



Master of IT in Business
ISSS606 Social Analytics & Applications
Final Report

Marvel Cinematic Universe Network Analysis

Group 8

Chen Jin Chuan Brian
Clara Chua Kiah Hwii
Jin XiYuan
Lee Kern Choong

Submission Date

26 April 2019

Contents

1. Background and Project Objectives	3
2. Dataset	3
3. Analysis of MCU movies over time	3
3.1 Data Collection & Processing	3
Data Cleaning	4
Data Processing.....	4
3.2 Approach and Analysis.....	4
Approach.....	4
Visualisation & Findings	4
3.3 Key Results & Findings	6
4. Analysis of Comic Database	8
4.1 Data Processing.....	8
4.2 Approach, Analysis and Results	9
Approach.....	9
Step 1: Pick a Complete Trilogy.....	9
Step 2: Apply different measures to comic dataset to predict characters	10
Step 3: Compare Accuracy Rates	12
Step 4: Predict Characters in upcoming movies	12
4.3 Using Node2Vec.....	13
5. Conclusion.....	13
6. Contributions	14
7. References	14
Appendix 1: Predicted Characters of Upcoming Movies	15

1. Background and Project Objectives

The Marvel Cinematic Universe (MCU) is a shared universe that is centred on a series of superhero films, produced by Marvel Studios, based on characters in comic books published by Marvel Comics. MCU movies have grossed over \$17.9 billion worldwide, from Iron Man in 2008, to Captain Marvel in March 2019. The culmination of the Infinity Saga with '*Avengers: Endgame*' in April 2019 means we will say goodbye to the current batch of superheroes; whilst we look forward to different stories and continuation of some of the other superheroes' stories such as Black Panther and Doctor Strange.

If we consider the entirety of movies covering Marvel comic superheroes, they would add an additional 29 movies, from Blade in 1998 to X-Men in 2016 to Deadpool 2 in 2018. These movies also come from the Marvel Comics canon, but do not have as crossovers in the MCU shared universe, as they were produced by different movie studios.

Our project has two main aims:

- i. To explore the importance and influence of the characters in the MCU movies and how they have changed over the years
- ii. To predict what new characters will come with the next stage of the MCU movies, using Marvel Comics as the base. The predictions could help script-writers and producers as a first step to identify potential key characters for upcoming franchises, and explore their storylines further.

2. Dataset

To address the aims laid out above, we drew information from two datasets. The first dataset is the list of Marvel Superheroes (Comics) as seen in the comics from 1939 to 2018, which was obtained from Syntagmatic (Chang, 2019). The second was obtained from the Internet Movie Database (IMDb) (ninewheels0, 2019), consisting of the amount of time that each character spent on-screen in the movies they appeared in. The data was intended for use as follows:

Dataset	Use
Marvel Superheroes (Comics)	Create nodes and edges based on characters existing in the same comic
Character screen-time (Movies)	Create nodes and edges based on characters in the same movie

3. Analysis of MCU movies over time

3.1 Data Collection & Processing

For the movie dataset, the following fields were included in the table:

Variables	Data Format
Movie Title	Title of Movie
Year of Release	YYYY
Opening Weekend Box Office US\$	Value in millions
Final Box Office US\$	Value in billions
Phase	1, 2, 3

The MCU movies were split into the 3 phases as follows:

Phase 1 (2008 – 2012)	Phase 2 (2013 – 2015)	Phase 3 (2016 – April 2019)
Iron Man	Iron Man 3	Captain America: Civil War
The Incredible Hulk	Thor: The Dark World	Doctor Strange
Iron Man 2	Captain America: The Winter Soldier	Guardians of the Galaxy Vol 2
Thor	Guardians of the Galaxy	Spiderman – Homecoming

Captain America: The First Avenger	Avengers: Age of Ultron	Thor: Ragnarok
The Avengers	Ant-Man	Black Panther
		Avengers: Infinity War
		Ant-Man & The Wasp
		Captain Marvel

Non MCU movies were split into the 3 phases according to the movie release dates.

Data Cleaning

One of the main challenges in data cleaning was the need to match characters which were recorded differently. For example, Nick Fury has many roles, and is sometimes listed as any one of the following:

Director Nick Fury / Director Fury / Agent Nick Fury / Nicholas J. Fury / Nick Fury

As the dataset was not overly large, the team decided not to approach the data cleaning process using fuzzy matching or other character matching techniques – it was simply more efficient to quickly run through and standardize the names manually. For example, all the different versions of the character above were simply recoded as “Nick Fury”.

Data Processing

For the purposes of creating the network graph, our team defined the following parameters:

- Node: A node will be created for each superhero in the movies
- Edge: An edge will be created between two characters if they appear in the same movie together

To create a network graph from the above definitions, we needed to obtain the combinations of hero-hero interactions from the list of characters in each movie. This was done by using the “**combinations**” function from the **itertools** package.

3.2 Approach and Analysis

Approach

To measure importance and influence, we looked at different measures of centrality: degree centrality, closeness centrality and betweenness centrality. First we created the network graph for each phase of both the MCU and non-MCU movies using the **Networkx** and **pandas** packages. We then used the networkx functions to compute these measures and extracted them into a dataframe to see how the measures changed over the 3 phases.

