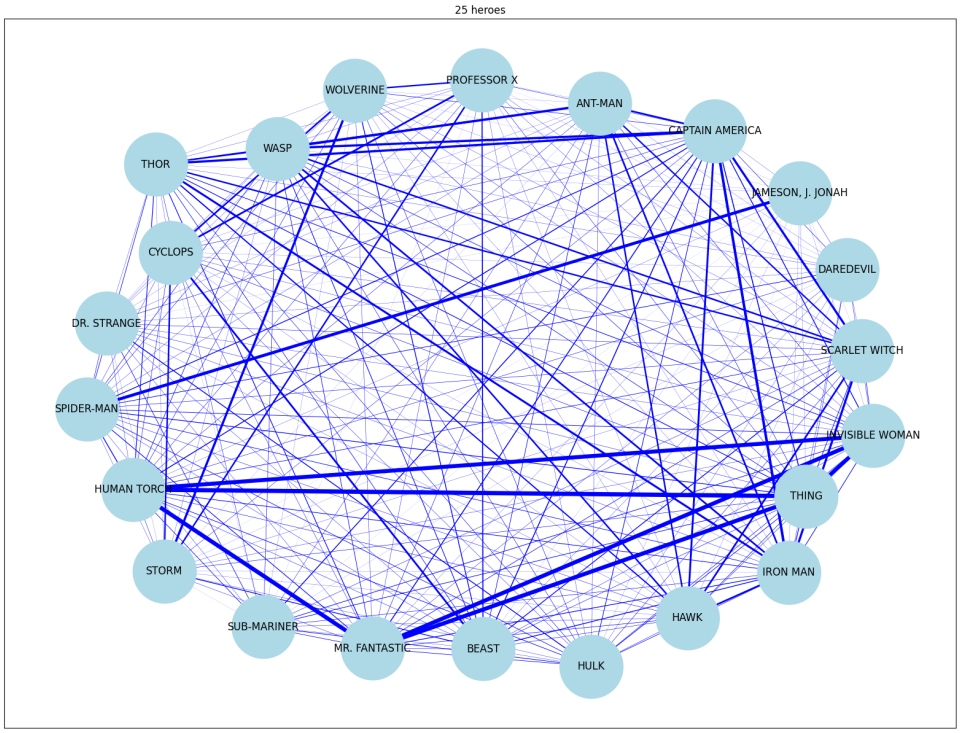

Figure 1: Graph representation of 25 Heroes

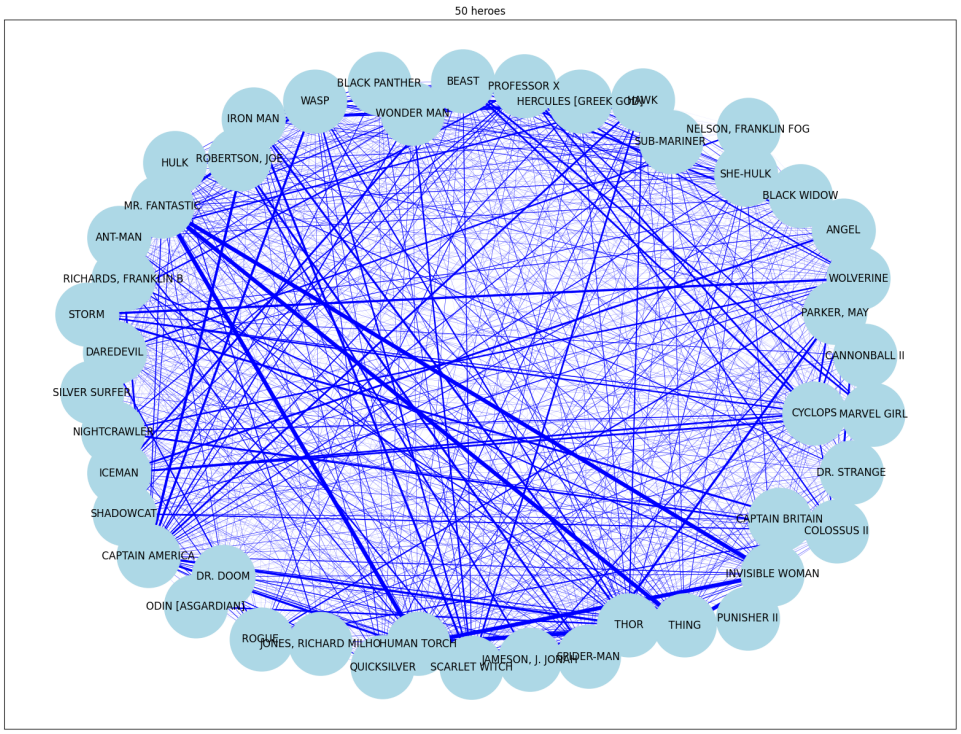

Figure 2: Graph representation of 50 Heroes

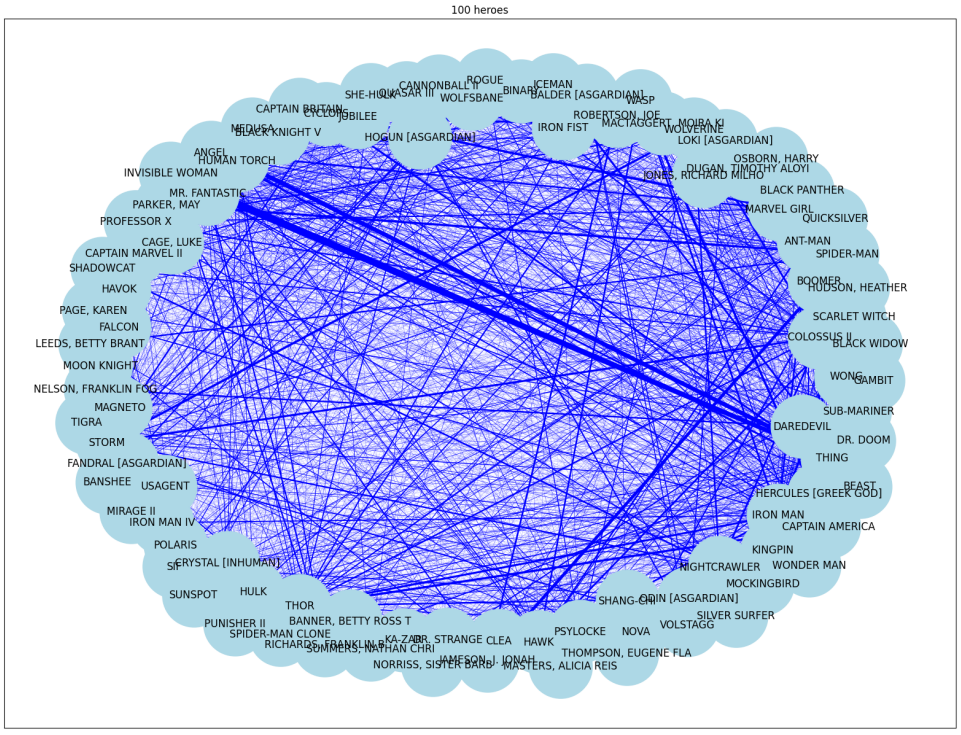

Figure 3: Graph representation of 100 Heroes

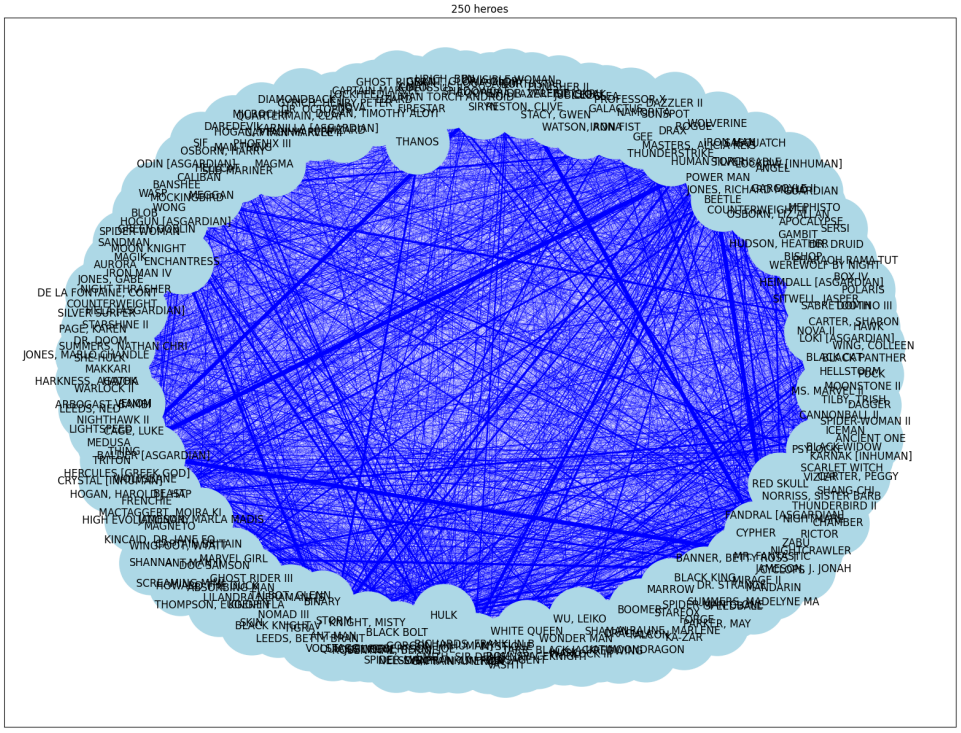

Figure 4: Graph representation of 250 Heroes

5 Measures and Results

5.1 Centrality

These metrics will reveal the influence, connectivity, and key positions of nodes within the network, providing insights into the network's structure and dynamics.

- **Weighted Degree Centrality:** Degree centrality measures the number of connections a node has in a network. Weighted degree centrality takes into account the weights of the edges in a network, giving more importance to nodes with higher-weighted connections. It quantifies the centrality of a node based on the sum of the weights of its incident edges. Higher weighted degree centrality indicates that a node is more central in terms of the overall weight of its connections.

- **Betweenness Centrality:** Betweenness centrality measures the extent to which a node lies on the shortest paths between pairs of other nodes. It quantifies the importance of a node in facilitating the flow of information or resources between other nodes. Nodes with high betweenness centrality act as bridges or brokers between different parts of the network. They have the potential to control or influence the flow of information or resources in the network.

- **Closeness Centrality:** Closeness centrality measures how quickly a node can reach all other nodes in the network. It quantifies the average distance of a node to all other nodes in terms of shortest paths. Nodes with high closeness centrality are more central because they can quickly access or disseminate information to other nodes in the network. Closeness centrality is inversely related to the average distance or path length to other nodes.

