From comic to comments: Social network Analysis in Marvel Universe

MACS 36000: Computational Methods Using Online Social Media Data

Course Instructor: Prof. Wang

Student Author: Linhui Wu

1. Introduction and Background

Marvel Universe network (MU) is composed by actors (nodes)- the Marvel characters, and by the network where Marvel characters build undirected social ties if they co-appear in the same comic during the past sixty decades (Alberich et al. 2002). Gleiser (2007) find that the growth of Marvel universe is not like collaboration graph, but more like a social media network in internet which has exponential power. Using a weighted network model based on the frequencies of the co-appearance in a comic publication, Gleiser (2007) find that only small proportions of characters have strong interactions with others while the majority characters have weak interactions with others. Furthermore, the community structure reflects those characters who are villains in the comic are disconnected with small communities (Gleiser 2007). Previous study about Marvel Universe simply focuses on the network, few combine the opinions from fans on social media for analyze the relationship between characters and Marvel network. The development of social media platforms and the frequent participation of Marvel fans in discussion platforms such as Reddit, provides abundant textual resource which reflect how fans think about the characters of Marvel Universe, and most importantly, how these comments corresponding to each character work as additional features for us to understand the community structure and nodes in Marvel Universe network.

In literature of social network analysis, a field of scholars focuses on the social network in artificial work such as literature and movies other than Marvel Universe. Because these artificial works have relatively clear relationship between characters and the network shares certain features with realistic complex network. Li et al (2019) uses complex theory to analyze the popular movies Lord of the Rings and Harry Potter. By simulating the growth of complex network in these two movies, the authors find that the evolution of network follows the preferential attachment (if the character has more degrees, the character will receive even more links) (Li et al 2019). And some special plots in the story also contributes to the development of the network (Li et al 2019). Yet whether such preferential attachment theory works for the characters in the Marvel Universe remains unknown.

Mourchid et al (2018) constructs a multilayer network model to analyze the narrative of two popular movies, considering multiple components within a movie, that are people, location, and other semantic elements, in contrast to previous studies which only focus on a single layer network model. In their paper, they employ series Avengers' script and extract four main semantic components: the characters, the location of episodes, the subjects, and time (in response to who, when, where, and what four main elements in narrative) for network modeling (Mourchid et al 2018). There are three facets which represent the network of characters, the network of keywords in conversation and the network of location: for the character network, if two characters appear in the same conversation, there would an undirected and unweighted tie between these two characters; for the establishment of the location network, the authors refer to the succession of two movie scenes; Keyword links with each other if they co-appear in the same conversation; There is interaction (link) between each facet if all components happen in the same scene (Mourchid et al 2018). Such multi-facet network model does not only reveal information about the relationship between characters in the movie, but also reveals the interactions between characters and the interactions with narrative settings. Mourchid et al (2018)'s paper inspires to systematically analyze a literature or movie rather than focusing on simply characters.

2. Research question

This research project intends to investigate 1) Does characters with higher degrees have higher accuracy in link prediction? 2) What are the similarities and differences between the communities detected from Marvel Universe network (MU) and the communities from comments associated with each character? The significance of the research questions provides a comparative analysis of how the social network construction in this artificial world, especially the appearance of social ties, and encourages Marvel creators and writers of later generations for content creation through analyzing existing Marvel Universe Network and social media comments to continue this magic world. Furthermore, understanding the how characters are distributed and are connect with each other in the Marvel Universe, one of keys of Marvel's success, provides inspiration for comic creators, movie script writer and content creators for creative work beyond Marvel Universe.

3. Data

The data in my research project can be divided into two parts - Marvel Universe network and Reddit comments corresponding to characters.

3.1 Marvel Universe network

Marvel Universe network is retrieved from the Kaggle platform. The Marvel Universe network contains nodes and edges between nodes. Each node represents a character in any comic series of Marvel Universe during the past decades. Each edge between character represents a co-appearance of two character in the same comic series. The first dataset contains 6246 characters and 167219 undirected edges. The original dataset can be downloaded from Kaggle[1].

Figure two: Degree distribution of the Marvel Universe Network

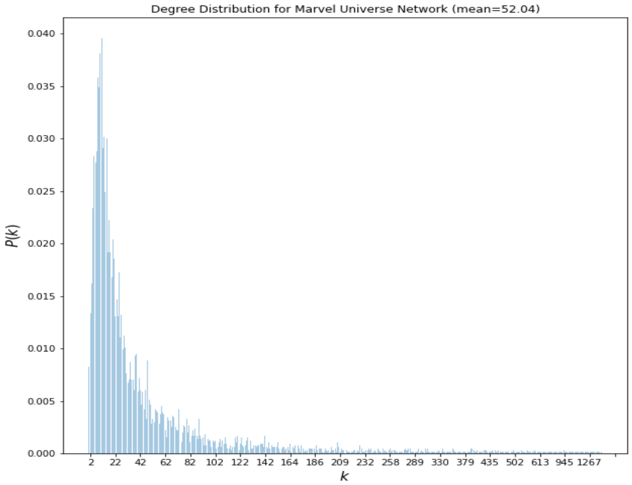


Figure two shows the degree distribution Marvel Universe, which has heavy-tail feature. Long-tail feature of degree distribution is a character of the complex network in social and natural world (Eom and Jo 2014). Thus, although Marvel Universe network is an artificial network, it could be regarded as a mimicry of our real social network.

3.2 Reddit comment data

Comment data is scraped from the Reddit platform. Based on the degree of each node in the Marvel Universe network, I select 100 characters who have the most degrees in the Marvel Universe. Then, I use Reddit Pushshift API to scrape comments which contain the name of the corresponding character from the subreddit "AskforScience". The reason why to select this subreddit instead of others is that people are discussing comic narratives and posting reflections on this subreddit, while other subreddits contain much more advertisements. I set the maximum

[1] https://www.kaggle.com/datasets/csanhueza/the-marvel-universe-social-network?datasetId=741

number of comments for each character as 2000. That means, for each character whose comments are fewer than 2000, the API would return all comments of that character.

This dataset contains 64435 pieces of comment for a total of 87 characters (the comments of 13 characters are not found in the subreddit). The second dataset is saved into a json file and can be downloaded from this drive[2], where the first dataset is also included.

