How the Avengers Assembled? Analysis of Marvel Hero Social Network

Xuan Yu, Ziyang Ren, Chongyang Shi

Abstract

The movies of Marvel universe are very popular among young people. Almost every young people nowadays know some of the heroes in Marvel universe, such as iron man and spider man. The data set named The Marvel Universe Social Network (MUSN) describe the social relationships of the heroes. By analyzing the MUSN, we establish a social network of Marvel heroes. We derive some basic statistics from the Marvel network, such as the number of nodes and links, the hubs, the components, the shortest path lengths and the diameter. In the next part, we analyze the structure of the Marvel network and obtain some results of the connectedness, the clustering, the degree distribution, the degree correlation. Meanwhile, we fit the power law and divide the network into different communities. In this process, we not only find that the network appears to have the small world nature, since it is obviously a scale-free network, but also find that it is very similar to the real-world social network. Further, based on the work of Loverkar et al., we do a hypothesis test on the small world nature of the network. We find that the Marvel network does have small world property. In the end, we visualize the Marvel network to give a better understanding of the network.

Keywords: Marvel University, Scale-Free Network, Null Model, Hypotheses Testing

1 Introduction

The Marvel Universe is an overhead world composed of many original characters and stories based on series of comics, movies, and animations. As the Marvel series became popular around the world, more and more people became fans of these heroes, worshiping them who defend the justice of the world, and caring about their disputes or friendship.

Therefore, we collect a data set of the relationship between Marvel heroes from **kaggle** and hope to study the social network formed by them. Then compare with the real social network and explore some characteristics of the hero social network. The dataset is from Claudio Sanhueza (2016). *The Marvel Universe Social Network An artificial social network of heroes (version 1).* Retrieved June, 2021 from *https://www.kaggle.com/csanhueza/the-marvel-universe-social-network.*

2 Basic Statistics

2.1 Nodes and links

We believe that there is a connection between heroes who appear in the same comic or movie. So, there is a link between each of these heroes.

Then, in the Marvel data set, there is 6582 nodes in the Marvel Universe network and 167219 links in total. The average degree is 50.81.

2.2 Hubs

By order the number of neighbors, the top 5 nodes are

Nodes	Captain America	Spiderman	Ironman	Thing	Mr.fantastic
Number of Neighbors	1907	1737	1522	1416	1379

Those 5 heroes can be regarded as the center of the 5 largest hubs. So if danger comes and you want to gather the Avengers, you can save a lot of energy by contacting these five people first.

2.3 Component

The Marvel network (denoted by G_0) is composed of 160 components in total. The largest one has 6408 nodes and 167163 links, while the second one has only 9 nodes and 34 links. The remaining 156 components are all scattered points.

Component	Node	link
Largest one	6408	167163
The Secend	9	34
The third	7	21
The fourth	2	1
5th-160th	1	0

Thus, we find that the network consists of a very large component and many small components. This shows that a few heroes don't like to interact with other heroes. Maybe they prefer to save the world alone.

2.4 The Shortest Path Lengths and Diameter

Since the largest component (denoted by G) is not much different from the overall network G_0 and G is connected, **from now on, we only focus on the structure and characteristics of G.**

The diameter of G is 5. And the average shortest path length of G is 2.63843 (as Fig.8 shows in appendix), which means each person is separated from the others by 2 people. According to the principle of Six Degrees of Separation, a certain hero can know any other hero through a maximum of three people.

3 Structure

3.1 Connectedness

Transitivity is used to measure the connectedness in this report, which measures the density of loops of length three (triangles) in a network. It is the overall probability for the network to have adjacent nodes interconnected, thus revealing the existence of tightly connected communities. As Rodrigue (2002) stated,

$$T = \frac{observed\ number\ of\ closed\ triplets}{maximum\ possible\ number\ of\ closed\ triplets\ in\ the\ graph}$$

In the Marvel network, the transitivity is 0.19453, which is quite small. There is few loop between heroes, maybe because they have to contact one-line to keep secrets.

3.2 Clustering

Clustering Coefficient for Iron Man

According to the Marvel comics, Anthony Edward "Tony" Stark was a billionaire industrialist, a founding member of the Avengers, and the former CEO of Stark Industries, who have wide connections in Marvel world. Thus he is chosen for us to analyze.