- **PageRank Centrality:** PageRank centrality measures the importance of a node in a network based on the concept of "random walk." It originated from Google's PageRank algorithm, which ranks web pages based on their importance on the internet. Nodes with high PageRank centrality are those that are frequently visited or reached by random walks on the network. It considers both the direct and indirect connections of a node, giving higher importance to nodes with incoming links from other important nodes.

- **Eigenvector Centrality:** Eigenvector centrality assigns a score to each node based on the concept of eigenvectors. It considers both the number of connections a node has and the importance of those connections. Nodes connected to other important nodes receive a higher eigenvector centrality score. Eigenvector centrality measures the influence or importance of a node in the network based on the centrality of its neighboring nodes.

After computing these centrality measures for each of the four datasets, it became evident that they exhibited notable variations in values, even across different scales. Consequently, to ensure comparability and meaningful analysis, a standardization process was implemented using a Min-Max Scaler.
The Min-Max Scaler, a commonly employed statistical technique in data preprocessing, serves the purpose of transforming data values to a consistent scale, typically ranging between 0 and 1.
With all centrality values harmonized to a uniform scale through the Min-Max Scaler, a further step was taken. An aggregate measure of general centrality was computed by averaging the

standardized centrality values for each superhero in each dataset. It is crucial to note that each dataset contributes its own general centrality measure for heroes shared among them. This means that if a hero appears in all four datasets, there will indeed be four separate general centrality measures, reflecting the hero's significance within each dataset.

CAPTAIN AMERICA	0.942383
THING	0.866509
HUMAN TORCH	0.846791
MR. FANTASTIC	0.826381
INVISIBLE WOMAN	0.773970
IRON MAN	0.723825
SCARLET WITCH	0.625494
WASP	0.614907
THOR	0.588941
HAWK	0.501198

(a) First 10 heroes according to general centrality with the 25-hero configuration.

CAPTAIN AMERICA	0.985183
THING	0.862353
HUMAN TORCH	0.847411
MR. FANTASTIC	0.829678
INVISIBLE WOMAN	0.792115
IRON MAN	0.782286
SCARLET WITCH	0.739102
THOR	0.714095
WASP	0.704966
BEAST	0.664177

(b) First 10 heroes according to general centrality with the 50-hero configuration.

CAPTAIN AMERICA	0.995617
THING	0.861850
HUMAN TORCH	0.844544
MR. FANTASTIC	0.823227
THOR	0.812398
IRON MAN	0.809004
INVISIBLE WOMAN	0.788319
SCARLET WITCH	0.760382
WASP	0.726475
BEAST	0.711987

(c) First 10 heroes according to general centrality with the 100-hero configuration.

CAPTAIN AMERICA	0.999512
THING	0.856972
HUMAN TORCH	0.841439
MR. FANTASTIC	0.816735
IRON MAN	0.806881
THOR	0.799014
INVISIBLE WOMAN	0.790711
SCARLET WITCH	0.743796
SPIDER-MAN	0.742155
WASP	0.702969

(d) First 10 heroes according to general centrality with the 250-hero configuration.

Figure 5: General centrality of each DataFrame

For the sake of clarity and readability, only the centrality histogram and graph for the first 25 superheroes will be shown below.

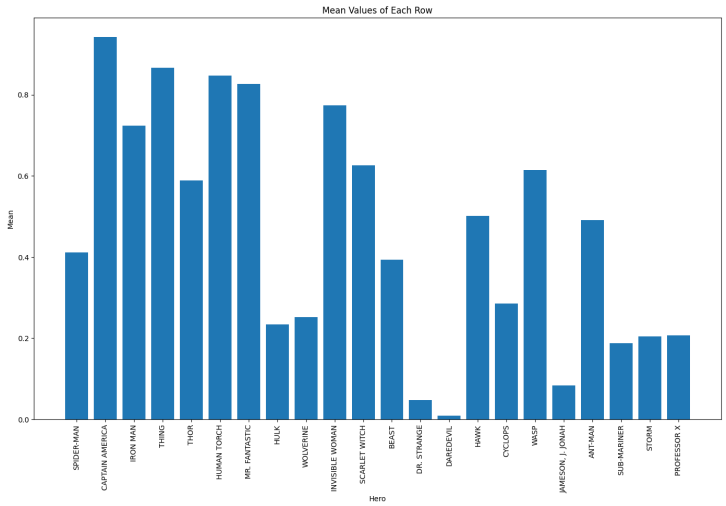

Figure 6: Average centrality of 25 Heroes

Below is the new graph, with nodes sized diagonally according to the centrality of the hero.

