DSA5101 - Introduction to Big Data for Industry

National University of Singapore

Faculty of Science

Marvel Universe Hero Network Analysis

Authors:
Yu Haowei (ID: A0330270W)

Abstract

This study investigates the structural properties and community dynamics of the Marvel Universe hero network, constructed from 574,467 co-appearance records involving 6,426 unique heroes. We employ network analysis techniques, including degree, betweenness, closeness, and PageRank centralities, to identify the most influential heroes and assess their strategic importance within the network. Using the Louvain algorithm for community detection, we uncover 28 distinct communities, corresponding to recurring teams or tightly connected hero groups. Analysis of inter-community edges highlights frequent collaborations and cross-over events among major communities. Network visualizations further illustrate the distribution of influence and community structures, revealing dense clusters anchored by central heroes. Our findings provide a comprehensive view of hero connectivity, influence, and community organization, offering insights into the underlying social structure of the Marvel Universe.

1 Introduction and Dataset Overview

1.1 Dataset Description

The dataset used in this study is the *Marvel Universe Social Network*(available on Kaggle). It contains 574,467 records of hero co-appearances in Marvel comics. Each row represents a pair of heroes who appeared together in the same comic issue. The dataset allows for the construction of a social network where nodes represent heroes and edges indicate co-appearances.

Table 1: Sample records from the Marvel hero co-appearance dataset

hero1	hero2
LITTLE, ABNER	PRINCESS ZANDA
LITTLE, ABNER	BLACK PANTHER/T'CHAL
BLACK PANTHER/T'CHAL	PRINCESS ZANDA
LITTLE, ABNER	PRINCESS ZANDA
LITTLE, ABNER	BLACK PANTHER/T'CHAL

The dataset contains 6,426 unique heroes across the two columns, with no missing values detected. Preliminary data cleaning included removing duplicate entries and self-loops to ensure network consistency.

1.2 Graph Network Construction

An undirected graph G was constructed using the hero co-appearance data. Each node represents a hero, and each edge represents a co-appearance between two heroes within a comic issue. Self-loops were removed to maintain consistency.

The resulting network has the following structural properties:

- Total nodes: 6,426
- Total edges: 167,207
- Network density: 0.0081
- Number of connected components: 4
- Largest connected component: 6,408 nodes (99.7% of all nodes)
- Average clustering coefficient: 0.7747

The network is relatively sparse but exhibits a highly interconnected core. The high clustering coefficient indicates strong local cohesion and potential community structures.

1.3 Network Statistics and Degree Analysis

Node degree, defined as the number of connections a hero has, was computed to assess connectivity patterns. Key statistics are summarized in Table 2.

Table 2: Degree statistics of the Marvel hero network

Statistic	Value
Average degree	52.04
Median degree	20
Maximum degree	1906
Standard deviation	113.64

The network exhibits a heterogeneous degree distribution. While most heroes have relatively few connections, a few hubs such as *CAPTAIN AMERICA*, *SPIDER-MAN*, and *IRON MAN* have extremely high degrees. These hubs play a central role in maintaining network cohesion.The top 5 heroes by degree are listed in Table 3.

Table 3: Top 5 heroes by degree

Hero	Degree (connections)
CAPTAIN AMERICA	1906
SPIDER-MAN/PETER PAR	1737
IRON MAN/TONY STARK	1522
THING/BENJAMIN J. GR	1416
MR. FANTASTIC/REED R	1379

1.4 Degree Distribution Visualization

The degree distribution was visualized on both linear and logarithmic scales (Figure 1). The linear-scale histogram shows the overall spread of hero connections, while the log-scale highlights the presence of highly connected hubs. These visualizations confirm that the network exhibits a *scale-free* structure: a few heroes act as highly connected hubs, while most have relatively few connections.
Insights from the degree distribution:

- The network exhibits a heavy-tailed distribution.
- Few heroes have very high degrees (hubs).
- Most heroes have relatively few connections.
- This pattern is characteristic of scale-free networks.

