Network Analysis Project

Stefania Bernabè - matricola 986543
LM Informatica, Curriculum B: Informatica per il Management

16.08.2022

1 Context

For this Social Network Analysis project, I decided to study a literature case, based on the relations of the main characters of Marvel superheroes comics (here is a nice list of them: `marvel.com/comics/characters`).
The legendary Stan Lee has created for comics lovers several stories and superheroes, with the help of his fabulous creativity.
As we know, there are a lot of characters in the comics that interact with each other in different stories, and I think it might be interesting to analyze their relationships.

2 Problem and Motivation

Having the full amount of superheroes and their relationships with each other, I thought it might be interesting to analyze their connections.
For example, are there one or more characters in the network that are considered "central" as they appear in many comics and relate to other heroes?

Making logical reasoning, since there are many stories created in the Marvel World, these characters could belong to the leagues of superheroes, as you have more characters fighting together to save the situation and humanity.
And speaking of the groups of heroes, is it possible to find them within the network based on the number of relationships they have with others? That is, for example, by performing an analysis on the net, can we quite easily identify the "Avengers" or "The fantastic 4"?

And following the most recent cinematic events of the Marvel world, if Thanos snaps the fingers pulverizing half of humanity (and then hopefully halving the number of heroes), the results of the above analysis would still be valid?
These are some questions that I asked myself for the development of this project and that I would like to verify.

3 Datasets

The dataset I want to use for this analysis is available online on the Kaggle website with license Attribution 3.0 Unported `https://www.kaggle.com/datasets/csanhueza/the-marvel-`

`universe-social-network?select=edges.csv`.

The data is collected in 3 files with the extension '.csv' containing the names of the Marvel characters present in the dataset, and the relationships with the characters of the Marvel world if they appear together in a comic book.
In the graphical representation of the network, the nodes will represent the characters present in the dataset. The links between these nodes will indicate that the two connected figures have appeared at least once together in the same comic.
For the network study, I decided to use the networkX library available for Python.

4 Validity and Reliability

The dataset includes heroes, comics, and information about their relationships, as we have already said. Three separate files make up the dataset:

- **Nodes.csv**: This file has two columns (node, type), each of which lists the name and kind of the nodes (comic, hero).
- **edges.csv**: This file has two columns (hero, comic), which list the comics where each hero can be found.
- **hero-edge.csv** : The list of heroes who frequently appear together in comic books. The source of this document was located at the following link [1].

During the pre-analysis phase, redundant data had to be removed because the same hero's name was duplicated in the files in two distinct ways (for example, SPIDER-MAN/PETER PARKER and MAN/PETER PAR as a hero name/character name combination).
Following a check of the maximum average length of the hero-network names, the name section was eliminated after the slash.
Going to check and count the connections between 2 heroes defined in *hero-network.csv*, it occurs that the value is not alway equal to the number of intersections present in the other edges.csv file.
This is why the *edges.csv* file was used to rebuild the data. Character names were taken out, and the number of times they appeared together was tallied. With this information, a new undirected graph was created, in which the nodes represent the heroes and the edges reflect the number of comic books in which they appear together.

Regarding the reliability of the data, Cesc Rosselló, Ricardo Alberich, and Joe Miro from the University of the Balearic Islands created the Marvel Comics character collaboration graph in its original form. They contrast the features of this universe with networks of teamwork that exist in the real world, such as the Hollywood network or the network built by scientists who collaborate to produce research publications.
Here are the sources that they cite [2]. The authors used this information to produce the paper *"Marvel Universe looks almost like a real social network"*[3].

5 Measures

The measures that I consider to use for the study of the net are mainly of two types:

- measures of centrality
- measures of grouping

Below is possible to find a brief description of the measures chosen for the analysis and why they will be useful for our study.