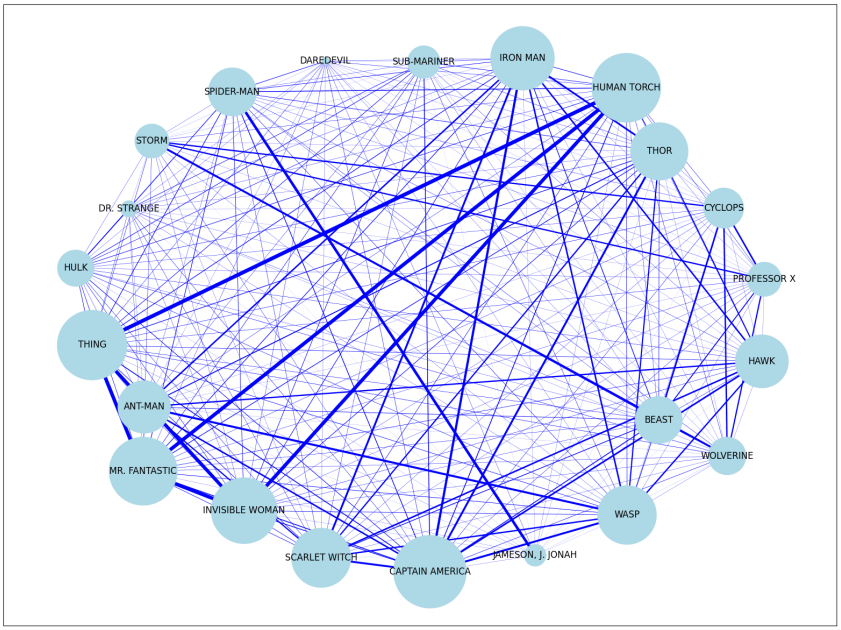

Figure 7: Final graph of 25 Heroes

As we can see, the most important hero in the Marvel network (according also to the other datasets) is **Captain America**!

5.2 Communities Detection

We employ a community detection algorithm from the NetworkX library to explore the formation of character communities within the superhero network. This process reveals tightly-knit groups of superheroes who frequently appear together or share common storylines, offering valuable insights into superhero alliances and dynamics. In detail, here below we can see 4 important communities, easily recognizable thank to the color differentiation. If two heroes belong to the same community, it means that they are used to appearing together, so most likely they are in the same team or squad, or they can be enemies, and so on. For the sake of clarity and brevity, we are going to show the communities of the first 25 heroes; however, it is important to note that the analysis can be extended to encompass a broader array of characters, allowing for a more comprehensive understanding of the intricate social dynamics within the Marvel superhero universe.