Visualisation & Findings

We wanted to visualize how the network graph evolved over the three phases, and from the network graphs created, imported them into Gephi for better visualization¹.

We used the Fructerman-Reingold algorithm to visualise the network graph, so nodes with high centrality are in the centre, while nodes with lower centrality are at the outer edge of the circle. We colour-coded the nodes in descending order of centrality (red being the most central, green the least central). The colour splits were made on 25th percentiles. The evolution of the network can be seen in Figure 1 below.

¹ See ‘MCU graphs.gephi’ for detailed visualisation.

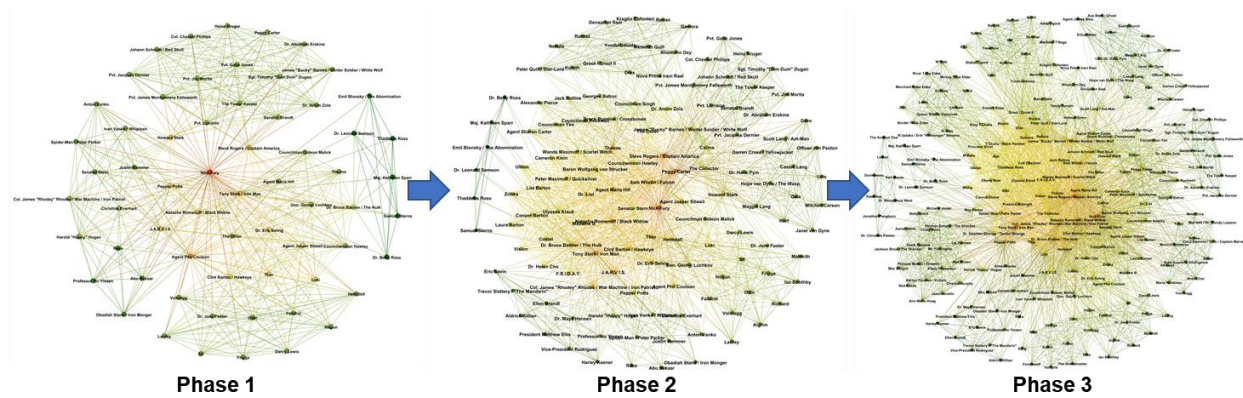


Figure 1: Evolution of MCU movie network graph

The network graph is fairly small with less character interactions in Phase 1. As expected, the graph grows more complex with each phase, and many of the central characters have many more edges with each other in the middle (core Avengers characters), while maintaining edges with their respective movie storyline characters on the outside.

We also plotted the same 3 phases, including the non-MCU movies (Figure 2). The Avengers are the most interconnected section of the graph (not surprising as it was done by design). However, we also see that the X-Men are also fairly connected, due to characters such as Professor Charles Xavier, and Wolverine, who had 3 standalone movies (X-Men Origins: Wolverine, The Wolverine, Logan). Both X-Men super-heroes appear in the top 10 central characters.

It is interesting to note that even though Marvel Studios sold off the rights to characters from X-Men, there is one sole connector between the Avengers franchise and the X-men movies - Quicksilver. He appeared in 2 Avengers movies (Captain America: The Winter Soldier, Avengers: Age of Ultron), and 2 X-Men movies (X-Men: Apocalypse & X-Men: Days of Future Past). This is borne out by the Betweenness centrality score – Quicksilver has the highest betweenness score (0.293) which is almost 3-4 times higher than the next highest betweenness score.

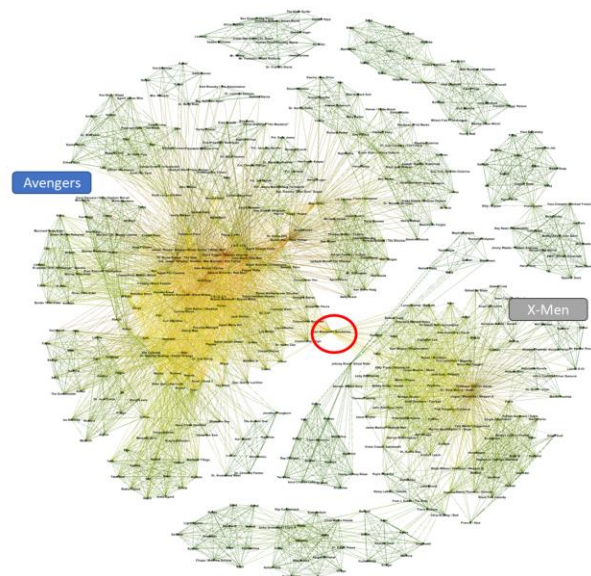


Figure 2: Network graph of all Marvel related movies (Phase 3)

3.3 Key Results & Findings

Sorting characters by their centrality scores in descending order, we get similar characters returned in the top ranks regardless of centrality measure used. The top 15 characters in Phase 3 of the movies by measure are as follows:

Rank	Top Centrality	Top Closeness	Top Betweenness
1	Steve Rogers / Captain America	Steve Rogers / Captain America	Steve Rogers / Captain America
2	Nick Fury	Tony Stark / Iron Man	Tony Stark / Iron Man
3	Tony Stark / Iron Man	Natasha Romanoff / Black Widow	James "Bucky" Barnes / Winter Soldier / White Wolf
4	Natasha Romanoff / Black Widow	Nick Fury	Nick Fury
5	Dr. Bruce Banner / The Hulk	James "Bucky" Barnes / Winter Soldier / White Wolf	Dr. Bruce Banner / The Hulk
6	James "Bucky" Barnes / Winter Soldier / White Wolf	Col. James "Rhodey" Rhodes / War Machine / Iron Patriot	Thor
7	Col. James "Rhodey" Rhodes / War Machine / Iron Patriot	Dr. Bruce Banner / The Hulk	Natasha Romanoff / Black Widow
8	Thor	Sam Wilson / Falcon	Col. James "Rhodey" Rhodes / War Machine / Iron Patriot
9	Sam Wilson / Falcon	Thor	Pepper Potts
10	Pepper Potts	Pepper Potts	Scott Lang / Ant-Man
11	Heimdall	Spider-Man / Peter Parker	Sam Wilson / Falcon
12	Spider-Man / Peter Parker	Wanda Maximoff / Scarlet Witch	Dr. Stephen Strange / Doctor Strange
13	Wanda Maximoff / Scarlet Witch	Heimdall	Spider-Man / Peter Parker
14	Loki	F.R.I.D.A.Y.	Heimdall
15	Thanos	Vision	Drax

Figure 3 shows the movement of the top 15 central characters (by degree centrality) from Phase 1 to Phase 3. It shows that the most central characters included most of the core Avengers members (excluding Hawkeye): Captain America, Iron Man, Black Widow, The Hulk, and Thor.

Nick Fury, Bucky Barnes and James Rhodes, who are key supporting characters were also unsurprisingly central to the MCU movies. Some characters such as Sam Wilson / Falcon, or Wanda Maximoff, only appeared from Phase 2 onwards.

