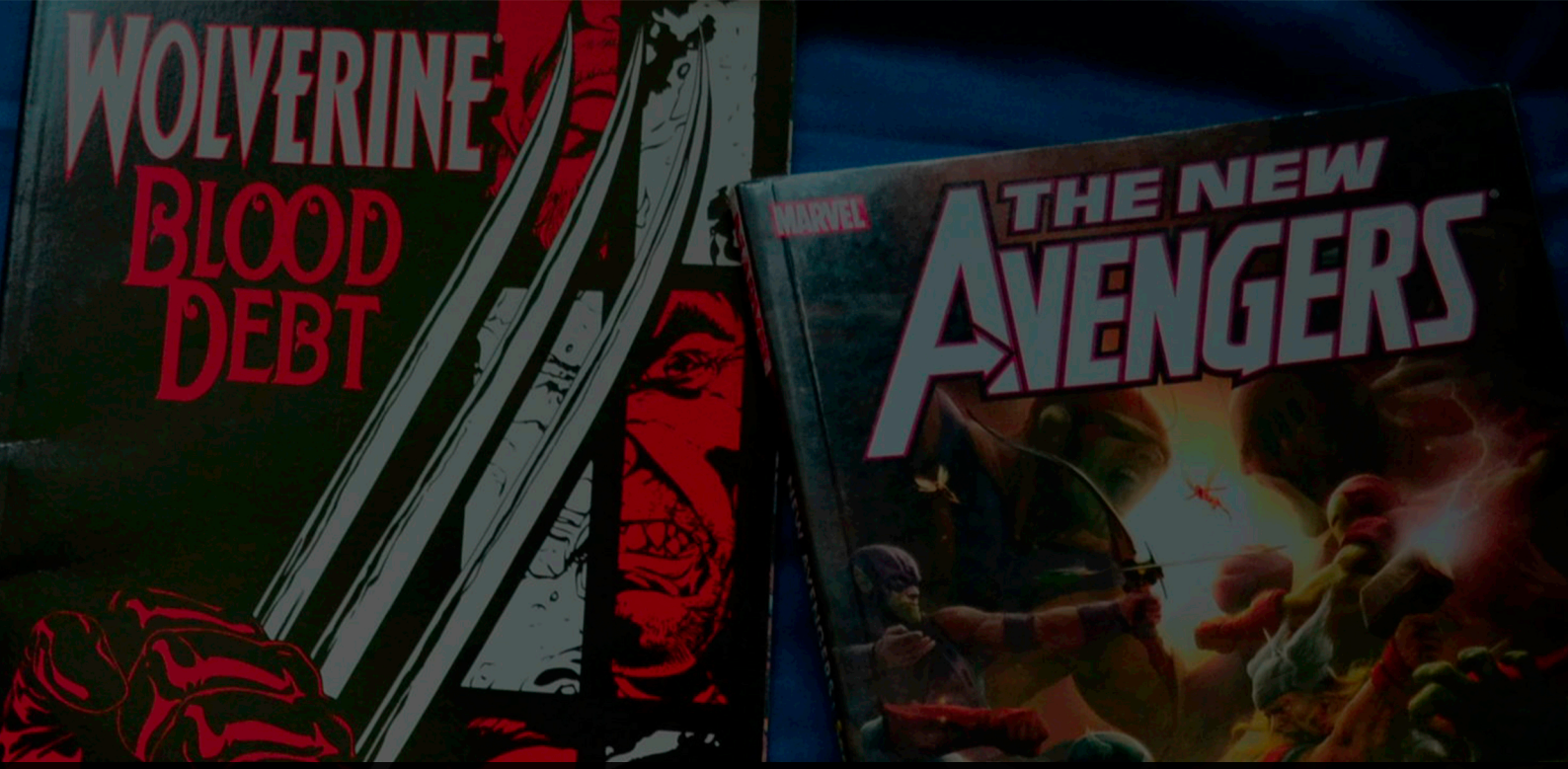