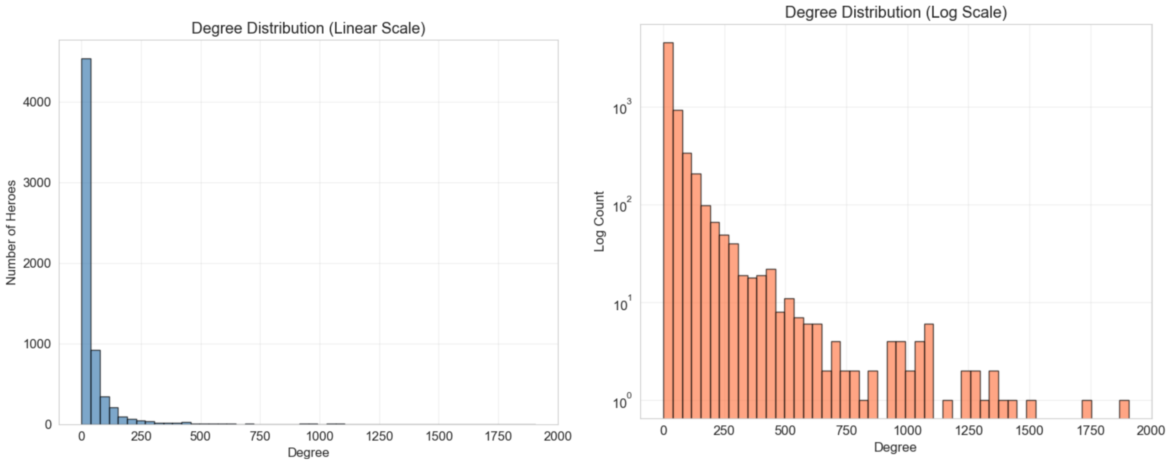

Figure 1: Degree distribution of the Marvel hero network. (Left) Linear scale; (Right) Logarithmic scale.

2 Centrality Analysis of Marvel Heroes Network

To understand the influence and strategic importance of heroes within the Marvel Universe network, we compute several centrality measures. These metrics reveal not only the most connected heroes, but also those acting as bridges between communities or possessing strong potential to propagate influence.

2.1 Centrality Measures

2.1.1 Degree Centrality

Degree centrality counts the number of direct connections a node has. For node i, the degree centrality $C_D(i)$ is defined as:

$$C_D(i) = \frac{deg(i)}{N-1} \tag{1}$$

where $deg(i)$ is the degree of node i, and N is the total number of nodes. Heroes with high degree centrality are directly connected to many other heroes, serving as prominent hubs.

2.1.2 Betweenness Centrality

Betweenness centrality measures the frequency a node lies on the shortest paths between other nodes:

$$C_B(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2}$$

where σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(i)$ counts those passing through i. Heroes with high betweenness centrality bridge different communities and facilitate information flow.

2.1.3 Closeness Centrality

Closeness centrality quantifies how close a node is to all others:

$$C_C(i) = \frac{N-1}{\sum_{j\neq i} d(i,j)} \tag{3}$$

where $d(i,j)$ is the shortest path distance between nodes i and j. Heroes with high closeness centrality can quickly interact with or influence the entire network.

2.1.4 PageRank Centrality

PageRank evaluates node influence by considering the importance of its neighbors. For node p:

$$PR(p) = \frac{1-\alpha}{N} + \alpha \sum_{q\in M(p)} \frac{PR(q)}{L(q)} \tag{4}$$

where α is the damping factor (typically 0.85), $M(p)$ is the set of nodes linking to p, and $L(q)$ is the number of outgoing links from q. In the Marvel hero network, PageRank identifies heroes connected to other influential heroes, even if their direct degree is moderate.

2.2 Top Heroes by Centrality Measures

Table 4: Top Heroes by Different Centrality Measures (Top 5)

Rank	Hero	Degree	Betweenness	Closeness	PageRank
1	CAPTAIN AMERICA	0.2967	0.0552	0.5837	0.00512
2	SPIDER-MAN/PETER PAR	0.2704	0.0766	0.5741	0.00532
3	IRON MAN/TONY STARK	0.2369	0.0320	0.5614	0.00411
4	THING/BENJAMIN J. GR	0.2204	0.0239	0.5578	0.00367
5	MR. FANTASTIC/REED R	0.2146	0.0242	0.5561	0.00357

Table 4 presents the top five heroes for each centrality metric in the Marvel hero network. These centrality measures provide different perspectives on a hero's importance.