Table one: description of comment data from reddit

Corpus (total pieces of comment)	64435
Tokenized word per comment	121.84
Vocabular size	7850746
Normalized words size	3552351
average normalized tokens for each comment	55.13

[2] https://drive.google.com/drive/folders/1HBE0mzvWdBSthaOmc2n4nLgbq8pKqz1A?usp=sharing

Figure one: subgraph of TOP 100 hero in Marvel network Universe

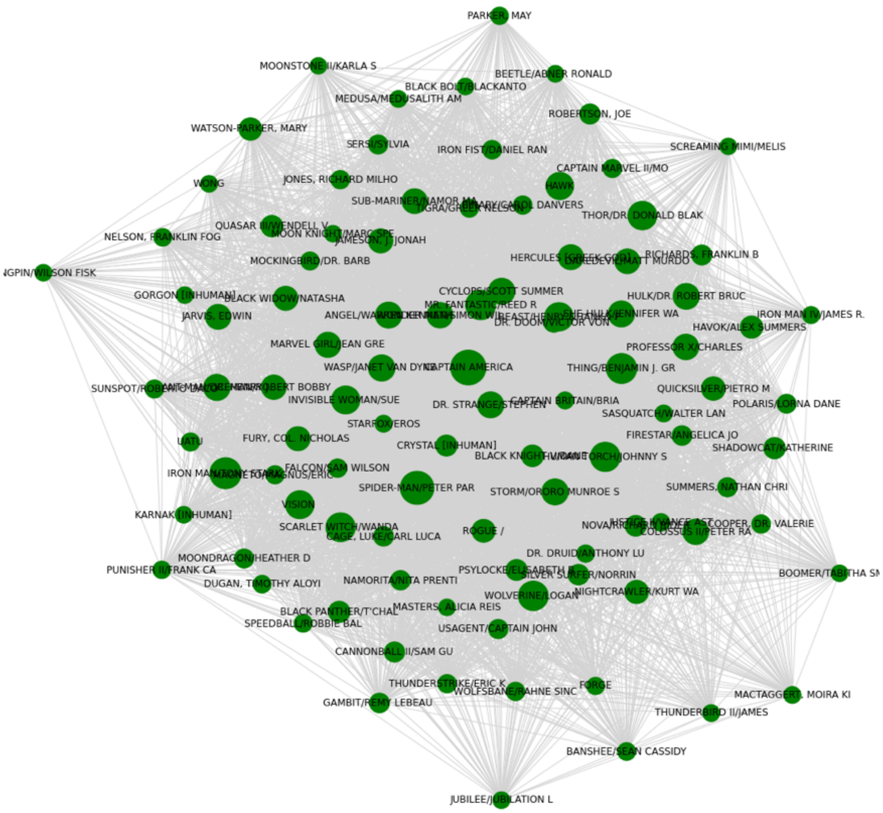


Figure one shows a subgraph of Marvel Universe network which includes 100 characters with highest degrees. The node size represents the degrees of character in the whole Marvel Universe. Considering that the network is established by the co-appearance in same comic series, the larger of node size represents more frequent appearance in the narrative. In other words, this network could be regarded as the subgraph of main characters in the Marvel Universe. Characters with more degrees are closer to the center position of the network structure, such as “Captain America”, “Mr Fantastic” etc. Yet some characters (e.g., “Captain Britain”, “Black Knight”, “Starfox” etc.) with relatively lower degrees more frequently co-appear with these main characters but not co-appear with other characters outside of this subgraph. After the detailed looking of these characters, I find these characters appear as early as other popular characters, but their contribution or appearance in comic is not as frequent as popular characters during the recent decades. It also inspires those certain characters are more tightly and densely connected with some characters than others, and this in agreement with definition of small community. The subgraph partially reflects the time-evolving, unpredictability, and complexity nature of Marvel Universe Network.

4. Method

4.1 Link Prediction

In network theory, Link prediction aims to predict whether there is a potential link between two nodes based on existing network structure (Hasan et al 2011). Adamic–Adar measure is a topology-based method for link prediction, assuming that nodes with similar network structure are more likely to be linked (Adamic and Adar 2003). The Adamic–Adar index is defined as:

$$A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$$

where $N(u)$ is the set of the adjacent nodes of u. If the index between two nodes is higher, the two nodes are closer.

To answer the first question, I will use the Adamic–Adar index algorithm for each node. Link prediction is widely employed for studying the structure and predicting the evolution of complex social networks, but the accuracy of link prediction is very dependent on the structure, size, and context of the network. By randomly removing 50% of existing edges for target characters, the Adar rank method predicts the most likely linked friends and least linked friends. The measurement is the prediction accuracy which is calculated as the matching between predicted friends and true friends whose tie with target character is removed previously. Then, I exclude nodes with prediction accuracy lower than 40%. Because these nodes perform worse in link prediction and are outliers for further regression analysis. I calculate the average prediction accuracy of nodes with same degrees and intend to regress the average accuracy on the standardized degrees and square of standardized degrees.

4.2 Nonlinear Regression

Nonlinear regression is a regression model for revealing nonlinear relationship between dependent variable and independent variable. In this study, I observe a non-linear relationship between degree and average link prediction accuracy of nodes with same degree. Thus, I construct a model:

$$A_i = \alpha + \beta_1 * d_i + \beta_2 * {d_i}^2 + \varepsilon$$

Where d_i is the standardized degree number and A_i is the average link prediction accuracy of nodes with d_i, α is the intercept, β_1 is the coefficient of the degree and β_2 is the coefficient of degree square. I first employ polynomial transformation to transform the standardized degrees and then employ Python statsmodel OLS linear regression to explore the coefficients and their statistical significance. The significance of coefficient is measured by the p value.

4.3 Community Detection – Network

Community detection aims to find subgroups or small communities in the complex social network. Characters within the smaller group are closer to each other in terms of their network structure, compared with the nodes outside the small community. Popular method for community detection includes Louvain method (Blondel et al 2008), Girvan–Newman algorithm (Girvan and Newman 2002) and K-means clustering.

To answer the second research question, I will use the K-means clustering based on eigenvectors for community detection of Marvel Network Universe subgraph which contains 100 most popular characters. I determine the number of clusters based on elbow method. I will measure the accuracy based on Sum square of error which calculates the within cluster variation. Besides, I intend to shed light on what are the characteristics of each small community. According to Girven and Newman (2002), characters in a graph are more tightly connected to other characters within the community but are more loosely connected with other characters outside the community. Thus, an interpretation of the small community with high density in this research context might be that these characters are more likely to co-appear in the comic series or share with similar characteristics.