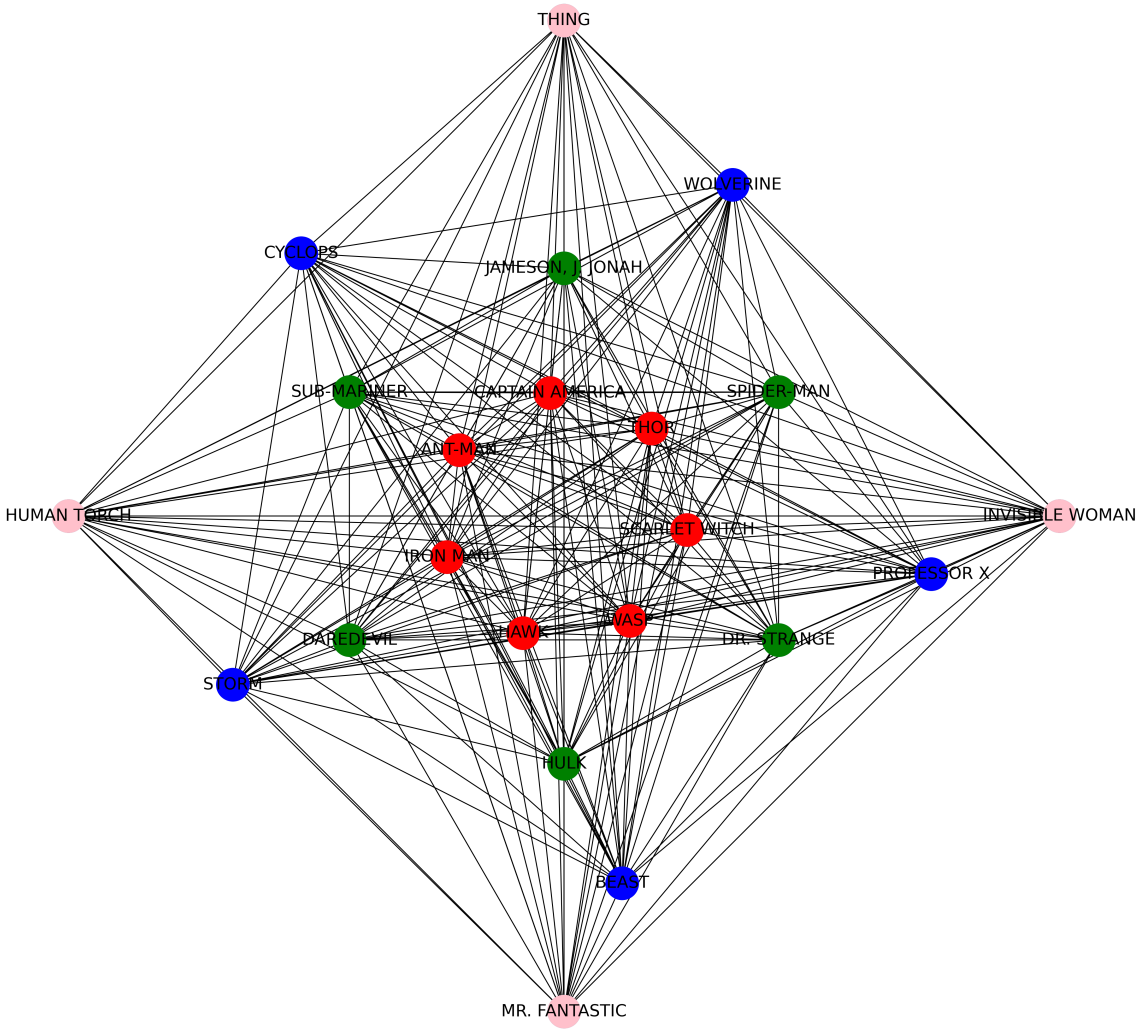

Figure 8: Communities Detection, 25 Heroes

5.3 Cliques Detection

To identify powerful superhero teams and partnerships, we apply thresholding to the edge weights, thereby detecting cliques - dense clusters of superheroes with numerous co-appearances. Indeed, each of these cliques represents a group of superheroes who frequently appear together in various storylines and have established strong connections and relationships within their respective universes. These cliques are likely to be powerful superhero teams or partnerships, as their frequent collaborations suggest a strong ability to work together effectively to tackle various challenges and adversaries. By using thresholding techniques on the edge weights of the network, we have effectively filtered out less significant connections and highlighted the most densely interconnected groups of superheroes. This provides valuable insights into the dynamics and alliances within the superhero world. Overall, the analysis of cliques in this superhero network sheds light on the influential superhero teams and partnerships within the fictional universe, revealing key alliances that contribute to the collective strength and impact

of these iconic characters.
After several attempts the right threshold was found, such that all the graphs show the same biggest cliques.

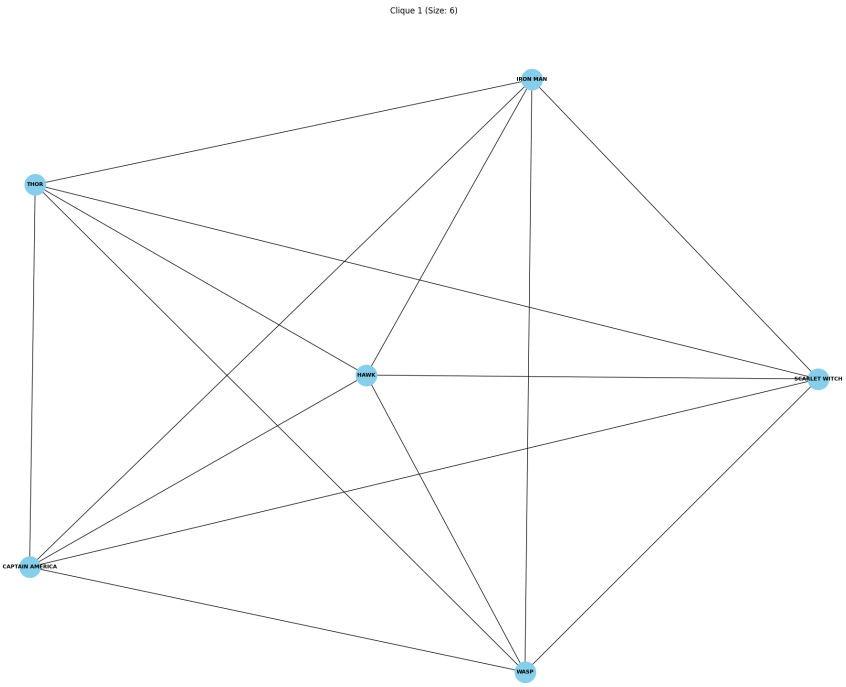

Figure 9: Clique 1

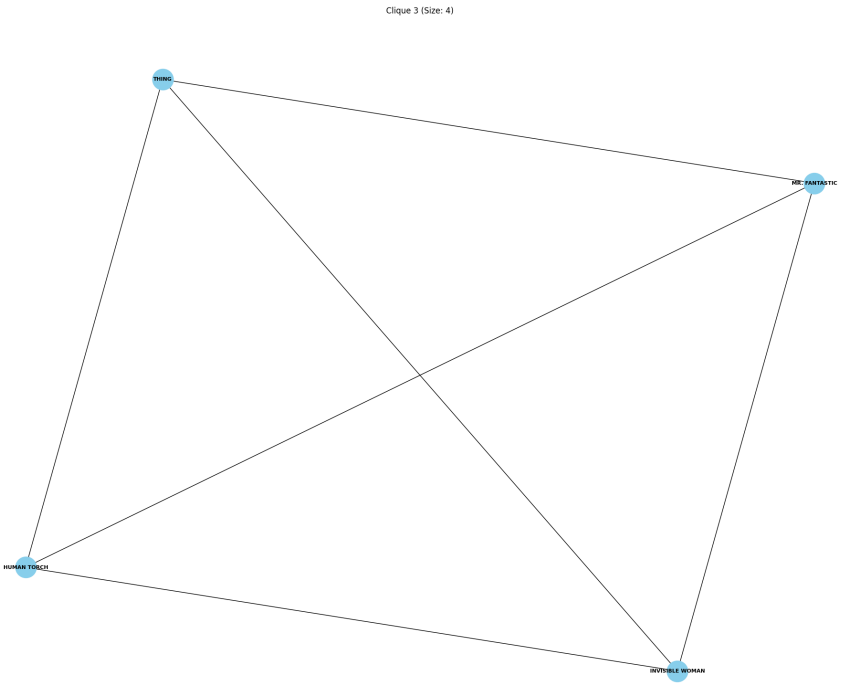

Figure 10: Clique 2

5.4 Robustness

Robustness refers to the ability of a social network to maintain its structural integrity and continue to function effectively in the face of various perturbations, disruptions, or attacks. Robustness is an important concept as it helps us understand the resilience of a network and its ability to withstand different challenges without collapsing or losing its essential properties. In order to understand the pure robustness, we are going to consider only nodes and edges, by ignoring edge weights and centrality measures.
Considering as we mentioned above we do not use the centrality or weight criteria to distinguish the most important node or edges, we are going to base out robustness analysis only on node connections. The 25-hero configuration and the 50-hero one are fully connected, and it is more interesting monitoring a bigger network. For these reasons, I will examine the robustness of the 250-hero configuration.