5.1 Measures of centrality

Regarding the **centrality** of the characters to understand if they are particularly important and well central nodes of the net, I'm going to use the following indexes

- **Degree centrality**: indicates the number of edges connected to a node. Is the simplest measure of centrality, and usually, nodes with a high degree centrality are those considered the most important within the network.
- **PageRank**: computes a ranking of the nodes in graph G based on the structure of the incoming links. Originally used as an algorithm to rank web pages, it allows giving a value of importance to a node of a network.
- **Eigenvector centrality**: computes the centrality for a node based on the centrality of its neighbors. The eigenvector centrality for node i is the i-th element of the vector x defined by the equation : $\mathbf{Ax} = \lambda \mathbf{x}$, where the value A is the adjacency matrix of the graph G with eigenvalue λ.
- **Betweenness centrality**: is a way of detecting the amount of influence a node has over the flow of information in a graph. It is often used to find nodes that serve as a bridge from one part of a graph to another. The algorithm calculates like $g(v) = \sum_{sd} \frac{n^i_{sd}}{g_{sd}}$, where n^i_{sd} is the number of shortest path from s to d passing from i, and g_{sd} is the number of shortest path from s to d.
- **Closeness Centrality**: Closeness centrality is a way of detecting nodes that can spread information very efficiently through a graph. The closeness centrality of a node measures its average farness (inverse distance) to all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes.

5.2 Measures of Grouping

Instead, for the **grouping** analysis that can be formed in the network, I think it might be useful to use the following measures:

- **K-core**: is a connected set of nodes where each node binds at most to other k nodes. Nodes that are located within the highest level cores are called network cores, while external ones are called peripheral.
  The core, therefore, defines a sort of centrality of the nodes that are part of it.
- **K-component**: is a set of nodes where each member is reachable by the others from at least k unique paths

- **Transitivity**: is the overall probability for the network to have adjacent nodes interconnected, thus revealing the existence of tightly connected communities (or clusters, subgroups, cliques). Is the fraction of all possible triangles present in the graph.
  Possible triangles are identified by the number of "triads" (two edges with a shared vertex).
  The transitivity is $3\frac{tot-triangles}{tot-triads}$.

- **Assortativity**: refers to the tendency of nodes to connect with other 'similar' nodes over 'dissimilar' nodes.
  Assortativity has values between -1 and 1. Values nearly to -1 indicate that we have a negative correlation between the two variables in consideration, otherwise closer to 0 indicates that there is no correlation. In the end, positive values have a positive correlation.

- **Coefficient clustering**: is the number of paths of length two that are closed, divided by the total number of paths of length two in the network.
  The clustering coefficient value can be between 0 and 1. Values closer to 0 mean that neighbors of a node tend to be disconnected, instead, values near 1 indicate a very clustered network. Social networks have a high value of cluster coefficient. In general, nets with a dense number of edges and nodes maintain high values

5.3 In case of Thanos finger snap and final considerations

Calculating the measurements for a standard network with all heroes, I want to try to ask myself instead: what would happen if 50% of the nodes present were removed randomly? The nodes with high centrality in the base case, will continue to have it? And will the different groupings still be present or will they undergo high variations?
During the pre-analysis, in this situation in the optimistic case, I assume that a minimum of the characters belonging among the 20-50 heroes with higher values of centrality, can remain in the system and that therefore in that case all the analysis and the attention will move on these surviving nodes, changing in better the values of centrality. I assume that the same thing can apply to groupings: in the optimal case, the halving of the network would entail the corresponding halving of any groups that have been created and values related to them, and in the worst case can lead to their inevitable disappearance.

Using these values (and possibly others in the course of the research), I believe that a good study of the network and its components can be carried out.

6 Results

6.1 Measures used in the first phase of the study using entire networks

I made some preliminary assumptions based on the recovered data. The thickness of the arcs between nodes is likewise directly proportional to the frequency with which the two nodes appear together. The size of the node is related to its page rank score within the network, and hence its centrality.
These initial observations led us to notice larger points and thicker arches than the others. In order to make it easier to understand visually, some of the nodes' colors have been altered in addition to being specified, taking into account some of the most well-known figures in the

Marvel universe (i.e., CAPTAIN AMERICA, SPIDERMAN, HUMAN TORCH, WOLVERINE, DR.STRANGE, THOR). Initially, I take the 30 top heroes of the net.

The graph created contains 30 nodes and 435 edges and there are 0 hero pairs that were never released together.