From the table, we observe that some heroes, such as *CAPTAIN AMERICA* and *SPIDER-MAN/PETER PAR*, consistently rank highly across multiple centrality measures, indicating that they are not only well-connected but also strategically positioned to influence the network. Other heroes, like *THING/BENJAMIN J. GR* and *MR. FANTASTIC/REED R*, may have lower degree but maintain notable roles in terms of bridging or closeness.

2.3 Centrality Correlation Analysis

To investigate how centrality measures relate, we compute the correlation matrix among Degree, Betweenness, Closeness, and PageRank scores. Figure 2 shows the heatmap of correlations. The extremely high correlation (0.993) between Degree and PageRank suggests that well-connected heroes are also highly influential according to the PageRank algorithm. In contrast, the low correlation (0.357) between Betweenness and Closeness indicates that heroes who act as bridges do not necessarily reach all other nodes efficiently, highlighting that these metrics capture distinct roles. Overall, the results confirm that each centrality metric captures complementary aspects of hero importance in the network.

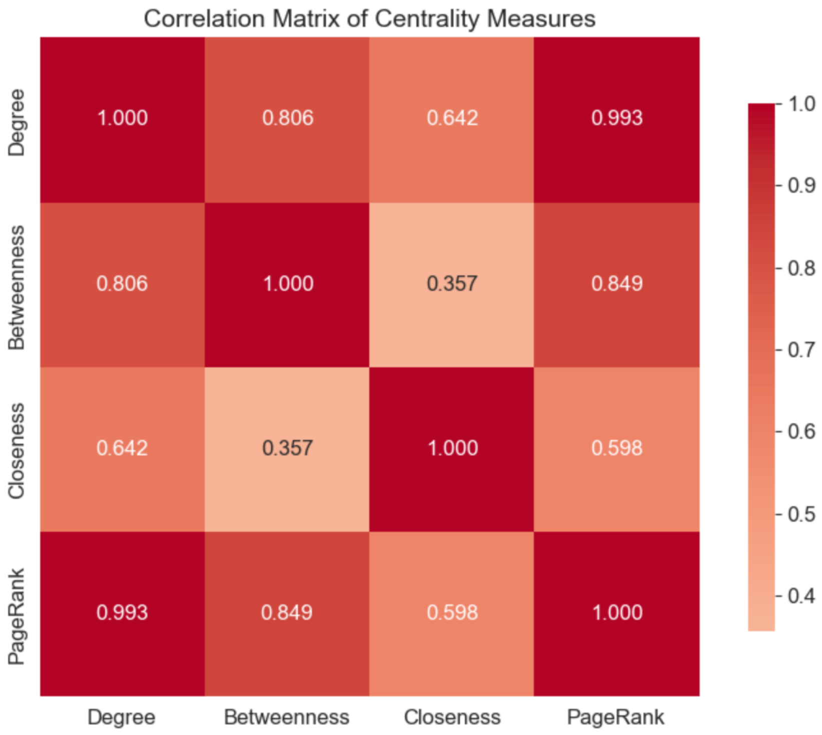

Figure 2: Heatmap showing correlations among Degree, Betweenness, Closeness, and PageRank centrality measures.

2.4 Comprehensive Hero Ranking

To assess overall hero importance, we aggregate rankings from all four centrality measures. Each hero's rank for Degree, Betweenness, Closeness, and PageRank is averaged to obtain a composite score. Table 5 lists the top 5 heroes based on this comprehensive ranking. Rank order is shown first for compactness, followed by the hero name, average rank, and individual centrality ranks.

Table 5: Top 5 Overall Central Heroes by Average Rank

Rank	Hero	Avg. Rank	Degree	Betweenness	Closeness	PageRank
1	SPIDER-MAN/PETER PAR	0.5	2	1	2	1
2	CAPTAIN AMERICA	0.5	1	2	1	2
3	IRON MAN/TONY STARK	2.2	3	4	3	3
4	THING/BENJAMIN J. GR	4.8	4	10	4	5
5	WOLVERINE/LOGAN	5.0	6	8	6	4

This analysis highlights the most influential and strategically important heroes in the Marvel Universe network, integrating multiple centrality perspectives for a comprehensive understanding.