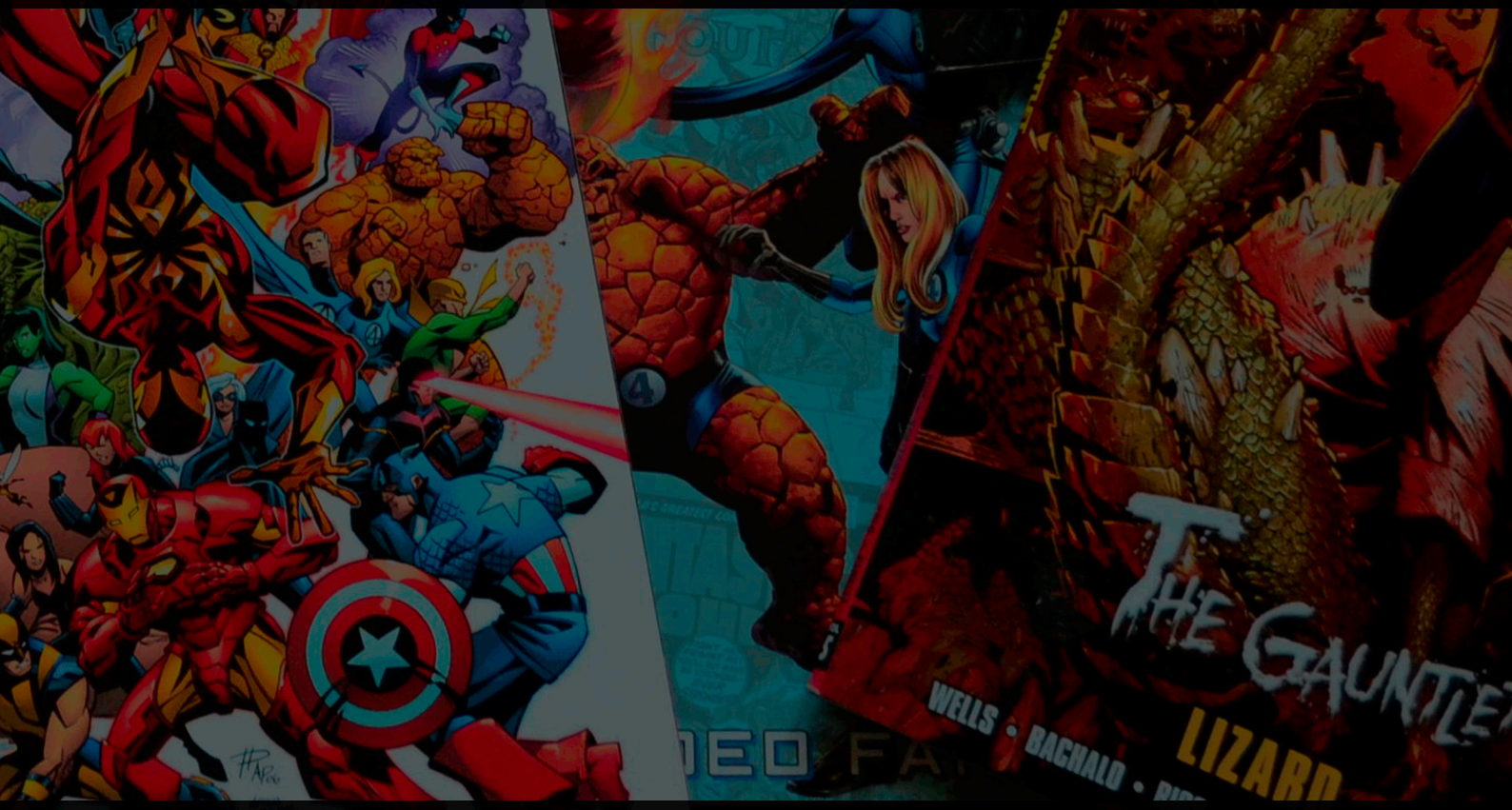