Iron man has 1522 acquaints and the clustering coefficient of him is 0.062, which reveals that there is little connection between his acquaints.

Clustering Coefficient with Degree

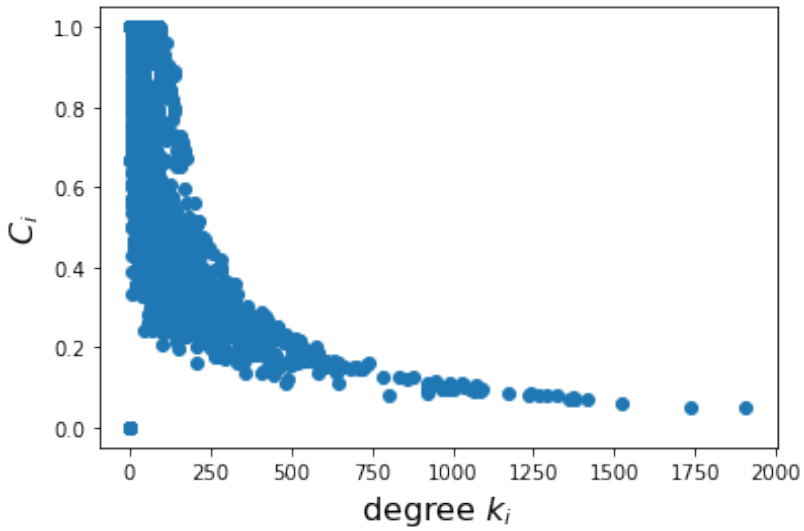
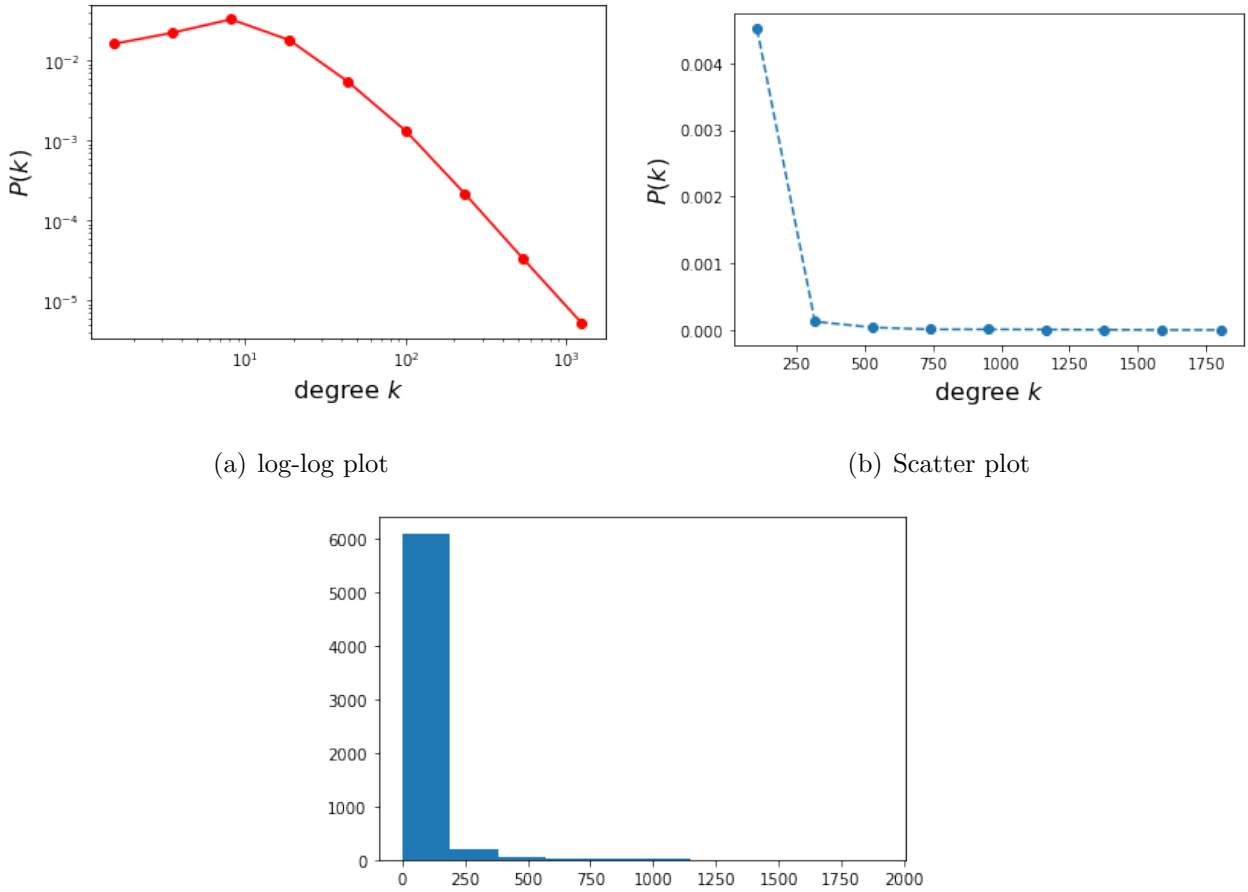