4.4 Community Detection - Word Embedding and Topic Modeling

Word embedding is a popular representation method in natural language processing by transforming the word in a sentence or document into a vector where the dimension of vector refers to the meanings and context of the word (Jurafsky and Martin 2000). Topic modeling is a text-mining method to find the hidden semantic structure of each sentence or document by clustering similar words in each piece of text for topic generation (Blei et al 2003). Latent Dirichlet Allocation is a technique for topic modeling based on Dirichlet distributions. Each topic is a probabilistic distribution of words, and each document is a probabilistic distribution of topic. With large collection of unorganized textual data, topic modeling helps to find probabilistic topics for each document.

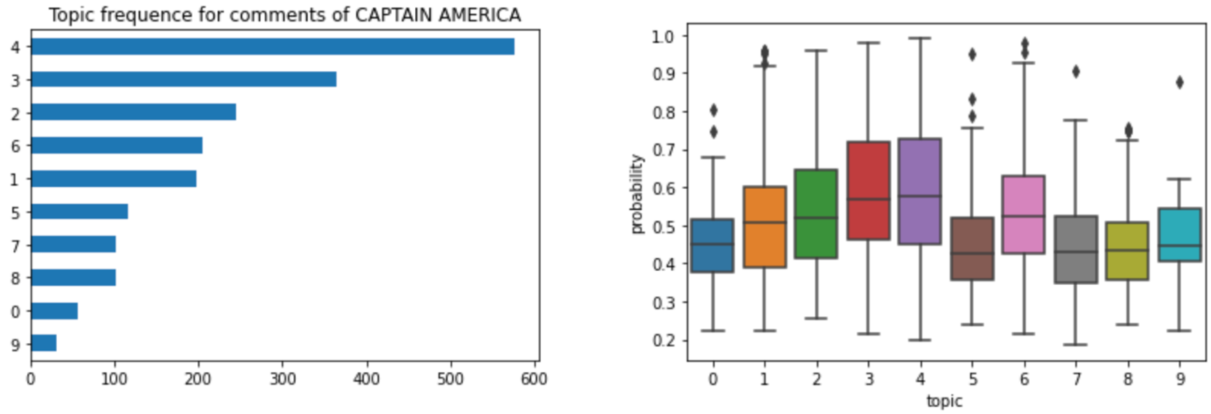
To answer the second research question, the comments for each character from subreddit platform can be used for detecting topic-based and semantic-based communities. I will employ topic modeling to find topic-based communities and word embedding method to find semantic-based communities. By grouping all characters into several small communities, we can compare these communities with network-based communities. Since there is no numerical accuracy measurement for the community detection, I will put more weight on comparing members within communities.

Topic-based communities

For the lad topic modeling, I start to tokenize and normalize the corpus and then use the topic modeling to generate 10 topics. When using the topic modeling, each piece of comment may be associated with different topics with different likelihood, I select the most frequent topic for each character as label of each character. For each character, I will calculate the average probability associated with each topic. If the most frequent topic of two characters is the same, there would be an undirect edge between these two characters. If there is more than one most frequent topic for a character, I use both topics to build edge. For example, if A's most frequent topic is [0, 4], B's most frequent topic is 4, and C's most frequent topic is 0, then A and B would be connected through an edge, and A and C would be connected through an edge. Therefore, If characters share similar topics, they are in the same community.

Taking Captain America as an example, figure three (left) shows the frequencies of topics of comments which relate to Captain America. Topic 4 is the most frequent appeared topic for Captain America and comments which are identified as topic 4 have highest average probability association.

Figure three: topic frequencies of comments and probability distribution of topics about Captain America



Yet, topics generated from LDA topic modeling is hard to interpret. To have a better definition of each topic, I print out top 10 pieces of comment which have the highest probability of each topic, manually analyze each piece of comment, and extract commonalities for the definition of each topic. For example, for topic 1, LDA modeling generates:

(top1, “0.010*"marvel" + 0.008*"comic" + 0.008*"year" + 0.008*"earth" + 0.007*"hulk" + 0.007*"time" + 0.006*"aveng" + 0.006*"univers" + 0.006*"charact" + 0.006*"ag")

where it is difficult to interpret indirectly from the words. Thus, some representative pieces of comments associate with high probability of topic 1 help for topic definition and clarification.

Table two: Top 4 highest probability comments associated with topic one

Comment	Probability
Doom is **legitimately** better than many **authoritarians** in that his desire to give the people he brutally rules over with a literal and figurative iron fist a better life is sincere. But he will not accept any course of action that involves him doing so by *not* **brutally ruling** over them with an iron fist and doing as much violence as necessary to **establish and enforce that rule**.	0.9678
Captain America believes in what **America is supposed to represent,** not what the country really is. He's perfectly aware of the many fallings of his nation and is **ashamed** of them. He likes the american dream and the **american values** Freedom, Security and the pursuit of happiness?	0.9590
He’s a veteran. At no point has it shown (that I could remember) that Capt. still has fulfill any **military obligations**. He’s Captain America because he **earned that rank** and can still use it. If he was promoted post humous, yes, it would still be **honored**.	0.9499
He genuinely cares about everyone and their rights he just hates Spider-Man being a masked vigilante who answers to no body, he’d very much be called a **liberal in real world USA**	0.9470

In comments for topic 1, I find most of them mention the political ideology (liberal US world, military obligations, American values) and reflect the high-status of character (honor, earn that rank and rule, and etc.). Thus, I annotate the topic as political ideology and high-status. The characters associated with this topic are more likely represent the political ideology or have higher status in the Marvel Universe.

For example, for topic 6, LDA modeling generates:

(topic 6: '0.053*"http" + 0.038*"com" + 0.026*"wiki" + 0.011*"org" + 0.011*"marvel" + 0.011*"www" + 0.010*"wikipedia" + 0.009*"amp" + 0.009*"earth" + 0.009*"vader"')

Table three: Top 4 highest probability comments associated with topic 6

Comment	Probability
Time travel doesn't work that way in the MCU It’s not really true time travel in the MCU though, it's more like jumping into parallel universes that haven't caught up with our own time stream; You can't change your own past **by going back** and **altering things**. Captain America wouldn't be changing the past; he would still be changing the **future** of the parallel universe since all the events mentioned in the original post wouldn't have happened in the **alternate timeline**.	0.9774
I assume that you don't have to **travel to the past of your own timeline**, but can travel to the **past/present/future of any timeline**, With that, would it be possible for a non-original MCU **timeline traveler** or a Variant, to end up traveling back to the *true* *original* MCU timeline, which could in a way Watsonianly/in-universely explain why there was an Ant-man actor in the particular stage play, because *an* Ant-Man could very well infact be there after all in the *original* Battle of NY?	0.9749

It would be infinite variations of every one of the variants seen. Holland-Spidey has several **What If episodes**. Presumably there's What If worlds for McGuire and Garfield versions of Spidey, and all of them were getting pulled in. **Infinite variations** all would've been pulled into that world, and I guess the worlds where everyone knows about Spider-Man would've been pulled in as well. Destroying both the MCU world and those versions where Peter Parker is public knowledge.	0.9727
Your **timeline is a bit off.** Peter Parker graduated high school in the 60s. He was out if college by the end of the 70s. Other than that, the **sliding time scale** is meant to explain these things in Marvel. DC is another beast because they have not been **running for the same time period.** Every reboot resets the timeline and thus there is no such issue of past conflicting tech because that **time** no longer exists.	0.9718

In comments for topic 6, I find time-period is the most frequent words. Time travel and alternative universe are the popular story narrative of Marvel Universe and one of the most interesting discussion topics among fans. Thus, I annotate the topic as story narrative: time travel and alternative universe. The character relates to this topic may be more likely to involve such narrative in the past or future.