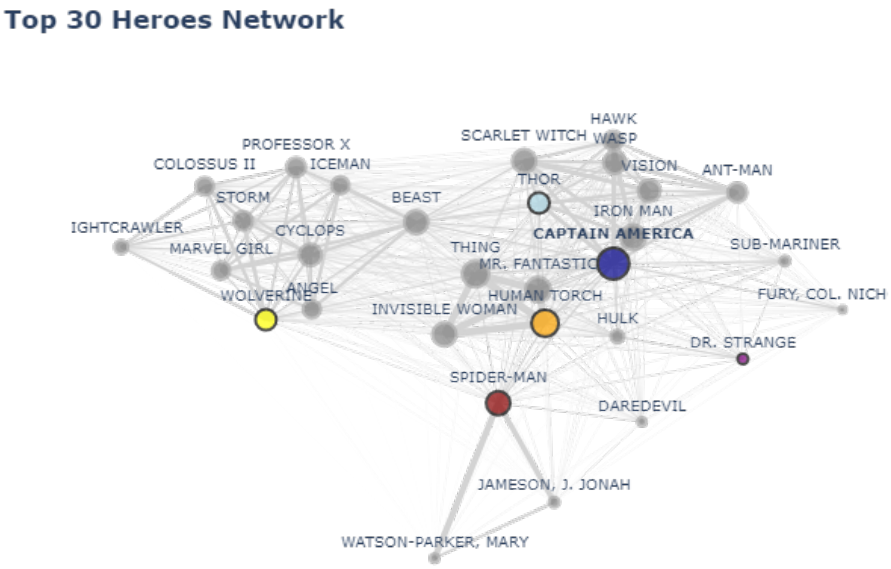

Figure 1: Top 30 Heroes Network

As we can see from the graph, the node that has obtained a higher page rank is the one that represents CAPTAIN AMERICA. So you can consider him "the captain" and then the central point of the network of characters of the Marvel world?
To answer this question, several possible calculations have been applied to determine the centrality of a node. In addition, an average of the results obtained has also been defined. Let's see below.
Since each node's centrality is expressed with large values, they have all been scaled using the MinMaxScaler function to make reading easier.

The betweenness centrality does not provide especially meaningful data for the study given the network's dense composition, therefore I choose to omit it.
Here is a table with the rankings of the first five heroes in the comparison.

hero name	pagerank	eigenvector	degree	closeness	mean
CAPTAIN AMERICA	1.000000	0.983444	1.000000	1.000000	0.995861
THING	0.822634	1.000000	0.842938	0.553226	0.804699
HUMAN TORCH	0.807099	0.991079	0.827660	0.527425	0.788316
IRON MAN	0.740548	0.812783	0.750737	0.756152	0.765055
MR. FANTASTIC	0.770533	0.973639	0.793085	0.497815	0.758768