Figure 1: Clustering Coefficient with Degree

The clustering coefficient is large (close to 1) at small k, which is corresponding to the rules that the number of occurrences of hero behavior should have a certain close relationship between each stage (Fig.1).



(a) log-log plot

(b) Scatter plot

(c) histogram

Figure 2: Degree Distribution

3.3 Degree Distribution

The degree distribution of the Marvel Universe network shown in Fig.6. The observed degrees vary between k=1 (isolated nodes) and k=1908, which is the degree of the most connected node. There are also wide differences in the number of nodes with different degrees: Large amount of the nodes have degree smaller than 100, while the $P(k)$ increase slightly at first and decline with the increase of k. This is quite similar to that of the scale free network.

3.4 Fit the Power Law

In the network, we have $p_k \sim k^{-\gamma}$. In Marvel network, 110 is the point from which the data displays the power-law behavior. γ is quite close to 2.5. Then we do a comparison between the power law and the log normal positive model, it turns out that the p-value is $1.59 * 10^{-13}$. The power law fits well.

The fit result can be shown in (Fig.3):

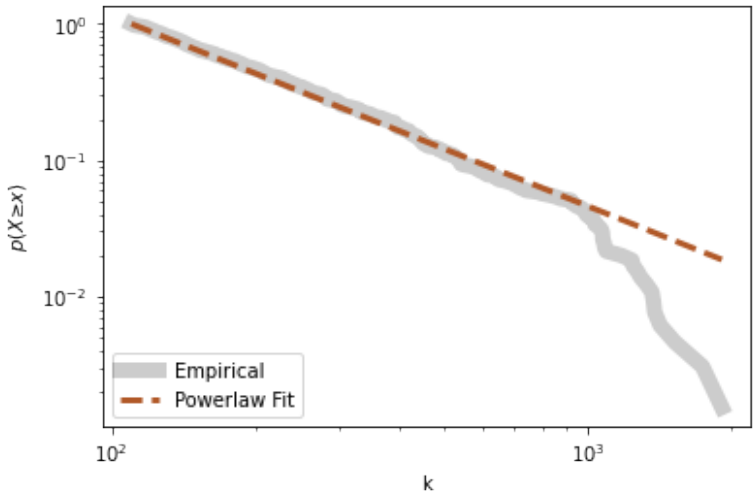


Figure 3: Fit the Power Law

It is obvious that there is large errors at the tail of the curve. At large k in our observation, the probability is relatively smaller than that of the prediction, which can be inferred that there are fewer heroes having wide connections than we expected.

3.5 Community

Greedy algorithm is an algorithm that takes the best or optimal (that is, the most advantageous) choice in the current state in each step of the selection, so as to hope that the result is the best or optimal algorithm. In this report, the best partition found consists of the 66 communities, that is 66 groups of Marvel heroes. The modularity of this partition is 0.359. It is a relatively clear partition.

4 Exploration

4.1 Small World Property

According to the previous analysis, we get $\gamma = 2.39$ from which we indicate that it satisfies the small world property. Thus, we do a hypothesis testing and get the average distance of the heroes is 2.638, that is, a hero needs to contact 2 friends for a targeting strange hero.