- **Node Removal:** By simulating the removal of nodes (superheroes) from the network, we observe its impact on network connectivity, identifying critical characters whose absence might significantly affect the network's structure. In order to achieve this, we are going to remove one node only, for testing how much a single hero affects the network. This is done for all the heroes.

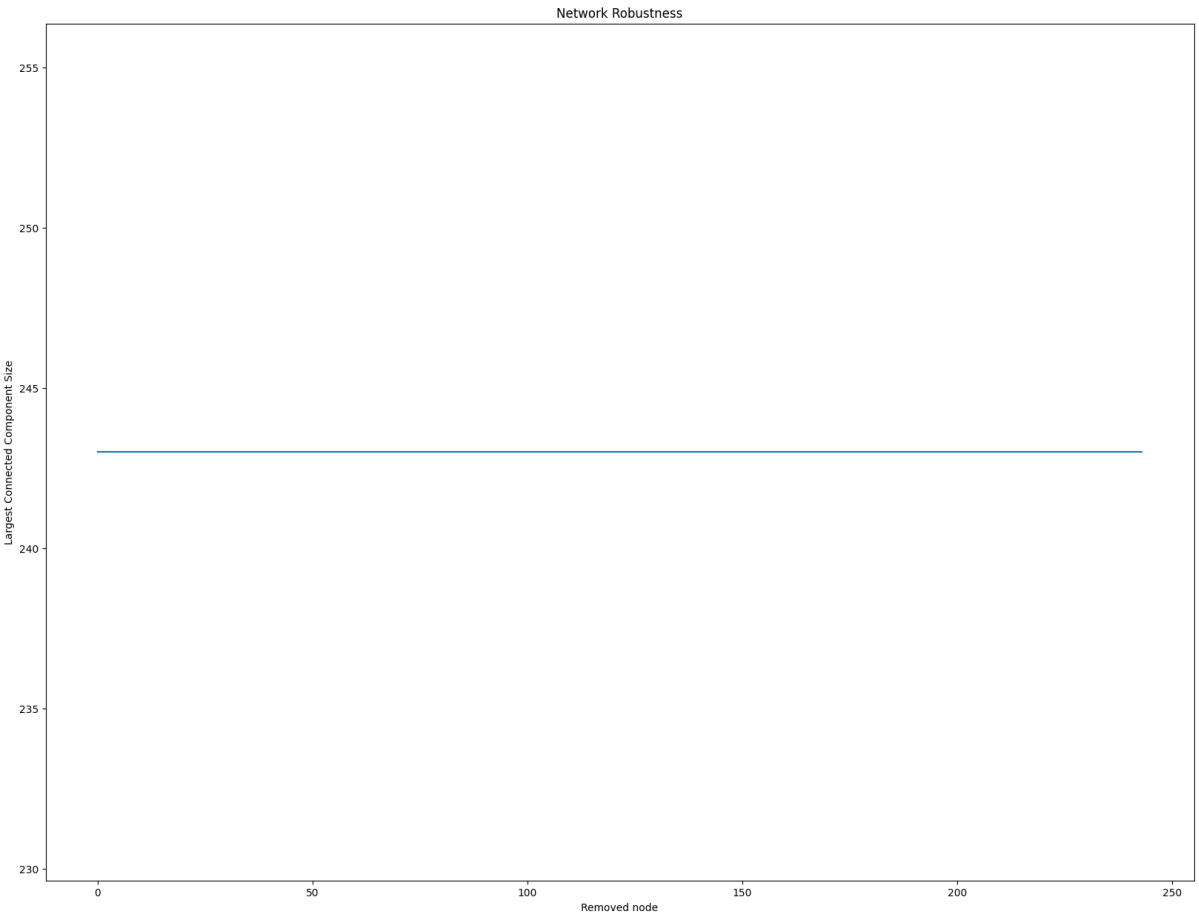

Figure 11: Node Removal

As we can see, even without a node, the network is still connected.

- **Edge Removal:** Examination of the effect of eliminating one edge between superheroes, identifying crucial connections that contribute to the network's overall cohesion. This is done for all the edges.