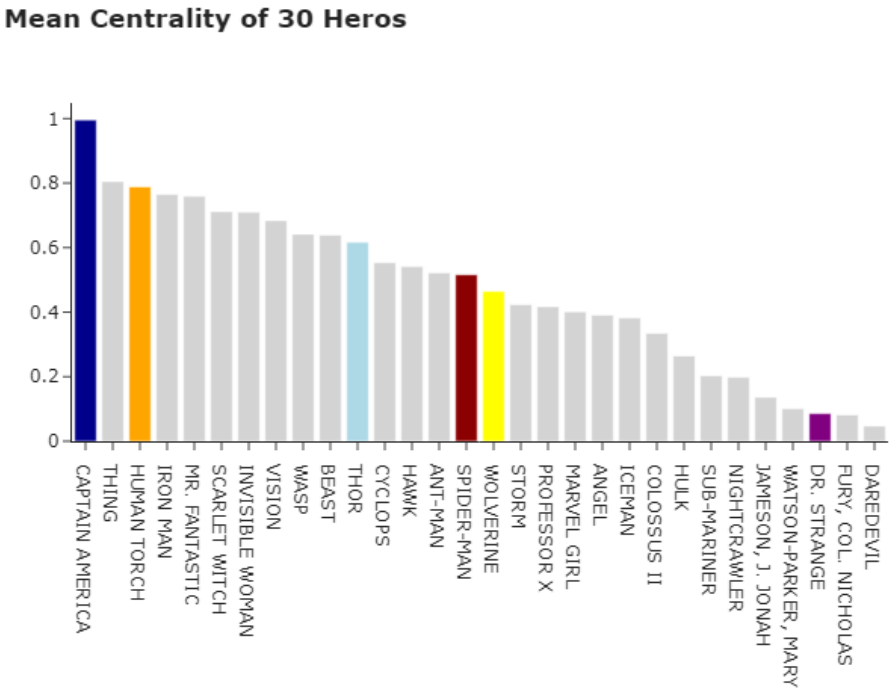

Figure 2: Top 30 Heroes Network

According to the table, Captain America turns out to be the hero with the highest values of centrality, which confirms his role as the most central character within the network.
For easier comprehension, note that its degree centrality is equal to 4646 (higher than all the others), as are its PageRank (0.053595), eigenvector centrality (0.292802), and closeness centrality (0.292801), (128.034244 without scaler).

It was also found a **Jaccard coefficient** for nodes ('CAPTAIN AMERICA', 'HUMAN TORCH'), the first and the third element of the rank, and the result is 0.9333333333333333.
This statistic represents the ratio of the total number of distinct neighbors shared by the two nodes, divided by the number of common neighbors.
Because they share practically all of their neighbors, it is obvious that it has a value very close to 1, further proving how dense of links the network is.
To summarize, the node with the lowest centrality values (not including the mean), ICEMAN, has a degree centrality of 2388.0 (about half that of CAPTAIN AMERICA) and a betweenness centrality of approximately Spider-Man.
Additionally, all this very high centrality values contribute to the concept's strengthening.

After analyzing the centrality measurement data, we check the values that apply to the network as a whole and any potential node groupings.
The results show nothing other than what can be expected from a particularly dense network.
The network's **k-core** value is 30, which is the same as the number of nodes, and for each **k-component** in a range between 1 to 29 we have 30 elements in the set.
The **average clustering** and the assessment of the **shortest path length** are both 1.0. The value of the centrality betweenness is also found in this (which before was not taken into account).
To verify the transactivity of the network, two indicators were used suitable for unconnected graphs such as these, clustering coefficients and triads.

The **clustering coefficient** turns out to be pairs to 1, which indicates a perfect transactiveness of the net.
To verify the **triads**, because of the non-directionality of the arcs, I considered counting the **number of triangles** of the network where a node turns out to be the vertex of the triangle under consideration. Each node of this network turns out to be the vertex of 406 possible triangles.
Nor is it surprising that the **eccentricity** value of each node is equal to 1, which indexes the maximum distance between a node v and any other node of the graph.
To further emphasize the network's connection, I confirmed that 29 nodes would need to be removed in order to produce an unconnected graph. This also represents the **network's degree** value.

The computations below have been performed on networks with various node counts, and even for the highest values (300–400 nodes), there is no difference in the outcomes.
This demonstrates that high levels of network connectivity persist even as the number of nodes rises.
After taking all of these factors into account, let's apply the same conclusions to a network of superheroes that has been cut in half. In the comic books, the story of the incident describes how the evil character Thanos wiped out half of the world's inhabitants with the snap of his fingers after obtaining the power stones.

6.2 Measures used in the second phase of the study using an half-net

To give a basic overview of the event, amid an epic battle between good superheroes and the evil Thanos, the latter gains infinite power and decides to wipe out half of the world's population, shocking his fingers.
I made the assumption that I would randomly delete half the heroes from the dataframe utilized for the analysis in order to recreate the occurrence. In this approach, by removing nodes from the network, their connections to the remaining nodes would likewise have been severed, and as a result, the final graph would differ from one that took into account all of the nodes.

I chose to start with a network of 60 nodes so that, after halving, the number of final nodes is equal to that examined earlier in order to compare it to the previous analysis. This makes it possible to compare the results of earlier calculations.