4.2 Null Model

In order to test the small-world property of the network, we refer to the work of Lovekar et al. (2021). For a particular network, we consider the number of nodes n, the expected degree 2δ, and the mixing proportion $\beta \in [0, 1]$ which is defined in the paper of Lovekar et al. (2021). The null model is defined as a pure Erdös-Renýi random graph with n nodes and $p = \frac{2\delta}{n-1}$. Under this model, we define the detection of small world property as the test of the hypothesis

$$H_0 : \beta \in \{0, 1\} \quad v.s. \quad H_1 : 0 < \beta < 1.$$

The null hypothesis asserts that the network is either a pure ER model graph with parameters $(n, \frac{2\delta}{n-1})$ or a pure ring lattice, while the alternative model denoted as NW-ER$(n, 2\delta, \beta)$ implies the presence of significant small-world property. We define $C = \frac{3T}{3T+V}$ is the clustering coefficient or transitivity of the graph and L is the average (shortest) path length of the graph. Lovekar et al. (2021) proposed a multiple testing procedure with two test statistics $[C, L]$, which we call the intersection test. In particular we reject the null hypothesis if,

$$\{C > K_1\} \cap \{L < K_2\}$$

for suitable choices of K_1 and K_2. The quantities K_1 and K_2 are determined by bootstrap method which involves fitting the respective null model to the observed data to estimate the parameters of the null model. Let K_1 be the 95th percentile of the distribution of C and K_2 to be the 99th percentile of the distribution of L. We generate 100 bootstrap simulations, the result is as follows.

Model	C	L	Decision
ER Network	0.008224	2.643	Reject
Hero Network	0.1945	2.638	

Table 1: The result of hypotheses testing

From the table 1, it's clear that $C > K_1$ and $L < K_2$, we reject the null hypothesis. Thus, the Marvel network has small world property.

4.3 Degree Correlation

After calculate the **degree correlation coefficient**, we obtain $r = -0.1620953256887782$, in which

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_r^2} \qquad -1 \leq r \leq 1$$

And through the definition of the degree correlation coefficient,

$$\begin{cases} r \leq 0 & disassortative \\ r = 0 & neutral \\ r \geq 0 & assortative \end{cases}$$

So the network should be disassortative (as Fig.4 shows). But we did not allow two nodes to have multiple edges, so we still cannot determine whether its disassortative is stuctural disassortativity caused by structural cutoff. Therefore, do **degree-preserving randomization** to decide whether the correlations observed in the Marvel network are a consequence of structural disassortativity, or are generated by some unknown process that leads to degree correlations.

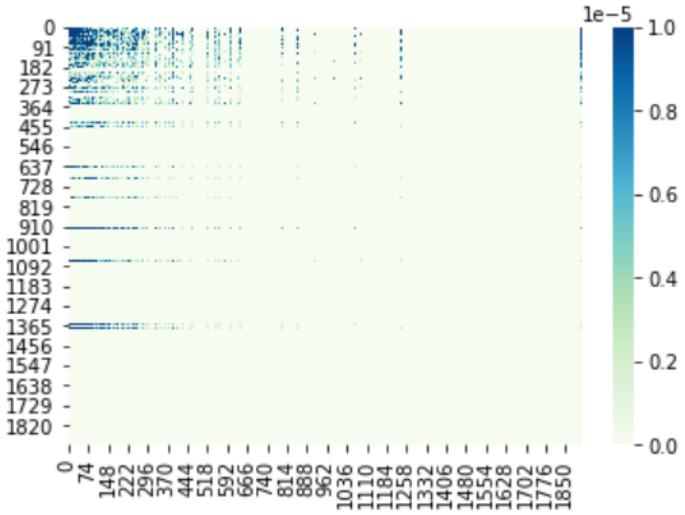


Figure 4: Degree Correlation Matrix

We apply degree-preserving randomization to the original network and at each step we make sure that we do not permit more than one link between a pair of nodes. Leung and Chau (2006) stated if the real $k_{nn}(k)$ and the randomized $k_{nn}^{R-S}(k)$ are indistinguishable, then the correlations observed in a real system are all structural, fully explained by the degree distribution.

After calculation, it is found that the degree correlation coefficient of the degree-preserving randomization network is $r = -0.10581477$, which is not far from the coefficient of the original

network. So, the Marvel networl is **structural disassortativity** and hubs tend to connect to small nodes.