Table four: the definition of the topics

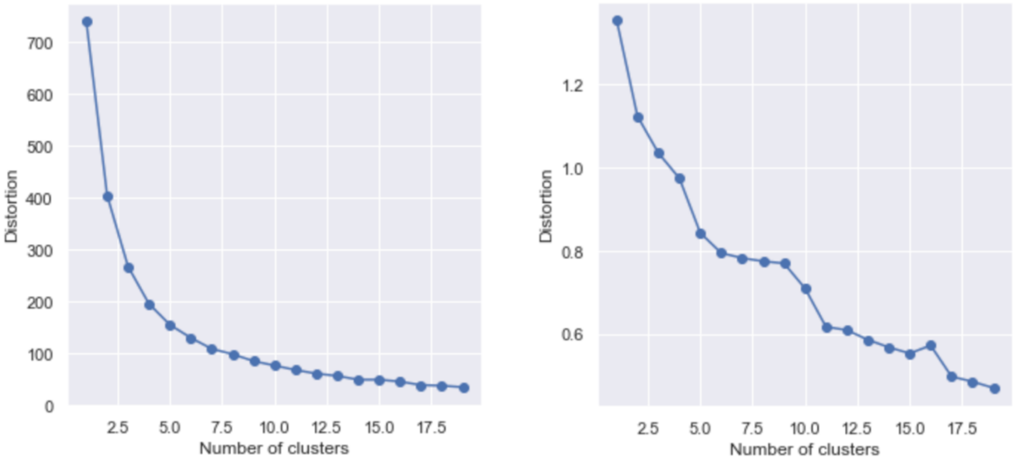
Topic 0	Weapon, war, universe

Topic 1	higher status, political ideology
Topic 2	team & communities & relationships
Topic 3	superpower and specialties
Topic 4	discussions & imagination about what will happen if XXX
Topic 5	civilian identity to against crime
Topic 6	time-infinite and alternative universe
Topic 7	mutant, non-human
Topic 8	magic, overconfidence
Topic 9	ultimate adamantium

Semantic-based community

I use pertained word2vec model to calculate the word embedding of each comment and then represent each character with the average word embedding vector of all comments associated with a character. Dimension reduction method would be used for reducing the dimensions of the average embedding vector of each character and then K-means clustering would help to find communities. High dimension data contains more information, but it may involve collinearity problem and becomes sparser in distribution. Dimension reduction can be used to improve the efficiency of clustering. I use Elbow method to determine the best number of clusters. Figure three shows the Elbow-method result for both dimensions reduced data and original high dimension data. The Error Sum of Squares (SSE) decreases steadily when the number of clusters reaches 10 for both dimension-reduced data and high dimension data.

Figure Three: Elbow-method (dimension reduced vs. high dimension data)



Beyond community detection, cosine similarity metric between embedding vectors of characters is used to compare the similarity of associated comments of each character on the subreddit platform.

5. Result & Discussion

5.1 Non-linear relationship

The regression result:

$$A_i = 0.5559 - 0.0802^{***[3]} \times d_i + 0.0304^{***} \times {d_i}^2$$

For nodes with degrees lower than a threshold, the more degrees the character has, the less likely the new links would be added to this character. For nodes with degrees higher than a threshold, the more degrees the character has, the more likely the new links would be added to the character. 0.0802 and 0.0304 are the standardized coefficients of the standardized degrees and

[3] *,**,*** represents p < 0.1, 0.05, 0.01 respectively

the square of the standardized degrees. The threshold could be calculated by taking derivative of the fitted function. The derivative function is 0.0608 × d_i – 0.0802.

Relationship between degree and link prediction accuracy is a concave curve, which means that the link prediction accuracy is high for the characters with low-degree and high-degree but low for characters with medium degrees in the universe. A possible explanation for this finding may that creator should maintain content creation around main characters (those with high degree) because these main characters have a large audience and are the main force to move the story forward. Whereas content creation which puts more weights on less popular and ignored characters who bring innovation and freshness to the narrative so that it raises the interest of the audience. For characters with medium degrees, they have been involved in old stories and may not be attractive and fresh enough to attract new audience.