A Network Tour of Data Science

A Growing Network of Characters In Marvel and DC Universes

TEAM 28
Maximilien Gomart - 272500
Ahmed Kooli - 269603
Antoine Schmider - 274431

EPFL



EE-558 Final Project - January 10th, 2020

1 Introduction and motivations

Marvel Worldwide Inc. and Detective Comics, also known as DC Comics are the two main comic book publishers in the world. They've become famous throughout the world thanks to their iconic characters and stories, first developed in comic books, then in many movies and TV shows. Over the years, these two companies have told the stories of thousands of characters, they have developed their personalities and made them interact with each other to build two gigantic universes: the Marvel Universe and the DC Universe.

The relationships between the characters in both universes present a huge potential for graph visualisations : the characters are nodes, and the links between them show how close they are to each other. Having access to huge online databases, i.e. Marvel and DC Fandom Wiki, information such as the relatives of every character, their affiliation, the year of their first apparition and the comic books in which they appear can be gathered. This will allow an overall visualisation of the evolution of the two separate universes over the years.

The aim of this Network analysis project is to find which characters are the key to the success of Marvel and DC universes, and thus to find which ones are more likely to appear in the next comics or movies, and what other characters will appear along with them.

2 Construction of the datasets

2.1 Gathering the data

The data sets to be analysed are obtained from two websites: Marvel fandom and DC fandom, which consist respectively of more than 26 000 and 13 000 characters. Let's first note that a choice was made to only focus on the Marvel characters coming from a universe called Earth-616, the most famous and important one in terms of characters, to avoid duplicates. For each character, the gathered information was its `Name`, `Current Alias`, `Relatives` characters, `Affiliation` (team, group or organization the character is affiliated to), and `URL` used as a unique ID. After that, some information had to be obtained from the comic books: `Publication date` (to have the characters first appearance dates), `Subcomics Name` (if a comic book is made of multiple other comic books), and a list of `Characters` that appear in the specified subcomic and that are also present in the characters dataset. This was also a big source of data since more than 50 000 comics were written by each publisher.

2.2 Treating the data

Raw data is never easy to handle, especially the one coming from Wiki Fandom (Wikipedia) pages where anyone can modify the information. It had first to be cleaned to make it ready to use. Data can be missing, and these pages are constantly modified with the release of new comic books, this is why the following analysis still has its limitations and is based on the moment the data was scraped from the web and the information it contained.

As we base our analysis on three main characteristics: the `Relatives`, the `Affiliation` and the `Comic books`, we built adjacency matrices for each of them. For the `Relatives`, the weight is set to 1 if the characters are relatives, and 0 otherwise. The `Affiliation` and the `Comic books` are a more complicated: if two characters appears in the same team or comics we increase the weight between them by one. Thus we can really play with the strength of their connections later. Finally, one last adjacency matrix is created, based on the sum of the previous ones.

These adjacency matrices are then treated using `Gephi`, a software used for graph visualisation.

3 Influence of the attributes

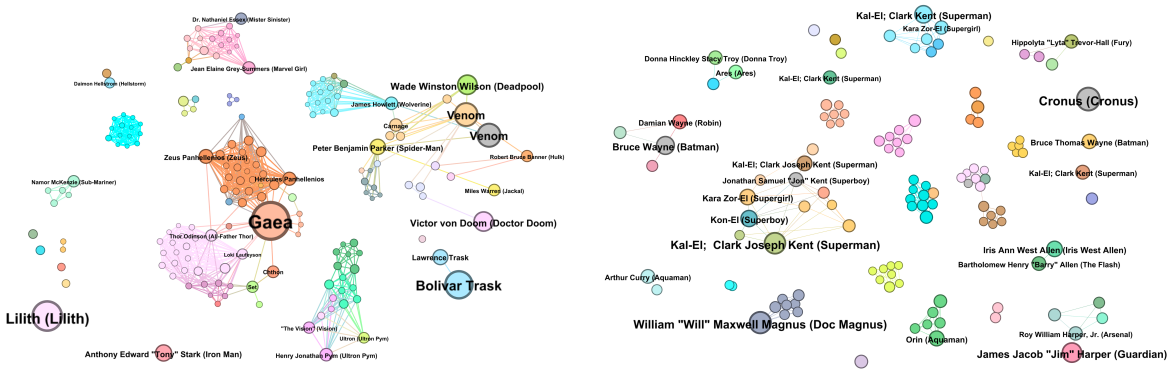
As explained previously, the network analysis was based on three different attributes: the `Relatives`, the `Affiliations` and the `Comic books`. What is interesting is that following the studied attribute, the analysis gives very different results.

3.1 Relatives analysis

The relatives graph works this way: the bigger a node is, the more relatives the corresponding character has. The motivation here is to build a graph highlighting the nodes that have the most connections to get a good overview of the characters having the biggest families.