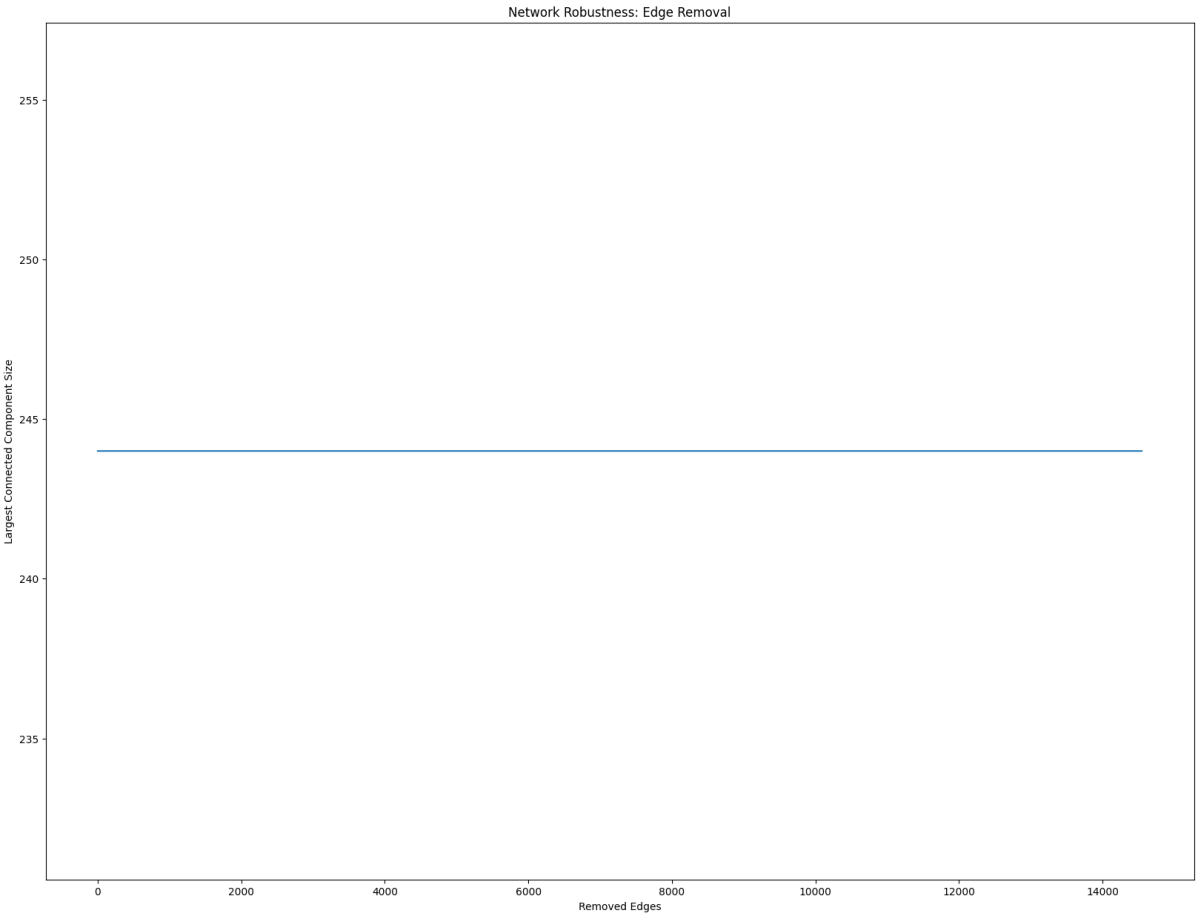

Figure 12: Edge Removal

Considering the total absence of crucial connections, and the equal result about Superheroes removal (node removal), we can say that the network is robust to single node (or edge) removal. So, random failure testing on a single element(node/edge) is useless. Another approach could be random failure testing based on several nodes/edges, But because of the complexity of the network and its connections, it is not possible to find vulnerabilities on the network itself. The next approach is Targeted Attacks.

- **Targeted Attacks:** Identification of critical nodes vulnerable to targeted removal, providing insights into potential weak points in the superhero network. In other words, the aim of this approach is to find the minimum number of nodes to remove in order to disconnect the network. The difference from before is to look for the less connected node and remove all the neighbors. In this way, we remove specific nodes and disconnect the network.

  Considering the nature of the robustness test, we can apply it on all the datasets (25, 50, 100, and 250 heroes), and plot a histogram that shows the minimum number of nodes to remove in order to disconnect the network.