The graph created contains 30 nodes and 416 edges and now there are 19 hero pairs that were never released together.
As we can already see in the next page, Captain America is one of the heroes who has vanished from the network as a result of Thanos' deadly rage. It is therefore important to confirm which other node assumes the characteristics of the network's central hero.

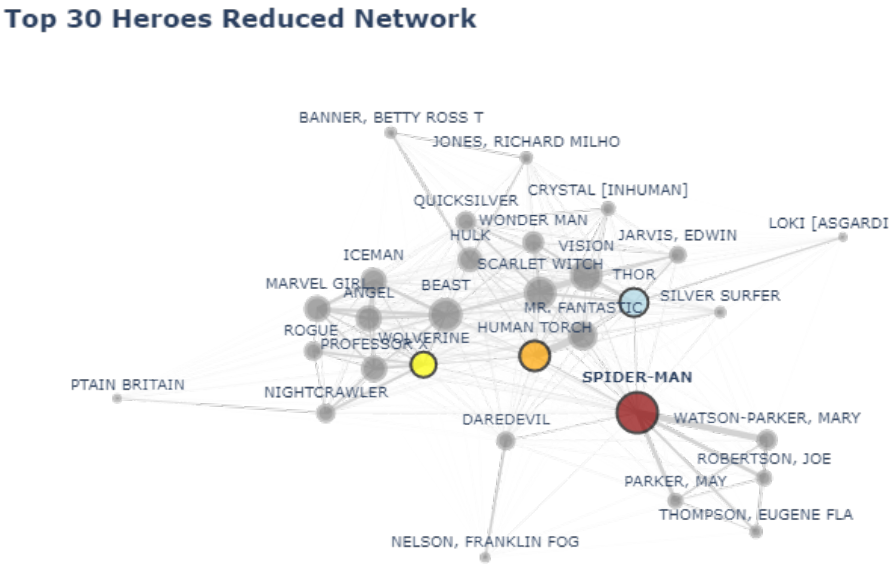

Figure 3: Top 30 Heroes Reduced Network

Visually, the largest node is SPIDER-MAN, but let's see how much its values differ from the other heroes in the top 5 of the centrality values, and how much its values have changed from the last analysis.

hero name	pagerank	eigenvector	degree	closeness	mean
SPIDER-MAN	1.000000	0.734829	1.000000	0.844511	0.894835
BEAST	0.761853	1.000000	0.849747	0.748536	0.840034
SCARLET WITCH	0.728449	0.876990	0.789801	0.952460	0.836925
VISION	0.702798	0.852803	0.762164	1.000000	0.829441
HUMAN TORCH	0.701429	0.854227	0.758661	0.961209	0.818882

The new top 5 predicts different heroes than the previous ones. When analyzing the centrality average, we observe that the value of the main node is also lower than in the previous graph, but overall, when the averages of the other nodes are taken into account, the values are dispersed over a range of $0.8 < x < 0.9$.
Even with the reduction, the network is still rather dense, as we can already see.
The degree centrality with the highest value is SPIDER-MAN (equal to 2810, closeness 68.000719), while the lowest is LOKI [ASGARDIAN] with value 346 and closeness centrality equal to 52.024820.
The **Jaccard coefficient** remained almost unchanged, with a value of 0.9 between nodes SPIDER-MAN and SCARLET WITCH.