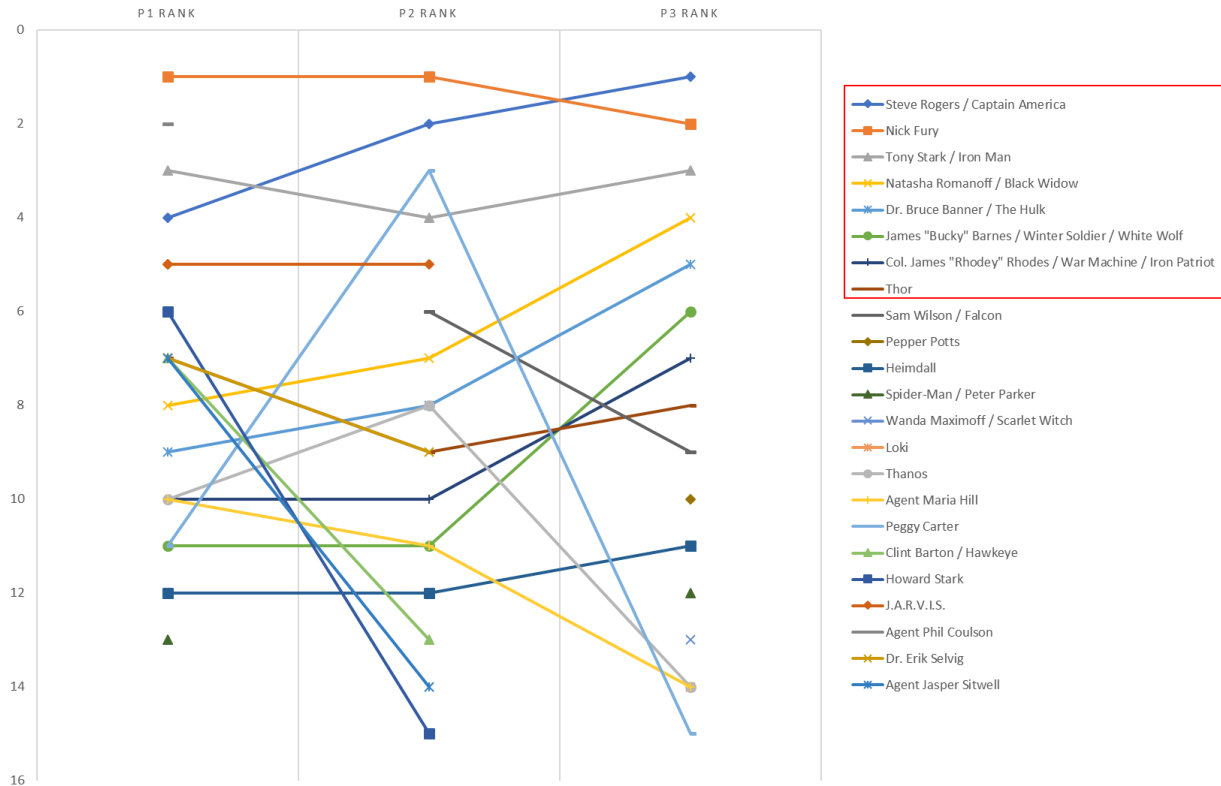


Figure 3: Movement of Ranking of Top 15 Characters by Degree Centrality

Captain America rose in ranks from 4th in Phase 1, to 2nd in Phase 2, to be the most central character by the end of Phase 3. This is no surprise as Captain America was the last movie before the first cross-over movie Avengers appeared in Phase 1. Although Thor only had 1 movie in Phase 1, Captain America had more character interactions than Thor in the movie, and therefore this accounted for his centrality score. This was also the case in Phase 2. Captain America: Civil War (Phase 3 movie) had one of the highest key character interactions in the movie franchise, outside of the Avengers cross-over movies (Avengers, Avengers Age of Ultron, Avengers Infinity War). It is therefore no surprise that Captain America would be the most central figure in the entire franchise by Phase 3. This also correlates with the rise in box office takings – from \$0.37 billion to \$0.81 billion to \$1.13 billion (total of \$2.21 billion).

Nick Fury started out as the most central character in Phase 1 and 2, but dropped to the second most central character in Phase 3. This is interesting as Nick Fury was not a major character in most of the movies in Phase 1, but he had appeared as a cameo in almost all the movies in Phase 1. He appeared in less movies in Phase 2 and 3, but his screen-time increased. However, his centrality rank is not surprising when we consider the storyline and comics database – it matches with the importance and influence of the character in connecting many of the superheroes.

4. Analysis of Comic Database

4.1 Data Processing

The comic dataset provides a list of comic character relationships, based on number of common occurrences across comics. Duplicates are expected as character-pairs can appear in multiple comics.

	hero1	hero2
0	24-HOUR MAN/EMMANUEL	FROST, CARMILLA
1	24-HOUR MAN/EMMANUEL	KILLRAVEN/JONATHAN R
2	24-HOUR MAN/EMMANUEL	M'SHULLA
3	3-D MAN/CHARLES CHAN	ANGEL/WARREN KENNETH
4	3-D MAN/CHARLES CHAN	ANT-MAN II/SCOTT HAR
5	3-D MAN/CHARLES CHAN	AURORA/JEANNE-MARIE
6	3-D MAN/CHARLES CHAN	BLACK PANTHER/T'CHAL

Figure 4: List of comic superhero interactions

The dataset was very large, spanning about half a million rows. In terms of visualization, it was not trivial to manually pick out observations and patterns as the graph is too large and dense.

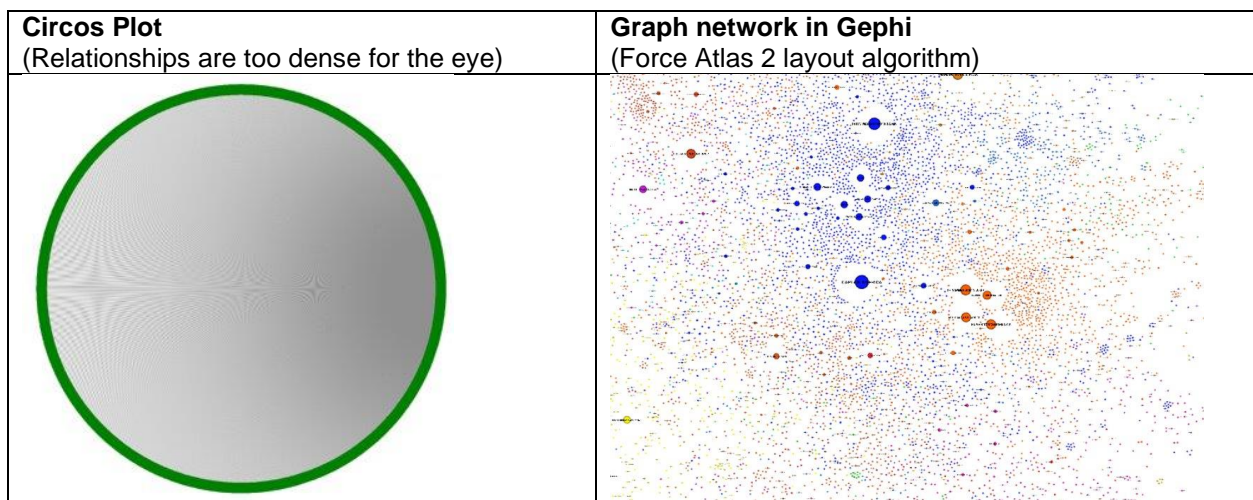


Figure 5: Visualisation of Comic Dataset


We analyzed the network to identify separate components and found 4 separate graphs to exist. However, we excluded the last 3 components from analysis as they were determined to be either erroneous or insignificant to the final results. There were a total of 6428 unique characters found in the dataset.