In order to clearly visualize the relatives network, some filters were applied. The first one is a `Degree Range` filter with a threshold of `17` relatives for the Marvel graph, and `7` relatives for the DC graph. This threshold defines the minimum degree (i.e. the minimal number of relatives) that a node has to have to be displayed on the final graph. Secondly, a modularity algorithm was applied which looks for the nodes that are more densely connected together than to the rest of the network. It will then highlight them by color and will thus ease the clustering process.

After that, the `ForceAtlas 2` layout was used. It is a force-directed layout that implements repulsion and attraction between points based on their degree or links between them. Highly linked points will be attracted, whereas high-degree nodes will repulse other high-degree nodes to ensure good cluster visualisation.



(a) Marvel Universe

(b) DC Universe

Figure 1: Relatives clustering visualisation with ForceAtlas2

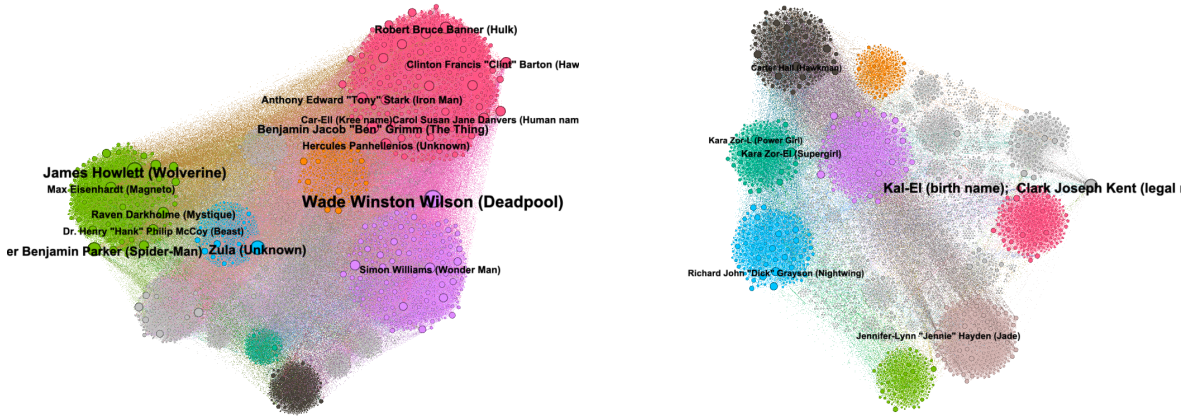
NOTE: multiple characters can be associated with a single alias: multiple Venom co-exist in the Marvel universe graph. This is because they correspond to different people that embodied the Venom character. It is the case for many other characters like Spider-Man, he died many times but it didn't cause the end of the character since other people incarnated it. Some characters have also been re-edited which create a new characters with similar characteristics

For the Marvel Universe graph, a higher degree threshold was needed as there are way more characters (more than twice the number of characters than in DC universe), and thus more characters to filter to ensure good visualisation of the main ones.
On the other side, as the degree threshold is lower for the DC universe, we would expect bigger clusters and families. It is not the case: main characters only have a few relatives and quite small families in general. It seems that writers of Marvel comic books care more about telling family stories and their members. DC writers, on the other hand, focus more on lone iconic figures that have only very few relatives such as Batman or Aquaman.

The relatives approach give us keys to understanding the main characters and families of both Marvel and DC universes, however it does present limits. This analysis can give too much importance to specific characters, just because they have many relatives. It is the case for Gaea, goddess of Earth, which is represented as the most important character of the Marvel universe in our graph, when we would rather expect Iron Man or Spider-Man to have this status. The same drawback appears in the DC universe graph, where Cronus, Gaea's son, is represented as the most important node in the graph.

3.2 Affiliations analysis

In the precedent approach, we saw that basing the analysis only on family ties between characters didn't give reasonable importance to the nodes in the graph. Relationships between characters in fact don't rely solely on family ties: strong connections can exist in many other ways. This is the reason why the affiliations were studied. An affiliation could be a team like the Avengers (a team of super-heroes), a place of origin (Asgardians), or a corporation (Stark Industries), among others. To make the clusters appears, this time we use the `Leiden's Algorithm` that assign each node to a cluster. Then we apply the `Cirle Pack Layout` to group this cluster together. This method is really convenient for large datasets as it outdoes the `Louvain's Algorithm` in false detection of cluster.