3 Community Detection in the Marvel Heroes Network

To uncover groups of closely connected heroes within the Marvel Universe network, we apply the **Louvain algorithm**. This algorithm identifies communities by optimizing the *modularity* of the network, allowing us to detect clusters of heroes that frequently co-appear or form teams.

3.1 Louvain Algorithm Overview

The Louvain algorithm operates in two iterative phases:

1. **Local Modularity Optimization:** Each node is initially assigned to its own community. Nodes are then iteratively moved to neighboring communities if the move increases the modularity of the network.

2. **Community Aggregation:** Once local modularity cannot be further improved, nodes in the same community are aggregated into a super-node, producing a reduced network. The process then repeats on this smaller network.

The procedure continues until modularity converges. The Louvain method is particularly suitable for large networks like the Marvel hero network due to its efficiency and its ability to reveal meaningful hierarchical community structures.

3.2 Community Detection Results

Applying the Louvain algorithm, we detected 28 communities with a modularity score of 0.4213, indicating a moderately strong community structure. Table 6 lists the top 10 largest communities by size.

Table 6: Top 10 Largest Communities in the Marvel Hero Network

Rank	Community ID	Number of Heroes
1	0	1,376
2	4	1,331
3	21	897
4	8	619
5	3	534
6	6	382
7	2	240
8	5	240
9	9	220
10	7	158

3.3 Community Structure Analysis

Using the Louvain algorithm, we detect communities in the Marvel hero network to uncover groups of closely connected characters. Communities often correspond to teams, recurring story arcs, or tightly knit clusters of heroes. For each of the largest communities, we identify central heroes based on *PageRank* scores, which highlight those with the greatest influence within their groups.

Table 7 lists the top 3 heroes in the five largest communities. These hubs serve as key connectors and often anchor the narrative of their respective groups.

We further examine edges connecting heroes across communities to reveal intergroup collaborations and cross-over events. The strongest inter-community connections occur between the largest communities, reflecting frequent interactions among central characters. Table 8 presents the top 3 inter-community links as illustrative examples.

These results indicate that central heroes not only dominate within their own communities but also facilitate connections across groups, underpinning the overall interconnectedness of the Marvel Universe network.

Table 7: Key Heroes in the Largest Communities (by PageRank)

Community	Hero	PageRank
0	CAPTAIN AMERICA	0.00512
0	IRON MAN/TONY STARK	0.00411
0	SCARLET WITCH/WANDA	0.00321
4	WOLVERINE/LOGAN	0.00387
4	BEAST/HENRY &HANK& P	0.00318
4	CYCLOPS/SCOTT SUMMER	0.00259
21	SPIDER-MAN/PETER PAR	0.00532
21	JAMESON, J. JONAH	0.00265
21	WATSON-PARKER, MARY	0.00251
8	DR. STRANGE/STEPHEN	0.00304
8	SHE-HULK/JENNIFER WA	0.00261
8	SILVER SURFER/NORRIN	0.00190
3	THING/BENJAMIN J. GR	0.00367
3	MR. FANTASTIC/REED R	0.00357
3	HUMAN TORCH/JOHNNY S	0.00350

Table 8: Selected Inter-Community Connections

Community 1	Community 2	Number of Edges
0	4	8,128
0	8	7,602
0	21	4,954

3.4 Network Visualization of Communities

To provide an intuitive understanding of the Marvel hero network, we visualize the network highlighting community structures and node influence. In the visualization:

- **Node colors** indicate the detected communities, highlighting groups of closely connected heroes.
- **Node sizes** are proportional to PageRank scores, reflecting the relative influence of each hero.
- **Edges** represent co-appearance relationships, with opacity reflecting connection density.

Figure 3 shows the resulting network. Dense clusters reveal tightly connected hero groups, while prominent nodes (larger in size) indicate central characters within and across communities. This visualization provides both analytical and visual insights into the network's structure.