Figure four: average link prediction accuracy of same degree nodes

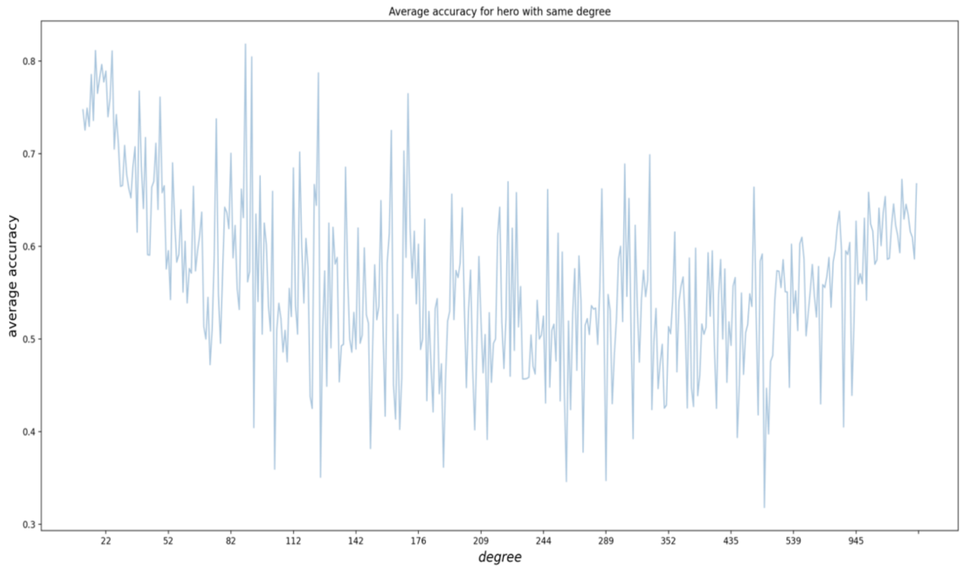
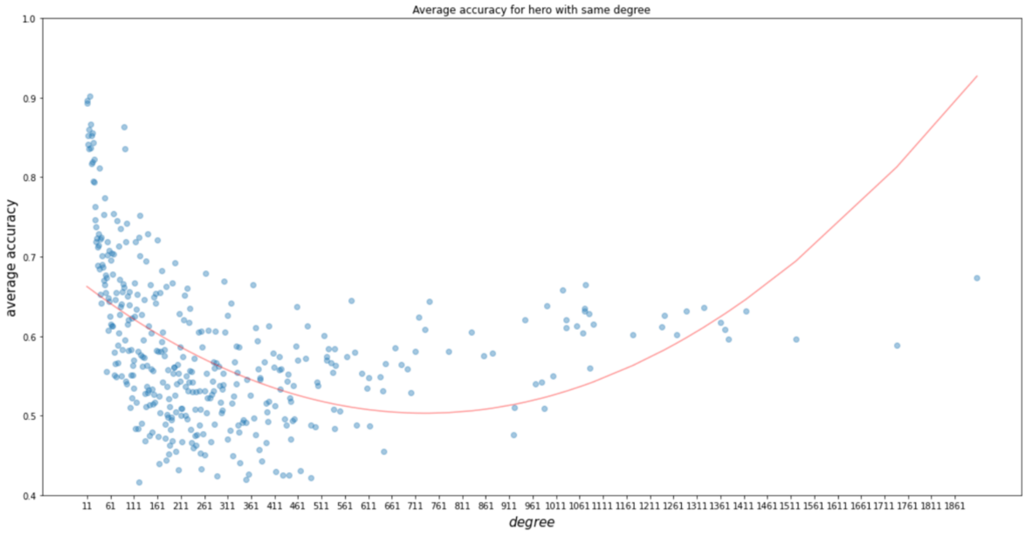


Figure five: average accuracy of same degree nodes (exclude nodes with prediction accuracy < 40%)



Yet some challenges remain unresolved. Firstly, the variance that can be explained by the model is low, which means there may existing some unobservable variables. Besides, whether such non-linear relationship would stay robust or not in the subgroup, for example within community detected from this Marvel Universe, is questionable. Secondly, the link prediction for some characters is extremely low, what are the share characteristics for these nodes with low prediction performance requires a further look. Thirdly, whether such non-linear relationship between degree and prediction accuracy could be generalized to other complex network, and how such artificial network similar or dissimilar with the real-life social network

5.2 Community detection from network

The second finding is that although community detected from the Marvel universe network does not perfectly match the semantic-based community and the topic-based community, these detected communities either contains characters from the same comic series or share similar characteristics (e.g., heroic or villain).

From the perspective of community size, network-based community and semantic-based community are more balanced. For topic-based communities, there are very large communities which contain 27 characters and there are very small communities which only have 2 characters. The methodological for community detection may explain such difference in community size. Topic-based communities are groups which contains characters who share the most frequent topic. Since the most frequent topic only reflects the most popular evaluation or description for a character from fans and topic modeling method emphasizes on the frequency distribution of words, it could not reflect either a multi-dimensional characteristic of characters or the semantic structure of comment. In addition, both semantic-based and network-based community detection uses K-means clustering, which helps to balance the community size.

From the perspective of community members, topic-based community members are more likely to co-appear in the same comic series. For example, one topic-based community contains ‘Banshee’, ‘Beetle, ‘Black Knight V’, ‘Captain Britan’, ‘Firestar’, ‘Gorgon’, ‘Havok’, ‘Herclues’, ‘Human Torch’, ‘Iceman’, ‘Invisible woman’, and etc. Most members within this community are from Fantastic four or X-man series where most of members are mutant with special power. The represented defined topic for this community is teamwork and relationship, which corresponds to what these characters perform in both Fantastic Four and X-man series. In X-man story, mutant collaborate with each other to fight against people who want to abuse them and try to achieve co-existence between human and mutant. In Fantastic four story, mutant works together to fight against devil and protect earth. However, topic-based communities do not differentiate heroic character and villains. Semantic communities help to differentiate hero from villain. One semantic community contains “Wolverine”, “Cyclops”, “professor X”, “Mageneto”, “Iceman”, “Shadow cat” and etc. They are all superheroes in X-man series, but these characters also have evil side. Network-based communities gather characters with similar characteristics, but they are from different comic series. For example, one semantic community contains 'Storm', 'Herclues [GREEK GOD]', 'Black Knight V', 'Nova', 'Gambit', 'Wolfsbane', 'DR. Druid', 'Forge', 'Medusa'. These villains appear in different comic series including X-man, Amazing Adventures, Avengers etc.

Figure six: community detection from Marvel universe network

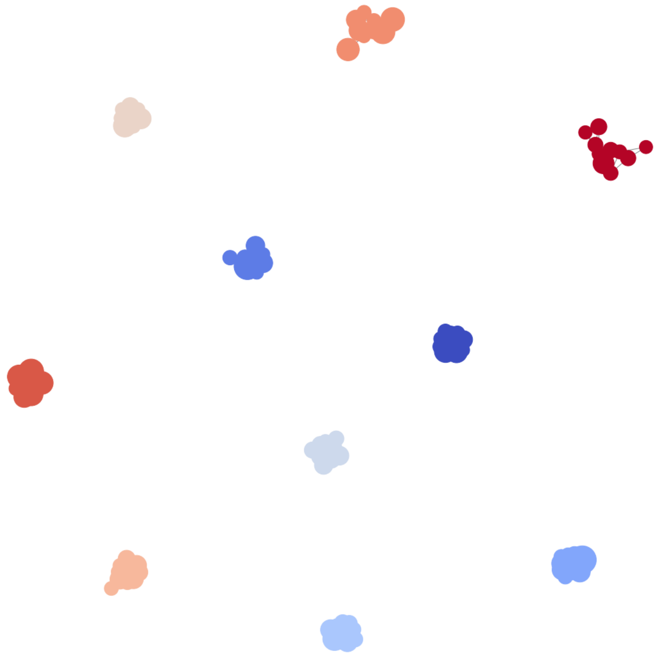


Figure six: community detection from Marvel comment network (topic modeling)