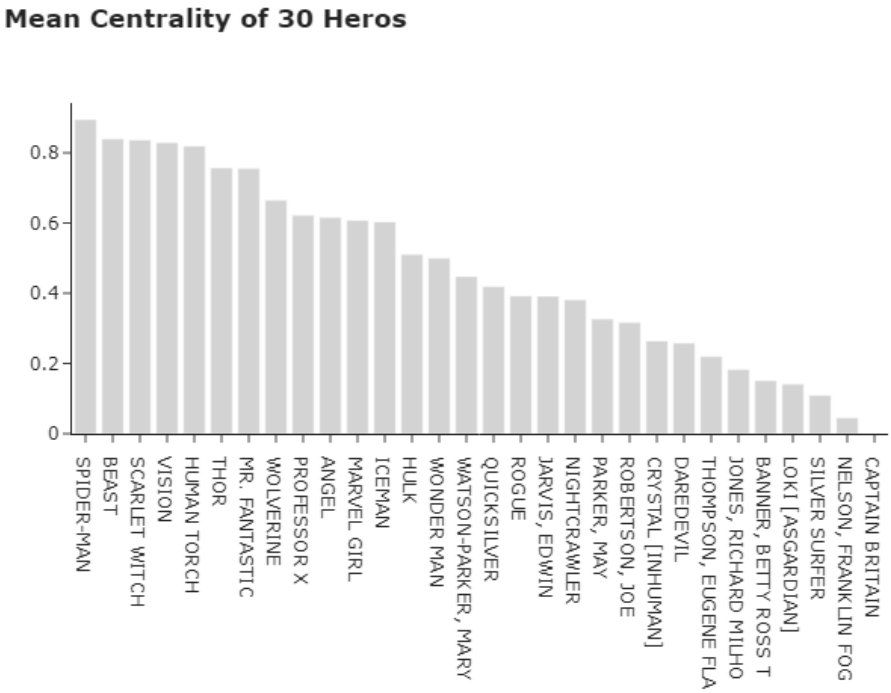

Figure 4: Top 30 Heroes Network

Instead, when we examine the grouping measures, we note some variations from the earlier analysis.
It is interesting to note that the maximum **K-core** for this network has a value k equal to 25 and contains a graph of 28 nodes and 371 arcs. Additionally, when k = 30, the **k components** display more ragged groups.
The **average clustering** and the evaluation of the **shortest path length** are worth about 0.96 and 1.04 respectively, so here too we do not notice too much variation, as well as the **clustering coefficent** still equal to 1.
In cases when not all nodes assume the same value, we discover some variances in the **number of triangles** that can be formed starting from a vertex. The primary node SPIDER-MAN (387) has the highest value, whereas node THOMPSON, EUGENE FLA has the lowest value (206).
Even, the **transitivity** varies slightly (around 0.96), as does the **diameter** of the net, which is equal to 2, and the **eccentricity** of the nodes, which can take not just 1 but also 2 value in some circumstances. The **net degree** also varies slightly from a maximum value of 29 to a minimum of 21 for the node THOMPSON, EUGENE FLA.
As a result, **the minimum number of nodes** that must be excluded from the network to create a disconnected graph drops to 21, but still indicates that the network remains robust.

These findings demonstrate that despite a sharp halving of nodes, the interconnectivity characteristics have remained mostly similar and have not undergone of interconnectivity have stayed mostly similar and have not undergone too significant decreases.

7 Critique

The first hypotheses proposed were excessively gloomy in light of the values discovered, which I did not anticipate. The groups of nodes not only continue despite halving, but they also continue to maintain extremely high values, demonstrating that the network has remained dense nonetheless.
The assortativity was the only variable whose calculation presented a challenge according to the project concept. The network function I used produced a problem that I was unable to fix and that I also discovered online where other users had experienced. As a result, I was unable to express the function's results in the report.
Also for the triads, I would have to recreate the net so that it turned out directed, but for the amount of work, I preferred to leave it undirected.

However, the outcomes are in accordance with my expectations, particularly for the first half. Calculations regarding the second part vary per attempt, as the nodes to be deleted are selected randomly with each execution of the code. However, I want to specify that before transcribing the results obtained on a specific execution, I started the code several times considering the same number of nodes and I did not find large variations in the results.

References

[1] Kai Chang, Tom Turner, and Jefferson Braswell. Marvel Universe Social Graph. `http://syntagmatic.github.io/exposedata/marvel/`, 2011.

[2] Cesc Rosselló, Ricardo Alberich, , and Joe Miro. Social characteristics of the Marvel Universe. `http://bioinfo.uib.es/~joemiro/marvel.html`, 2002.

[3] Cesc Rosselló, Ricardo Alberich, , and Joe Miro. Marvel Universe looks almost like a real social network. `https://arxiv.org/abs/cond-mat/0202174`, 2002.