(a) Marvel Universe

(b) DC Universe

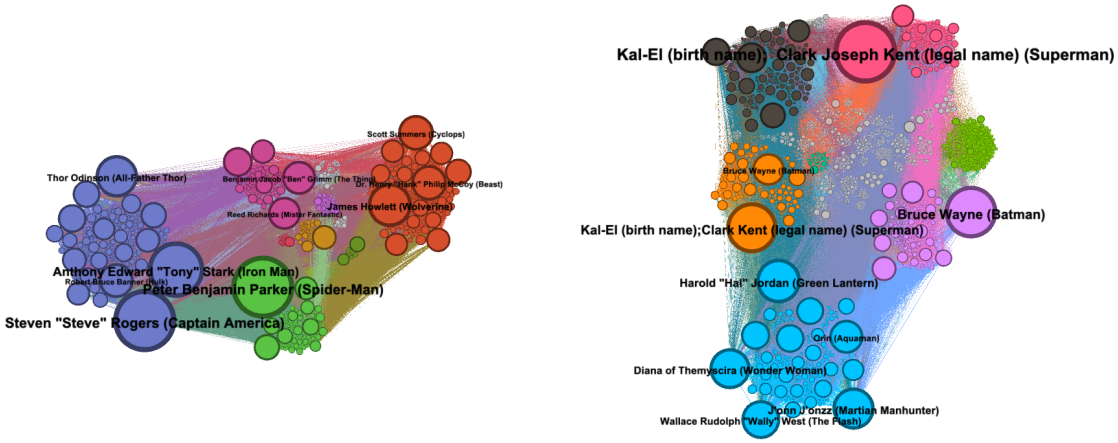
Figure 2: Affiliations clustering visualisation with Circle Pack Layout based on Leiden's Algorithm

The affiliations approach helps us visualising the main "clans" or "teams" in both universes. Moreover, the use of `PageRank` algorithm helps us filtering the most important characters in each of these teams. A recursive definition to this algorithm is that it gives more importance to nodes linked to other central (important) nodes. In other words, it represents the likelihood to arrive on a particular node after randomly walking through the network follwing links. We could thus infer main leaders, such as Wolverine, or Iron Man in the Marvel universe. In the DC universe, the main team-leaders are Kal-El (Superman), and his cousin, Kara Zor-El (Supergirl). This approach describes quite faithfully the structure of the teams in both universes.
However, it still present certain limits. For example, the Marvel graph gives too much importance to Dead Pool, because he belongs to a lot of different teams.

3.3 Comic book analysis

The comic book analysis can, in a sense, correct the problem that the relatives and affiliations attribute had. The graph based on this attribute not only links characters that appeared together in same comic books, but also give significance to characters that appear in a lot of comics in general.

As we are again dealing with huge data sets containing hundred of thousands of entries, we used the `Circle Pack` layout for visualisation with `Leyden's Algorithm`.



(a) Marvel Universe

(b) DC Universe

Figure 3: Comics clustering visualisation with Circle Pack Layout based on Leyden's Algorithm

The first observation that can be done using the comic book approach is that it represents the importance of the characters in both universes very faithfully. In the Marvel Universe, we would expect Iron Man, Spider-Man, Captain America and Thor to stand out as the most iconic figures, which is the case in the graph. Same for DC, where we would expect Batman, Superman, Green Lantern and Wonder Woman to be the main figures; the graph highlights it properly.
Moreover, for each iconic figure of both universes, the graph displays displays their own "universe", that is to say the main characters that they encounter the most during their stories. This approach, like the two precedent, also presents drawbacks. While the Marvel graph is in overall coherent with the clusters and the main characters of its universe, the presence of two clusters around two different Superman entities can be observed in the DC graph. As the comic book approach representation relies solely on the apparition of characters in comic books, and as Supermen from two different universes appear in comics along with new characters, they end up in two different clusters, even though they designate the same person.

4 Marvel and DC: two different writing approaches

In the previous sections, pros and cons of the different attributes were discovered. It could thus be a good idea to combine the three attributes and see what results are obtained. The straight-forward approach would be to simply sum up the contributions of each attribute. However two characters can only have one link as relatives, but they appear in hundreds of comic books together, which will create inconsistencies in the network if they are added without any treatment. This is why the links between characters need to be normalized by the maximum link value of the given attribute (12 and 5 for the affiliations, 2941 and 894 for comics, respectively for Marvel and DC; the relatives weights weren't affected since the maximum value is 1).