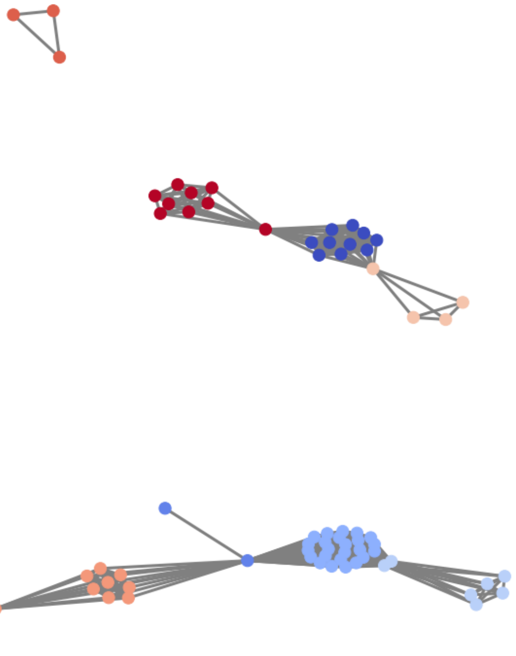


Figure seven: community detection from Marvel comment network (word embedding)

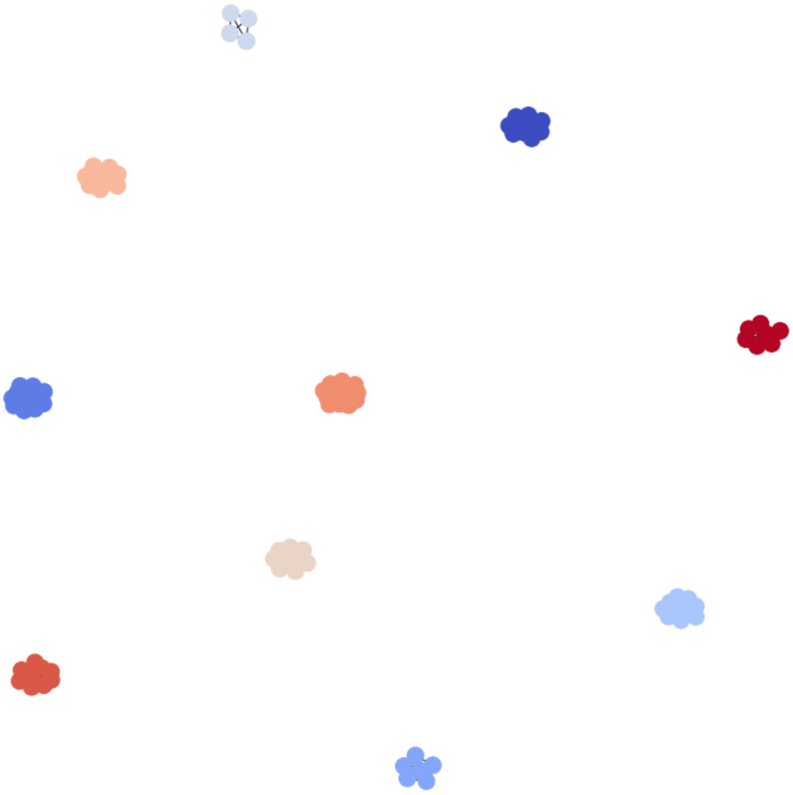


Table five: Community

	Topic-based	Semantic-based	Network-based
Community 1	'BANSHEE', 'BEETLE', 'BLACK KNIGHT V', 'CAPTAIN BRITAIN', ‘FIRESTAR', 'GORGON [INHUMAN]', 'HAVOK', 'HERCULES [GREEK GOD]', 'HUMAN TORCH', 'ICEMAN', 'INVISIBLE WOMAN', 'JUBILEE', 'MARVEL GIRL', 'MEDUSA', 'MOON KNIGHT', 'NIGHTCRAWLER', 'POLARIS', 'PSYLOCKE', 'SCREAMING MIMI', 'SHADOWCAT', 'SILVER SURFER', 'STARFOX', 'SUB-MARINER' ,'SUNSPOT', 'TIGRA', 'USAGENT', 'WOLVERINE'	'HUMAN TORCH', 'INVISIBLE WOMAN', 'HAWK', 'IRON FIST', 'FALCON', 'SUNSPOT', 'BANSHEE', 'BEETLE', 'BLACK BOLT'	'BEETLE ', 'GORGON [INHUMAN]', 'JUSTICE II', 'KARNAK [INHUMAN]', 'MARVEL GIRL', 'NELSON, FRANKLIN FOG', 'PSYLOCKE’, 'SCREAMING MIMI', 'WATSON-PARKER, MARY '
Community 2	'ANGEL', 'BEAST', 'BINARY', 'COLOSSUS II', 'FORGE', 'GAMBIT', 'JUSTICE II', 'NOVA’, ‘ROGUE ', 'SASQUATCH', 'SERSI', 'STORM'	'COLOSSUS II', 'SUB-MARINER', 'CAPTAIN MARVEL II', 'MOCKINGBIRD', 'TIGRA', 'JUSTICE II', 'PUNISHER II', 'CAPTAIN BRITAIN', 'SCREAMING MIMI'	ANGER, BLACK PANTHER, 'BOOMER', 'CAGE, LUKE’, 'HUMAN TORCH', 'IRON MAN’, 'JAMESON, J. JONAH', 'MOONDRAGON ', 'QUASAR III', 'SPEEDBALL', 'STARFOX/EROS'