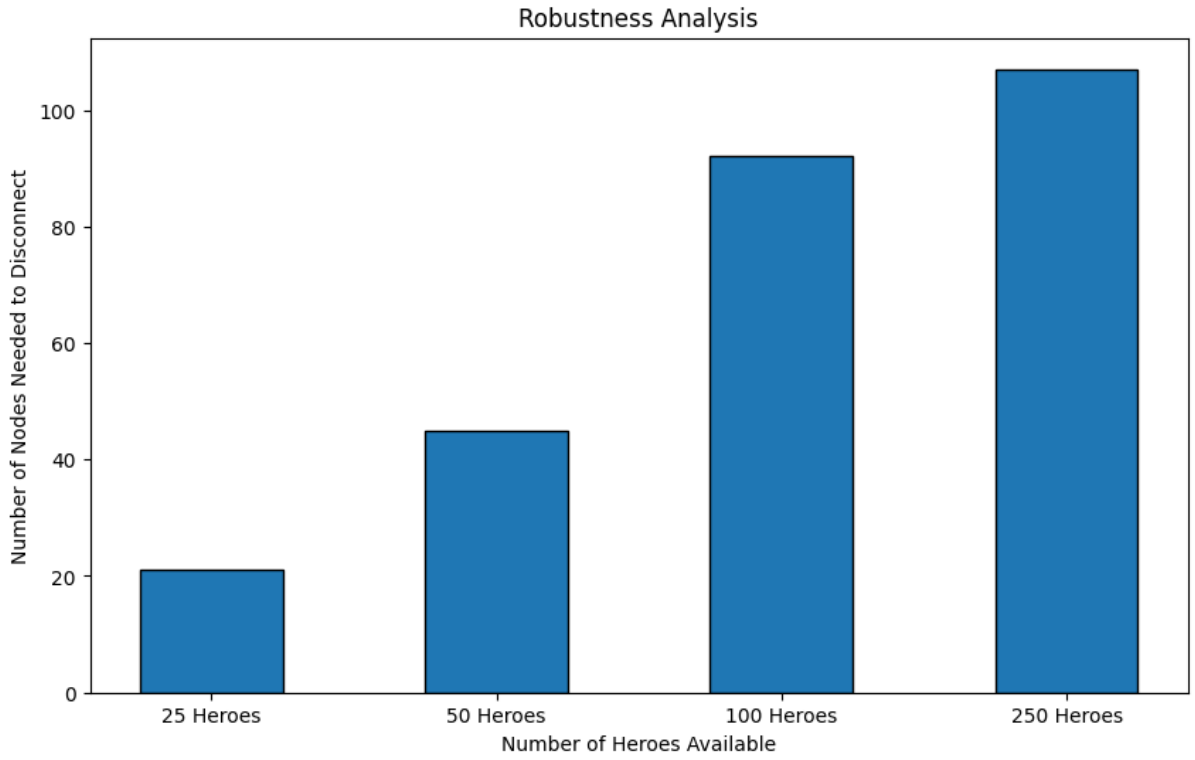

Figure 13: Targeted Attacks

- **Percolation Analysis:** Gradually removing nodes or edges, we are going to assess the network's degradation and identify key components crucial to maintaining network connectivity.

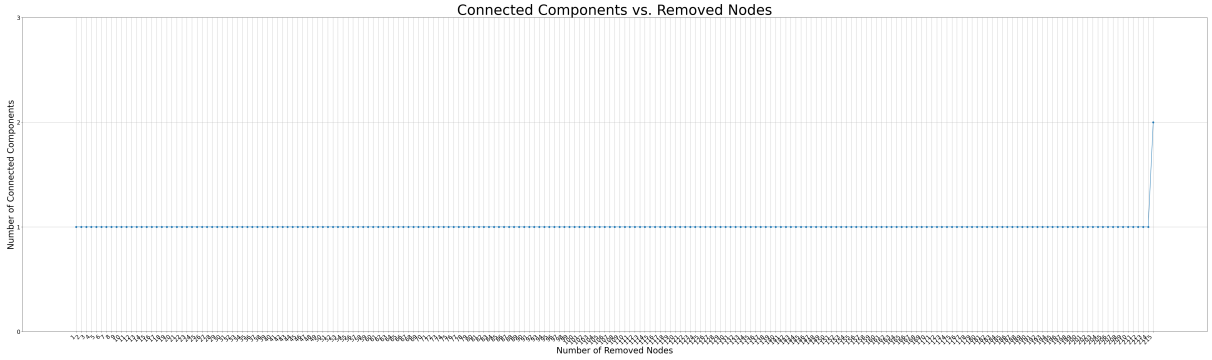

Figure 14: Percolation Analysis

As we can see, our percolation analysis has unveiled a network with exceptional robustness, shedding light on its underlying structure and the strategic nodes that ensure its continued functionality.

6 Conclusion

Throughout this study, a comprehensive investigation into the Marvel superhero network has been conducted, employing quantitative metrics and network analysis techniques. The study

primarily focused on examining the centrality of superheroes, the formation of character communities, and the identification of influential superhero teams and partnerships.
The quantitative findings of this study reveal nuanced insights into the Marvel superhero universe. Centrality measures, including weighted degree centrality, eigenvector centrality, betweenness centrality, closeness centrality, and Pagerank centrality, have been computed to assess the importance of individual superheroes within the network. These measures have unveiled key figures whose presence significantly influences the interconnections of the superhero network.
Furthermore, the study delved into character communities, employing community detection algorithms to identify clusters of superheroes with frequent co-appearances. This qualitative examination highlighted the existence of close-knit groups within the Marvel universe, revealing the collaborative dynamics that shape superhero narratives.
Additionally, by applying thresholding techniques to edge weights, cliques of superheroes with strong co-appearance relationships were identified. These cliques represent strong superhero teams and partnerships, reflecting the collective strength and collaborative efficacy of these iconic characters.
In conclusion, the quantitative findings of this study have enriched our understanding of the Marvel superhero universe. They have clarified the significance of individual superheroes, the formation of character communities, and the influential alliances that contribute to the vibrancy of this fictional world. These findings underscore the complex and interconnected nature of the superhero network and provide a foundation for further exploration and analysis.

7 Critique

In assessing the extent to which this project addresses the identified problem, it's important to acknowledge both its achievements and potential areas for improvement.
To begin, the project has made significant progress in comprehending the Marvel superhero network's structural and relational intricacies. The application of various centrality measures has allowed for a better evaluation of individual superhero importance within the network. This aspect effectively answers a portion of the project problem by shedding light on the significance of specific characters in the Marvel universe. Moreover, the exploration of character communities and the identification of influential superhero teams and partnerships contribute substantially to understanding the network's internal dynamics. In this regard, the project has successfully addressed key aspects of the problem.
However, it is crucial to acknowledge that the Marvel superhero network is exceptionally vast and complex. While this study has provided valuable insights, there remain opportunities for enhancement. Gathering additional data with more detailed character attributes, such as superpowers, affiliations, and character arcs, could have enriched the analysis. Such data could have facilitated a more extended understanding of superhero interactions and their evolution over time. Furthermore, alternative network models and measures could have been explored to capture different dimensions of superhero relationships.
In conclusion, a more extensive data collection effort and the incorporation of alternative measures and models could have further enriched the research. However, given the scope and resources available, the project has made commendable progress in addressing the research problem, offering valuable qualitative insights into the Marvel superhero universe's complex web of characters, relationships, and alliances.