Let's now visualize the results. First, the most important nodes are selected using a `PageRank` filter: only characters which `PageRank` is greater than a certain value (that depends on the graph) have their names displayed. The node size depends also on the `PageRank`, and even if the range of values is different for the graphs, the biggest node always has the same size (which serves as reference). To get a good network visualisation, we used `Openord` Layout in `Gephi`. `Openord`, just as `ForceAltas2`, is a force-directed layout algorithm specialised in large data sets and thus made to build large graphs. It is particularly convenient for this case since the the combination of all attributes increased a lot the number of links. The `OpenOrd` layout was also followed by a `ForceAtlas` to avoid the overlap of the nodes. Moreover, to see its evolution over the years, the characters network is analysed on three time periods: beginning till 1950, beginning till 1990, and beginning till nowadays (beginning designating the publication date of the first comic book).

The following network (DC universe) shows that the key characters of nowadays (Superman, Green Lantern or The Flash) were also the most important ones in 1950 *(you can zoom in to see more clearly the displayed names)*. DC Comics, over the years, has focused more on developing the characters that were the most successful since its foundation. However in 1990, it seems that many characters were added and had the approximate same importance in terms of numbers of connections. This allowed the emergence of new key characters in the DC universe such as Batman or Doll Girl. In the actual network (2019), these characters are still present, and the number of bigger nodes, i.e. most important characters, has actually decreased. One can thus say that DC went back to their old ways of creating characters by focusing only on a limited number of them. Finally one of the interesting results is that some characters with the same alias had close nodes in terms of distance like the different Superman or The Flash, which shows that combining different attributes still gives meaningful results, and corrected some inconsistencies like the one observed for in the comics analysis.