Community 3	'ANT-MAN', 'BLACK PANTHER', 'BLACK WIDOW', 'CAPTAIN AMERICA', 'CAPTAIN MARVEL II', 'DR. STRANGE', 'FALCON', 'HULK', 'IRON MAN', 'JARVIS, EDWIN ', 'MOCKINGBIRD', 'MOONDRAGON', 'NAMORITA', 'SCARLET WITCH', 'SHE-HULK', 'THOR', 'THUNDERSTRIKE', 'VISION ', 'WASP', 'WONG'	'CAPTAIN AMERICA', 'IRON MAN', 'SCARLET WITCH', 'WASP’, ‘ANT-MAN', 'MARVEL GIRL', 'BLACK WIDOW', 'BLACK PANTHER', 'CAGE’, ‘LUKE', 'USAGENT', 'MOONDRAGON', 'STARFOX', 'MOON KNIGHT'	'ANT-MAN', 'CANNONBALL II', 'INVISIBLE WOMAN ', 'JARVIS, EDWIN ', 'MASTERS, ALICIA REIS', 'MOONSTONE II', 'NOVA', 'SERSI', 'SILVER SURFER', 'WONG'
Community 4	'BANSHEE', 'BEETLE', 'BLACK KNIGHT V', 'CAPTAIN BRITAIN', 'FIRESTAR', 'GORGON [INHUMAN]', 'HAVOK', HERCULES [GREEK GOD]', 'HUMAN TORCH', 'ICEMAN', 'INVISIBLE WOMAN', 'JUBILEE', 'MARVEL GIRL', 'MEDUSA', 'MOON KNIGHT’, ‘NIGHTCRAWLER', 'POLARIS', 'PSYLOCKE', 'SCREAMING MIMI', 'SHADOWCAT', 'SILVER SURFER', 'STARFOX', 'SUB-MARINER', 'SUNSPOT', 'TIGRA', 'USAGENT', 'WOLVERINE'	'THOR', 'PSYLOCKE', 'NAMORITA', 'THUNDERSTRIKE', 'JUBILEE', 'GORGON [INHUMAN]'	'BANSHEE', 'BLACK KNIGHT V ', 'COLOSSUS II', 'CYCLOPS ', 'IRON MAN IV', 'MAGNETO ', 'NAMORITA', 'RICHARDS, FRANKLIN B', 'SHADOWCAT ', 'SPIDER-MAN', 'STORM', 'SUMMERS, NATHAN CHRI', 'WONDER MAN'
Community 5	'BLACK BOLT', 'BLACK KNIGHT V', 'CYCLOPS', 'GORGON [INHUMAN]', 'IRON MAN IV', 'KARNAK [INHUMAN]', 'QUICKSILVER'	BEAST', 'ANGEL', 'STORM', 'HERCULES [GREEK GOD]', 'BLACK KNIGHT V', 'NOVA', 'GAMBIT', 'SERSI', 'WOLFSBANE', 'DR. DRUID', 'FORGE', 'MEDUSA	'BEAST ', 'CRYSTAL [INHUMAN]', 'FORGE', 'IRON FIST', 'JONES, RICHARD MILHO', 'JUBILEE', 'MR. FANTASTIC', 'PROFESSOR X ', 'SCARLET WITCH ', 'THUNDERSTRIKE '

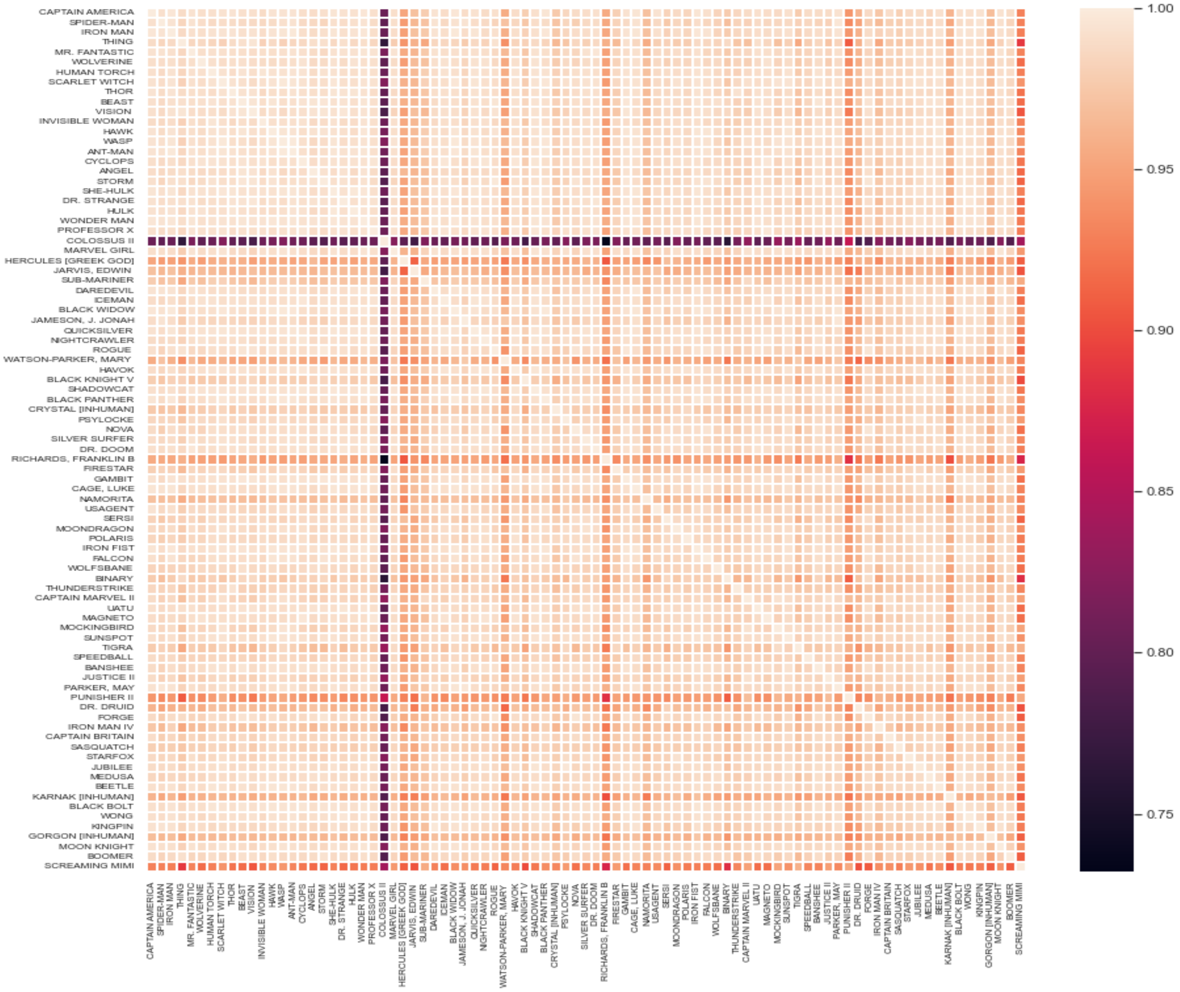
Community 6	'BOOMER', 'CAGE, LUKE', 'COLOSSUS II', 'DAREDEVIL', 'DR. DRUID', 'IRON FIST', 'SPEEDBALL', 'THING', 'WOLFSBANE'	'WOLVERINE', 'CYCLOPS', 'PROFESSOR X', 'ICEMAN', 'NIGHTCRAWLER', 'HAVOK', 'SHADOWCAT', 'POLARIS', 'MAGNETO'	'BLACK BOLT', 'DR. DRUID', 'HERCULES’, 'MEDUSA', 'PARKER, MAY', 'PUNISHER II', 'ROGUE', 'SUNSPOT', 'THUNDERBIRD II', 'USAGENT'
Community 7	'CRYSTAL [INHUMAN]', 'MAGNETO', 'PROFESSOR X’	'CRYSTAL [INHUMAN]', 'IRON MAN IV', 'SASQUATCH', 'KARNAK [INHUMAN]'	'COOPER, DR. VALERIE', 'FALCON ', 'ICEMAN ', 'MOCKINGBIRD', 'MOON KNIGHT', 'POLARIS, 'QUICKSILVER', 'SASQUATCH ', 'THING', 'WOLVERINE '
Community 8	'DR. DOOM', 'MR. FANTASTIC', 'NAMORITA', 'PARKER, MAY', 'PUNISHER II', 'RICHARDS, FRANKLIN B', 'SCREAMING MIMI', 'SPIDER-MAN', 'UATU', 'WATSON-PARKER, MARY '	'SPIDER-MAN', 'JARVIS, EDWIN ', 'DAREDEVIL', 'JAMESON, J. JONAH', 'WATSON-PARKER, MARY ', 'FIRESTAR', 'SPEEDBALL', 'PARKER, MAY', 'KINGPIN'	'CAPTAIN AMERICA', 'CAPTAIN MARVEL II', 'DR. STRANGE ', 'DUGAN, TIMOTHY ALOYI', 'KINGPIN', 'MACTAGGERT, MOIRA KI', 'SHE-HULK ', 'THOR'
Community 9	'HAWK', 'JUSTICE II', 'KINGPIN', 'WONDER MAN'	'MR. FANTASTIC', 'SHE-HULK', 'HULK', 'WONDER MAN', 'QUICKSILVER', 'SILVER SURFER', 'DR. DOOM'	'HAWK', 'HULK', 'NIGHTCRAWLER', 'SUB-MARINER', 'TIGRA', 'UATU', 'VISION ', 'WOLFSBANE'