Component #	Remarks
Component 1	Size of 6408 containing most of all characters
Component 2	Contains the following characters: {'AMAZO-MAXI-WOMAN/', 'DARLEGUNG, GEN.', 'MANT/ERNEST', 'MISS THING/MARY', 'PANTHER CUB/', 'STERLING', 'SWORDSMAN IV/'}
Component 3	Contains the following characters: {'ASHER, CARL', 'ASHER, DONNA', 'ASHER, MICHAEL', 'FAGIN', 'HOFFMAN', 'LUDLUM, ROSS', 'NILES, SEN. CATHERIN', 'ORWELL', 'OSWALD'}
Component 4	Contains the following characters: {'MASTER OF VENGEANCE', 'STEEL SPIDER/OLLIE O'}

Finally, we decided to collapse the flat list into a weighted list. This would not just simplify the analysis but would also make computation more efficient.

However, in terms of pairing, we found that character-pairs appear in the list regardless of order. For example, in Figure 6 below, you would want to also count A--B and B--A as the same relationship, since we do not consider the direction of the relationship. Assuming this is the full list, D--E and B--C would have weights of 2, and A--B a weight of 1.

hero1_hero2					
A--B					
B--C					
D--E					
C--B					
E--D					



	hero1	hero2	weight
213085	PATRIOT/JEFF MACE	PATRIOT/JEFF MACE	2550
186598	MISS AMERICA/MADELIN	PATRIOT/JEFF MACE	1894
213083	PATRIOT/JEFF MACE	MISS AMERICA/MADELIN	1894
186594	MISS AMERICA/MADELIN	MISS AMERICA/MADELIN	1344
292584	THING/BENJAMIN J. GR	HUMAN TORCH/JOHNNY S	744
128312	HUMAN TORCH/JOHNNY S	THING/BENJAMIN J. GR	744
193246	MR. FANTASTIC/REED R	HUMAN TORCH/JOHNNY S	713

Figure 6: Character-Pair weighting

#	Description
1	We prepare another set of the same data with columns switched to cater for the other relationship pair direction
2	We load each set of data and group by their rows and count the duplicates as weights
3	We combine both sets of data and again group by their weights
4	Leveraging on the fact that Networkx undirected edgelist method ignores same relationship pairs of opposite direction, we obtain a final graph with weighted edges

4.2 Approach, Analysis and Results

Approach

The second objective is to predict new characters in the next stage of the MCU movies using Marvel Comics as the baseline. We experimented and developed a few methods to fulfill this objective. The prediction process can be generalized into four steps:



Step 1: Pick a Complete Trilogy

We pick a complete trilogy to serve as both training and test dataset. We identify the unique characters in the trilogy to use for verification of our predictions. We take the number of unique characters (34) as the number of characters to predict for across the three movies.

There are 2 complete trilogies in the entire MCU movies to date – Iron Man and Thor. In this case we chose the Thor trilogy, which has 34 main characters (excluding Thor):

Main Characters in the Thor Trilogy

Agent Phil Coulson	Sif	Loki	Algrim	Korg
Agent Jasper Sitwell	Volstagg	Nick Fury	Ian Boothby	Surtur
The Grandmaster	Fandral	Darcy Lewis	The Collector	Topaz
Dr. Stephen Strange	Hogun	Dr. Jane Foster	Richard	Dr. Erik Selvig
Dr. Bruce Banner /Hulk	Frigga	Heimdall	Carina	Fenriswolf
Clint Barton / Hawkeye	Laufey	Malekith	Odin	Miek
Natasha Romanoff /Black Widow	Valkyrie	Hela	Skurge	

Step 2: Apply different measures to comic dataset to predict characters

(i) Top weighted neighbors

We created a weighted network graph, and ranked all neighbours of Thor by weight. We selected the top 34 neighbours. The ones who actually appear in the movies are in bold. They are:

Captain America	386	Hogun	186	Black Panther/T'Challa	103
Iron Man/Tony Stark	344	Loki	182	Quicksilver/Pietro Maximoff	99
Odin	266	Jarvis, Edwin	160	Spider-Man/Peter Parker	95
Vision	255	Mr.Fantastic/Reed Richards	129	Beast/Henry HankMcCooy	90
ScarletWitch/Wanda Maximoff	254	Thing/Benjamin J. Grimm	126	Captain Marvel II/Monica Rambeau	84
Wasp/Janet Van Dyne	238	WonderMan/Simon Williams	125	Enchantress/Amora	78
Hawk	210	HumanTorch/Johnny Storm	124	Karnilla	76
Balder	209	She-Hulk/Jennifer Walters	123	Hulk/Dr. Robert Bruce Banner	72
Sif	204	Hercules	114	Vizier	72
Ant-Man/Dr. Henry J. Pym	189	Kincaid, Dr. Jane Foster	112	Sub-Mariner/NamorMackenzie	62
Volstagg	187	Heimdall	111		
Fandral	186	Invisible Woman/Sue Storm	111		

(ii) Degree centrality

We created a weighted network graph, selected neighbours of Thor, and ranked them by degree centrality. We selected the top 34 neighbours. The ones who actually appear in the movies are in italics. They are:

Captain America	0.2970	Wasp/Janet Van Dyne	0.1698	Hercules	0.1542
Spider-Man/Peter Parker	0.2703	Ant-Man/Dr. Henry J. Pym	0.1684	Jarvis, Edwin	0.1535
Iron Man/Tony Stark	0.2369	Cyclops/Scott Summers	0.1682	Sub-Mariner/NamorMackenzie	0.1524
Thing/Benjamin J. Grimm	0.2204	Angel/Warren Kenneth Worthington III	0.1670	Daredevil/Matt Murdock	0.1505
Mr. Fantastic/Reed Richards	0.2146	Storm/Ororo Munroe	0.1668	Iceman/Robert Bobby Drake	0.1471
Wolverine/Logan	0.2134	She-Hulk/Jennifer Walters	0.1667	Black Widow / Natasha Romanoff	0.1435

Human Torch/Johnny Storm	0.2118	<i>Dr. Strange/Stephen Strange</i>	<i>0.1661</i>	<i>Fury, Col. Nicholas</i>	<i>0.1435</i>
Scarlet Witch/Wanda Maximoff	0.2062	<i>Hulk/Dr. Robert Bruce Banner</i>	<i>0.1642</i>	Jameson, J. Jonah	0.1432
Beast/Henry HankMccoy	0.1972	Wonder Man/Simon Williams	0.1608	Quicksilver/Pietro Maximoff	0.1362
Vision	0.1932	Professor X/Charles Xavier	0.1606	Nightcrawler/Kurt Wagner	0.1334
Invisible Woman/Sue Storm	0.1924	Colossus II/Peter Rasputin	0.1595		
Hawk	0.1829	Marvel Girl/Jean Grey	0.1564		

(iii) Eigenvector centrality

We created a weighted network graph, selected neighbours of Thor, and ranked them by eigenvector centrality. We selected the top 34 neighbours. The ones who actually appear in the movies are in italics. They are:

Captain America	0.2912	Beast / Henry Hank Mccoy	0.1572	Angel / Warren Kenneth Worthington III	0.1043
Thing/Benjamin J. Gr	0.2328	Wonder Man/Simon Williams	0.1434	Iceman / Robert Bobby Drake	0.1025
Human Torch/Johnny Storm	0.2287	Cyclops/Scott Summers	0.1357	<i>Hulk / Dr. Robert Bruce Banner</i>	<i>0.1000</i>
Iron Man/Tony Stark	0.2251	Wolverine/Logan	0.1315	Hercules	0.0968
Mr. Fantastic/Reed Richards	0.2248	Spider-Man/Peter Parker	0.1290	Patriot / Jeff Mace	0.0951
Invisible Woman/Sue Storm	0.2163	Storm /Ororo Munroe	0.1226	Sub-Mariner / Namor Mackenzie	0.0891
Scarlet Witch/Wanda Maximoff	0.2126	Professor X/Charles Xavier	0.1165	<i>Black Widow / Natasha Romanoff</i>	<i>0.0852</i>
Vision	0.2096	She-Hulk/Jennifer Walters	0.1147	Rogue	0.0839
Thor/Dr. Donald Blake	0.1995	Jarvis, Edwin	0.1092	Nightcrawler/Kurt Wagner	0.0777
Wasp/Janet Van Dyne	0.1971	Quicksilver / Pietro Maximoff	0.1075	Black Panther/T'Challa	0.0771
Hawk	0.1770	Colossus II / Peter Rasputin	0.1065		
Ant-Man/Dr. Henry J. Pym	0.1711	Marvel Girl /Jean Grey	0.1051		

(iv) Similarity score

We created a node vector model using the **node2vec** package and ranked all neighbours of Thor by similarity. We selected the top 34 neighbours. The ones who actually appear in the movies are in italics. They are:

<i>Executioner II/Skurge</i>	<i>0.9287</i>	<i>Fandral</i>	<i>0.8553</i>	Karnilla	0.8139
<i>Odin</i>	<i>0.9078</i>	Utgard-Loki	0.8519	Neffethesk	0.8138
<i>Loki</i>	<i>0.8981</i>	<i>Volstagg</i>	<i>0.8473</i>	Pentigaar	0.8128
Enchantress/Amora	0.8898	<i>Tyr</i>	<i>0.8443</i>	Designate / Tarene	0.8058
<i>Sif</i>	<i>0.8771</i>	Hobbs, Harris	0.8375	Nichols, Lorna	0.7890
<i>Heimdall</i>	<i>0.8732</i>	Vizier	0.8371	Horus	0.7863
Balder	0.8708	<i>Frigga</i>	<i>0.8365</i>	Toothgnasher	0.7850

Surtur	0.8683	Volla	0.8284	Malekith / Malcolm Keith	0.7846
Kincaid, Dr. Jane Foster	0.8630	Destroyer III	0.8259	Case, Col. Preston	0.7821
Krista	0.8621	Seth II	0.8245	Kurse / Algrim	0.7819
Harokin	0.8598	Hildegarde	0.8231		
Hogun	0.8588	Lorelei II / Melodi	0.8186		

Step 3: Compare Accuracy Rates

We compared the accuracy rate of the four measures to determine which one is the best. Accuracy rate is calculated as the matched number of characters / predicted number of characters (in this case 34 unique characters in the movies) to get the table below:

	Top Weighted Neighbours	Degree Centrality	Eigenvector Centrality	Node2Vec Similarity
Predicted number	34	34	34	34
Matched number	9	4	2	15
Accuracy	26.5%	11.8%	5.9%	44.1%

From the result, the similarity measure has the best performance with 44.1% accuracy rate. This is because the Node2vec model clusters the nodes that are similar to Thor, which means they are in the same dimension.