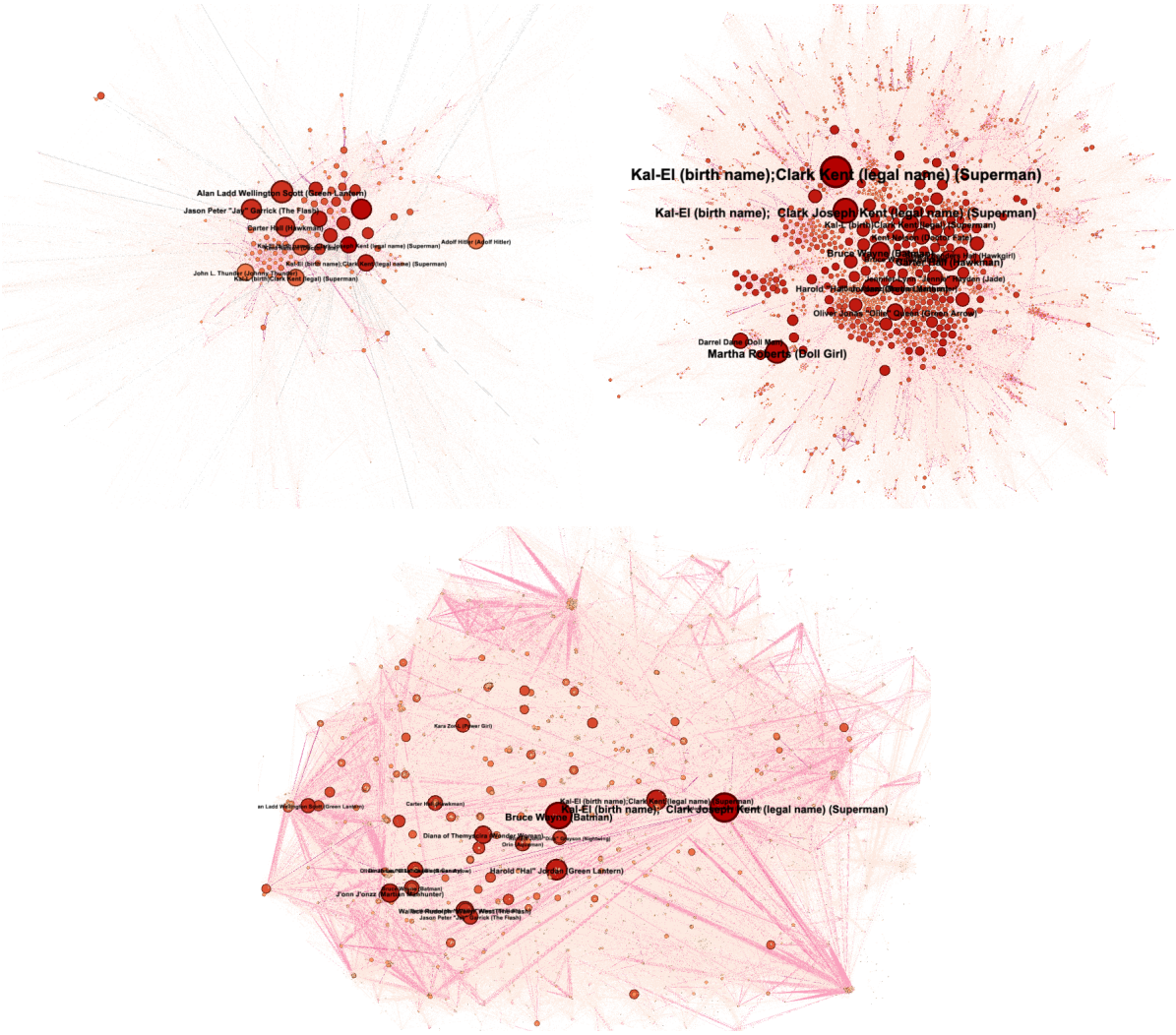


Figure 4: Evolution of the Marvel Universe (1950, 1990 and nowadays)

A similar phenomenon can be observed for the early years of the Marvel universe. The characters that were the most important ones are still highly represented in today's network. However, their importance has changed: for instance, the size of the Sub-Mariner's node has decreased whereas some other characters saw their importance grow over time like Captain America. This can be explained by the fact that other characters with a lot of links with the existing ones were created in the meantime. It thus had an effect on the importance of a character like Captain America since a whole network of characters around him was built over the years.
The Avengers is a meaningful example for this phenomena: many other affiliations linked to it were

created (S.H.I.E.L.D., Avengers Unity Division, Secret Avengers...) which increased the affiliations weight between the team members. Moreover, the characters that are affiliated to the same teams tend to appear in the same stories which again increases the comics edge weight (Captain America and Iron Man appeared in more than 2000 stories together!). See for example how successful were the recent comic books and movies about all the Avengers members together, which differs from those focusing only on one single character. Continuing with this example, the most recent network shows how the Avengers members got clustered together and are now the key players of the Marvel Universe: Spiderman, Iron Man, Hulk, the Thing and many others. This can also be seen in the earliest network with the nodes that are separated on the right. They actually represent a cluster of characters linked by the Imperial Japan affiliation. Note that this network follows directly the World War, which explains why so many Japanese characters were created at that time and why this cluster is not distinguished anymore in the recent networks.