Community 10	'JAMESON, J. JONAH', 'SCREAMING MIMI'	'THING', 'VISION ', 'DR. STRANGE', 'ROGUE ', 'RICHARDS, FRANKLIN B', 'BINARY', 'UATU', 'WONG', 'BOOMER'	'BINARY', 'BLACK WIDOW ', 'CAPTAIN BRITAIN/BRIA', 'DAREDEVIL/MATT MURDO', 'DR. DOOM ', 'FIRESTAR', 'FURY, COL. NICHOLAS ', 'GAMBIT ', 'HAVOK ', 'ROBERTSON, JOE', 'WASP '

5.3 Similarity

Figure eight shows the pairwise similarity between average word embedding vectors of comments associated with each character. Most similarity score ranges from 95% to 100%, while a few characters have around relatively lower similarity (90%-94%) with other characters such as Screaming Mimi and Punisher II. Only one character, Colossus II, has distinguished dissimilarity with all else characters. Based on the cosine similarity between characters, I find that all characters could be divided into two groups – superheroes and villains. In the heatmap, each row represents its similarity with other characters. If the color of the majority metric in a row is lighter, the representing character would be more likely in the superhero group. Most main characters are justice heroes. Unlike topic modeling or clustering which detects small communities, cosine similarity heatmap reveals to what extent the comments corresponding to a character differs that of other characters. Furthermore, the similarity between characters within the justice hero group have small variance, while the characters within devil group are not quite similar with each other. For example, “Captain America”, “Iron man”, “Spider man” are very similar with each other in terms of word embedding vector similarities (around 98%), but similarity score between embedding vector of “Screaming Mimi” and “Punisher II” is around 92%. This indicates that each antagonist is unique and is more likely to raise multi-faceted discussion among audience. In contrast, evaluations of justice hero are monotonic.

Figure eight: cosine similarity between represented vectors of characters



6. Conclusion

In conclusion, this paper employs link prediction and polynomial regression to find that future content creation is more likely to focus on popular characters and under-developed characters in Marvel Universe. For characters with medium level degrees, they would be less likely to receive new connections. In addition, this paper employs community detection method to detect network-based, topic-based and semantic-based communities in the most popular characters of Marvel Universe. Through comparing small communities, topic-based communities are more likely to co-appear in the same comic series and network-based communities are more likely to represent characters from different comic series but share similar characteristics, whereas semantic-based communities could represent characters from same comic series and share similar characteristics. This not only indicate that small communities exist in Marvel Universe, but also reflect how the characters connect with each other from both network in comic and comments from fans on the Reddit platform. The heatmap of cosine similarity between characters provides supporting evidence that there are small communities, and the characteristics of superheroes are more similar. But the characteristics of villains are different. Villains may serve a function of differentiating each comic series and make audience feel fresh about different comic series. These identifications help comic creators to understand how Marvel Universe designs characters and what are the interesting topics among the comments of fans in social media. Thus, content creators could generate more successful comic series. Yet there are still some future developments on this research topic. For example, the scripts or the communications of characters could be an additional textual resource to understand the interactions between characters. Furthermore, we can create a multi-facet social network considering not only characters, but also times, locations, universe, relationship to understand this complex system of artificial world. Since Marvel Universe has been developing over several decades, we can follow the locus of its development to simulate how each tie emerges and each character emerges as time going by to reveal insights about creativity and innovation in creative work.

Reference:

Alberich, Ricardo & Miro-Julia, J. & Rossello, Francesc. (2002). Marvel Universe looks almost. like a real social network.

Al Hasan, Mohammad; Zaki, Mohammed (2011). "Link Prediction in Social Networks" .

Adamic, Lada A; Adar, Eytan (2003). "Friends and neighbors on the web". Social Networks. Elsevier. 25 (3): 211–230. doi:10.1016/S0378-8733(03)00009-1. S2CID 2262951.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation". Journal of Machine Learning Research. 3: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Eom, YH., Jo, HH. Generalized friendship paradox in complex networks: The case of scientific collaboration. Sci Rep 4, 4603 (2014). https://doi.org/10.1038/srep04603

Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)

Gleiser, Pablo M. "How to become a superhero." Journal of Statistical Mechanics: Theory and Experiment 2007.09 (2007): P09020.

Jurafsky, Daniel; H. James, Martin (2000). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Prentice Hall. ISBN 978-0-13-095069-7.

Li, Jiarong, et al. "Complex networks of characters in fictional novels." 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). IEEE, 2019.

Mourchid, Youssef, et al. "Multilayer network model of movie script." International Conference on Complex Networks and their Applications. Springer, Cham, 2018.

Vincent D Blondel et al J. Stat. Mech. (2008) P10008