Insights:

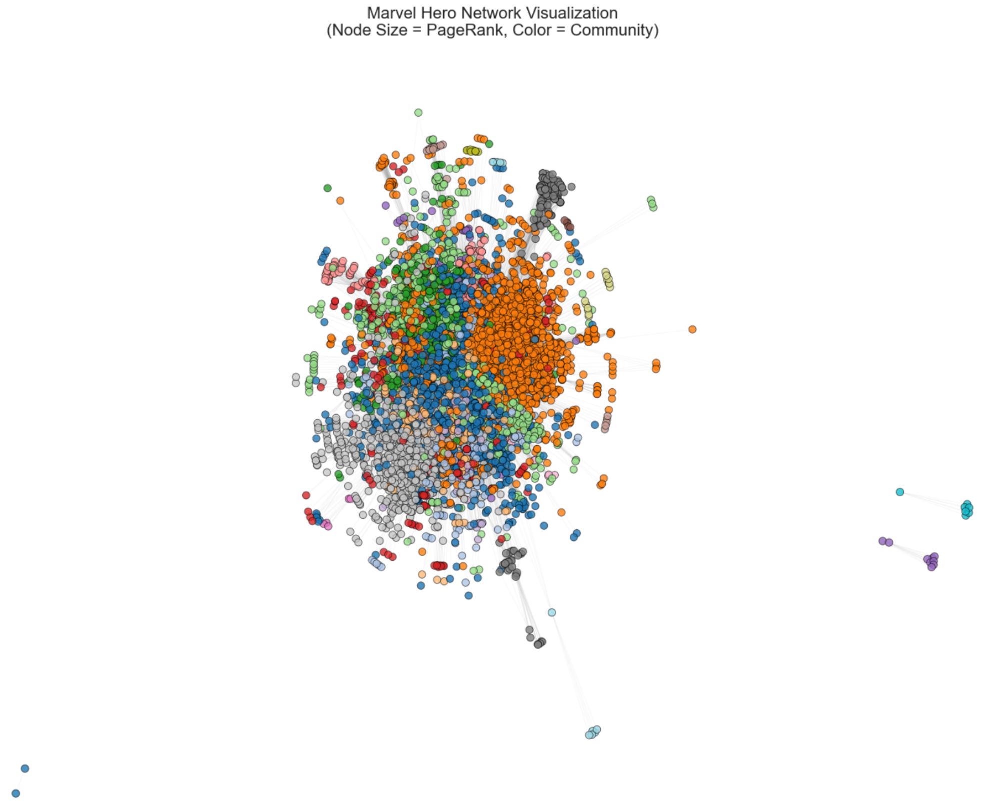

Figure 3: Marvel Hero Network Visualization. Node size corresponds to PageRank; node color indicates community membership.

- Node colors reveal distinct communities, corresponding to teams or recurring story arcs.
- Larger nodes highlight central heroes, consistent with high PageRank scores.
- Dense regions indicate tightly connected groups, emphasizing intra-community cohesion.

3.5 Discussion

The Louvain-based community detection reveals that the Marvel hero network is highly modular, with major communities corresponding to prominent teams or story arcs. Central heroes within each community act as hubs, while inter-community edges capture collaborations and cross-overs. These results provide a structural and functional perspective on hero interactions, complementing the centrality analysis.

4 Conclusion

In this study, we analyzed the Marvel Universe hero network using comprehensive network analysis techniques. Key findings include:

- **Hero Influence:** Centrality measures (Degree, Betweenness, Closeness, PageRank) consistently identify core heroes such as *CAPTAIN AMERICA*, *SPIDER-MAN*, and *IRON MAN* as highly influential within the network.
- **Community Structure:** Louvain-based community detection reveals 28 well-defined communities, with major communities corresponding to prominent teams or recurring story arcs.
- **Inter-Community Interactions:** Significant edges connecting different communities indicate frequent collaborations and cross-over events among central heroes.
- **Network Visualization:** Visualizations highlight dense clusters, community boundaries, and central hubs, providing intuitive insights into the network's modular structure and hero interactions.

Overall, the Marvel hero network exhibits a scale-free, modular structure with distinct communities anchored by central heroes. Future work could explore temporal dynamics of hero interactions, weighted networks considering co-appearance frequency, or influence propagation modeling to further understand the social structure of the Marvel Universe.