This shows that a superhero will help those who have little influence, or a small hero will seek the help of a superhero. Of course, it may originate from that comic books about little heroes are not selling well and need to be linked with superheroes to promote sales.

5 Visualization

Since the largest component is too large, only nodes with a degree greater than 100 are selected for visualization (Fig.5).

For the network made by selecting nodes, it contains 715 nodes, 54760 edges and 5 communities. The average degree is 153.175, the network diameter is 3, the average clustering coefficient is 0.601, and the average shortest path is 1.788.

So in this network, each hero is separated from the others by only one hero. According to the principle of Six Degrees of Separation, a certain hero can know any other hero through a maximum of two people. This shows that many heroes are closely connected, and it is easy for them to find heroes he didn't know before through their social networks.

6 Conclusion

According to the previous work of Alberich et al. (2002) and our work on the Marvel network, we note that the Marvel network is a scale-free network which has the small world property. It is very similar to the real-world social network and the average shortest path length is only about 2.6 which means only three heroes are needed to connect with another hero for a hero.

Recall the title *How the Avengers Assembled?* Here we may get the answer to this question. Because of the small world property and the tiny average shortest path, heroes can easily contact each other. And perhaps that is why avengers could assembled in such a fast speed and defend the world's justice.

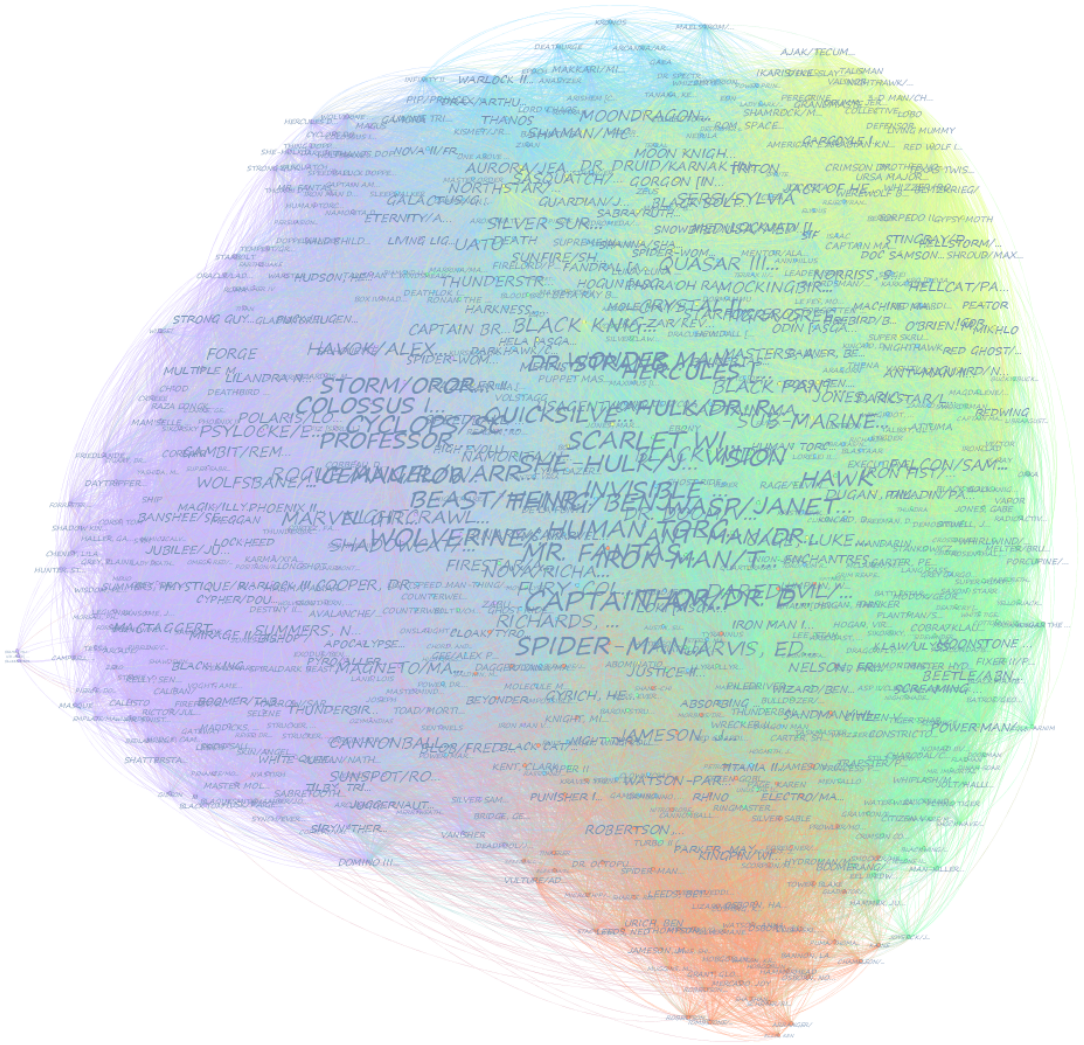


Figure 5: Visualization for those nodes with $degree \geq 100$ in the Marvel network

References

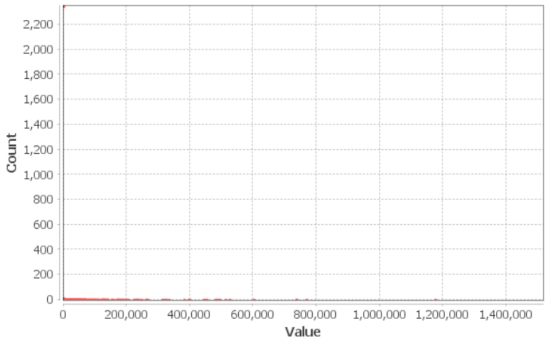
Rodrigue, J.-P. The Geography of Transport Systems FIFTH EDITION. 2002.

Lovekar, K.; Sengupta, S.; Paul, S. Testing for the Network Small-World Property. 2021.

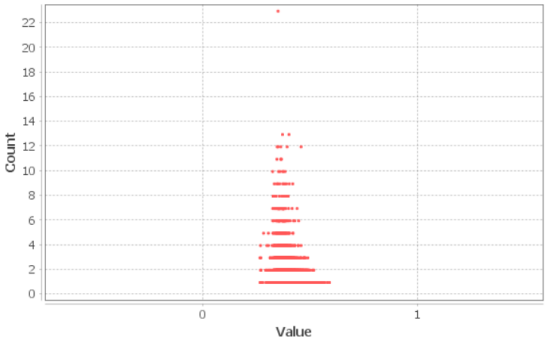
Leung, C. C.; Chau, H. F. Weighted Assortative And Disassortative Networks Model. 2006.

Alberich, R.; Miro-Julia, J.; Rossello, F. Marvel Universe looks almost like a real social network. 2002.

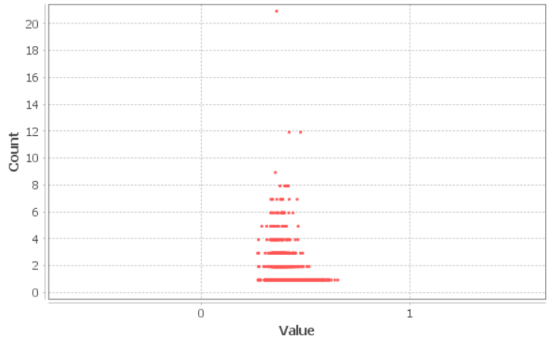
Appendix



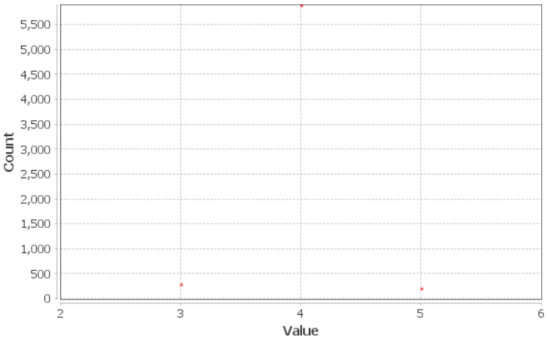
(a) Betweenness Centrality Distribution



(b) Closeness Centrality Distribution



(c) Harmonic Closeness Centrality Distribution



(d) Eccentricity Distribution

Figure 6: Graph Distance Report of G