Measure 1 (top weighted neighbours) also gives a good result (26.5%), because we get the characters connected to Thor directly and the top neighbours are the ones who appear most often with Thor in the comics. Thus, it is not surprising to find that they have a better chance of showing up in the movies.

In contrast, the degree centrality and eigenvector centrality measures are not performing well. This is not surprising as the degree centrality and eigenvector centrality also takes into account links to other characters, not only in relation to Thor.

Step 4: Predict Characters in upcoming movies

We apply our chosen measure (Node2Vec Similarity) to predict characters in 4 upcoming movies: (i) Black Widow (ii) Black Panther 2 (iii) Doctor Strange 2 and (iv) X-Men Origins: Gambit.

The following characters are the top 34 characters predicted for Black Widow. Characters in italics have already appeared in the franchise, which means there is a possibility for cross-overs. Characters highlighted in bold are the ones that our team predicts will have a high probability of appearing in the movie, typically due to their importance in the comic storyline (e.g. origin story, major foe / side-kick).

Black Widow

Character	Similarity	Character	Similarity	Character	Similarity
DEATHCRY	0.8272	MADAME MASQUE III	0.7346	TABULA RASA	0.7110
<i>ANT-MAN/DR. HENRY J. PYM</i>	<i>0.7781</i>	TUC	0.7337	STORM, CHILI	0.7105
HERCULES	0.7637	SWORDSMAN III/PHILIP JAVERT	0.7301	WATCHLORD	0.7067
<i>MANTIS</i>	<i>0.7598</i>	CARINA/CARINA WALTER	0.7300	<i>WASP/JANET VAN DYNE</i>	<i>0.7059</i>
IVAN PETROVITCH	0.7571	JOCASTA II	0.7289	PROCTOR	0.6990
NEUT	0.7556	<i>VISION</i>	<i>0.7264</i>	NELIT	0.6984
ZA'KEN	0.7348	T'KYLL ALABAR	0.7189		

The predicted characters for the other movies can be found in Appendix 1.

4.3 Using Node2Vec

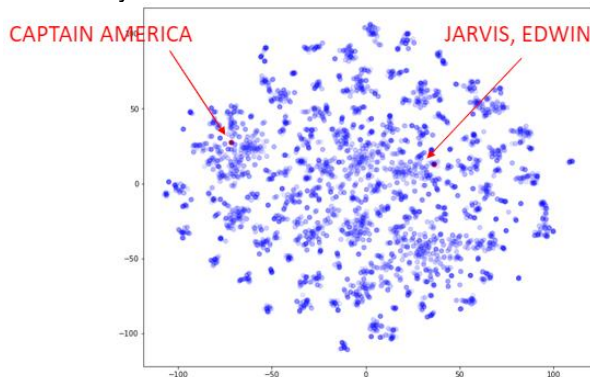
In view of the poor results from the weighted neighbours and centrality measures above, we wanted to explore if there were aspects of the graph relationship not captured by those measures alone. Node2vec was chosen as a method to explore these unexploited patterns, and boost prediction capability.

Node2vec makes use of the general principle behind the DeepWalk algorithms. By creating and taking random walks, it allows building of the model using a high-dimension vector space. T-distributed Stochastic Neighbour Embedding (t-SNE) is then used to reduce the dimensionality of data (Maaten, 2019), (Hoare, 2019), allowing for better understanding of the distance and similarity between nodes.

We trained several models, modifying the number of dimensions, steps and random walks. The final model that gave the best accuracy consisted of 20 dimensions, 25 steps and 125 random walks per node, which is what was used to predict characters from the upcoming movies.

TSNE Plot High - Similarity

In the plot below, t-SNE allows us to see that Captain American and Edwin Jarvis are not in close proximity; this makes sense as Edwin Jarvis is a character from Iron Man in separate community.



However, from the similarity ranking list we can understand that these two characters are quite similar.

```
model.wv.most_similar('CAPTAIN AMERICA', topn=15)
```

```
[('JARVIS, EDWIN ', 0.8652085065841675),  
( 'ANT-MAN/DR. HENRY J.', 0.8592347502708435),  
( 'SIKORSKY, RAYMOND', 0.8548966646194458),  
( 'FALCON/SAM WILSON', 0.8435205221176147),  
( 'WASP/JANET VAN DYNE ', 0.8273553848266602),  
( 'MCELROY, JAMES', 0.8218629956245422),  
( 'VISION ', 0.815720796585083),  
( 'RAVONNA LEXUS RENSLA', 0.7826374769210815),  
( 'RED SKULL/JOHANN SCH', 0.774689257144928),  
( 'DEMOLITION MAN/DENNI', 0.7746118307113647),  
( 'CARTER, PEGGY', 0.7686989307403564),  
( 'STANKOWICZ, FABIAN', 0.7686619162559509),  
( 'KORVAC, MICHAEL', 0.7571008801460266),  
( 'HAWK', 0.7521514892578125),  
( 'SENSATIONAL HYDRA/A ', 0.7510486245155334)]
```

5. Conclusion

In terms of limitations of our project, one of our original aims was to potentially identify new superhero franchises (i.e. main superhero) from the comic dataset. However, as we were not able to do any pattern matching to infer new superheroes not found in the current Marvel Cinematic Universe.