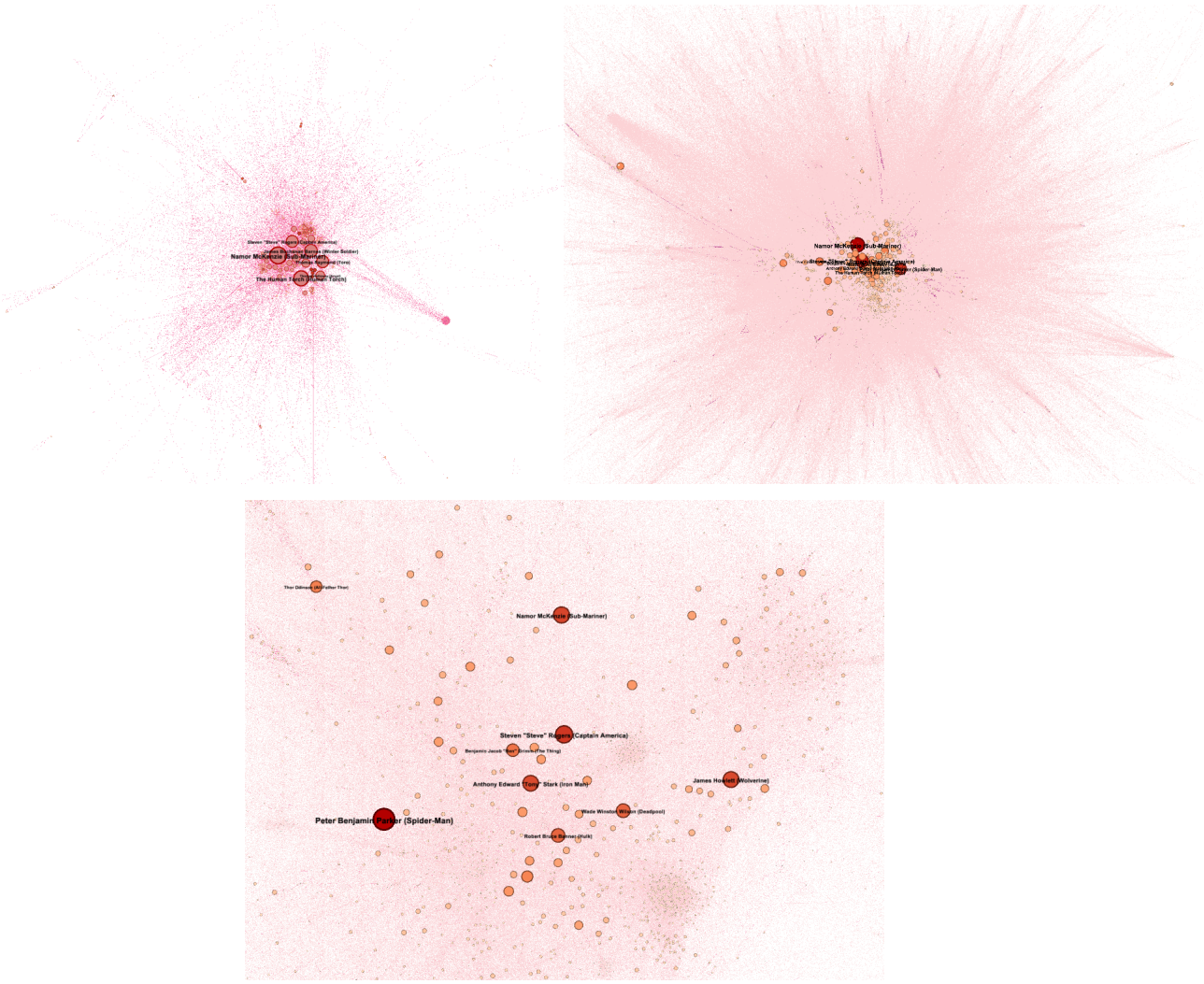


Figure 5: Evolution of the DC Universe (1950, 1990 and nowadays)

The graph analysis highlights a difference between Marvel and DC in their writing guideline: the first tends to focus on a team of characters and build networks between them, whereas the second seems to focus on characters by themselves and develop them as much as possible. This interpretation is confirmed by the the number of links in the Marvel network, that presents more than 1'200'000 links, representing more than 3 times the number contained in the DC network, which led to a higher average node degree (47.8 compared to 31.3). Moreover, Marvel's giant component's size is bigger than DC's (respectively 97.26% of the total characters and 94.76%), which again shows that it is a much more connected universe. Note that the actual network for Marvel goes largely beyond the displayed image, it however wouldn't be clear if all of it was shown.

5 Conclusion

The Marvel and DC universes represent very vast and complex networks, and it is thus essential to analyse them with the good approach to deliver a proper representation of them. It appears that using only one approach isn't the good way to go as each of them has its advantages and drawbacks. However, combining these different attributes with appropriate weights results in a much more accurate representation of the network. These representations are not only faithful to what we know about these universes, but also bring out new information about them: this is the power of graphs.

References

[1] Marvel Wiki Fandom
https://marvel.fandom.com/fr/wiki/Marvel_Wiki

[2] DC Wiki Fandom
https://dc.fandom.com/fr/wiki/Wiki_DC_Comics

[3] ForceAtlas 2 GitHub page
https://github.com/gephi/gephi/wiki/Force-Atlas-2

[4] Leiden Algorithm
https://www.nature.com/articles/s41598-019-41695-z

[5] OpenOrd GitHub page
https://github.com/gephi/gephi/wiki/OpenOrd

[6] PageRank
https://www.geeksforgeeks.org/page-rank-algorithm-implementation/