In terms of future work, we could explore how to use Node2vec with weighted edges, and experiment with different parameters to allow us to capture more latent patterns in the graph (e.g. other number of dimensions, length of walks, etc). We would also include other MCU properties, such as the Marvel TV Series (e.g. Agents of SHIELD, Jessica Jones, Daredevil, Iron Fist, Luke Cage, The Defenders) as well as other MCU produced One-Shots.

Marvel Studios has shown that the shared multi-verse with its cross-overs, contribute to greater audience engagement and high box-office takings. Avengers: Endgame is expected to rake in US\$300m in US for the opening weekend alone. We can therefore use our project findings to identify key character interactions to find more interconnected storylines to plan such cross-overs for greater audience engagement and potentially revenues for the studios.

6. Contributions

Team Member	Contribution
Brian	<ul style="list-style-type: none">• Data processing, analysis of Marvel Comic dataset• Prototype of node2vec model and model visualization• Presentation slides, Report writing
Clara	<ul style="list-style-type: none">• Collated Marvel Movie dataset• Data processing and analysis of Marvel movie dataset• Presentation slides, Report writing
Kern Choong	<ul style="list-style-type: none">• Experimentation and iteration of node2vec models for Marvel Comic dataset• Presentation slides, Report writing
Xi Yuan	<ul style="list-style-type: none">• Data processing, analysis of Marvel Comic dataset• Development of validation and scoring based on models• Presentation slides, Report writing

7. References

Chang, K. (1 February, 2019). *Marvel Universe Social Graph*. Retrieved from <http://syntagmatic.github.io/exposedata/marvel/>

Hoare, J. (1 February, 2019). Retrieved from <https://www.displayr.com/using-t-sne-to-visualize-data-before-prediction/>

Maaten, L. v. (1 February, 2019). Retrieved from <https://lvdmaaten.github.io/tsne/>

ninewheels0. (1 February, 2019). *MCU Movies Screen Time Breakdown*. Retrieved from IMDB: <https://www.imdb.com/list/ls066620113/>

Appendix 1: Predicted Characters of Upcoming Movies

Black Panther 2

Character	Similarity
LYNNE, MONICA	0.8212
NECRODAMUS	0.7893
ADAMS, NICOLE NIKKI	0.7757
SWORDSMAN/JACQUES DUQUESNE	0.7754
ROSS, EVERETT KENNET	0.7619
WHITE WOLF/HUNTER	0.7537
<i>KILLMONGER, ERIC/N'JADAKA</i>	<i>0.7512</i>
RED GUARDIAN II/ALEXEI SHOSTAKOV	0.7501
TAKU	0.7439
PRESTER JOHN	0.7430
STINGER II	0.7412
<i>ANT-MAN/DR. HENRY J. PYM</i>	<i>0.7402</i>
MASTER PANDEMONIUM / MARTIN PRESTON	0.7317
BLACK KNIGHT III/EOB	0.7280
GRIM REAPER/ERIC WILLIAMS	0.7280
<i>VISION</i>	<i>0.7268</i>
ZURI	0.7261
<i>WASP/JANET VAN DYNE</i>	<i>0.7253</i>
AMENHOTEP	0.7234
LLOIGOROTH	0.7214

Gambit

Character	Similarity
PSYLOCKE / ELISABETH BRADDOCK	0.8846
CHROME	0.8843
<i>ROGUE</i>	0.8745
DELGADO	0.8636
MACTAGGART, JOE	0.8599
HAZARD/CARTER RYKING	0.8536
<i>RASPUTIN, MIKHAIL</i>	0.8491
EJULP	0.8478
<i>STORM/ORORO MUNROE</i>	0.8434
TRION	0.8421
MARROW/SARAH	0.8420
MR. SINISTER/NATHAN	0.8387
SISTER MARIA	0.8299

Character	Similarity
BRAIN CELL	0.8242
WANDERER	0.8212
MEME	0.8184
WOLVERINE/LOGAN	0.8163
RYKING, ALEXANDER	0.8131
PAM	0.8126
AVATAR	0.8126

Doctor Strange 2

Character	Similarity
CLEA	0.967789
WONG	0.959589
CHANG, IMEI	0.904088
ANCIENT ONE	0.899713
RINTRAH	0.890948
WOLFE, SARA	0.885755
BLESSING, MORGANA	0.859316
DORMAMMU	0.855099
BARON MORDO / KARL MORDO	0.834619
BLACK, CYRUS	0.828678
ASMODEUS	0.825324
AGAMOTTO	0.820484
AZRAEL	0.820051
SHADOW QUEEN/SHIALMAR	0.817132
UMAR	0.816197
AGGAMON	0.816153
DR. STRANGER YET	0.814923
INTERLOPER	0.813463
SHUMA-GORATH	0.81115
APPALLA	0.810538