TEAM 48

Maximilien Gomart
Ahmed Kooli
Antoine Schmider

A Network Tour of Data Science

A GROWING NETWORK OF CHARACTERS IN MARVEL AND DC UNIVERSES

1) Story : motivation and choices

Since 1939, Marvel Comics have been telling stories about more than 65000 characters in more than 50000 comic books. Their main competitor, DC Comics, which has published around the same amount of comic books, has built a collective universe with more than 20'000 characters.

We can expect that all these characters won't be completely isolated from each other, and that there has to be some connections between them. Moreover, the writers started to create more and more common stories to the characters from different storylines over time. With this big amount of data, we expect to see some clusters appear between characters. An obvious one would be a cluster for *good* and *bad* characters. But there will probably be more subgroups that will appear depending on how we treat the data.

We will try to study the evolution of such networks, which even if they are fictive, can still be representative of real networks in our society. Constructing graphs will allow us to determine what characters seem to have the biggest importance for the Marvel or DC universes, i.e. the key characters to the success of Marvel and DC today. Moreover, being a bit cautious, we can predict what characters are most likely to appear in the next comics or movies, and what other characters will appear along with them.

2) Choice of data : origin and pre-processing

Our dataset comes from https://marvel.fandom.com/ and https://dc.fandom.com/, where all the Marvel and DC comics and characters are stored. The latter are stored along with their characteristics but most importantly their relatives (family, clones, duplicates, affiliations...). This will allow us to first construct connections between family-related characters. In addition to that, some characters have an "affiliation" to a company, a team, or another character, which will also allow us to define some relations based on that. However, this is not enough since if some characters always appear together but are not relatives or are not affiliated to the same team, they will not have any connection. This is why we will also use the comics as a dataset since we have all the appearing characters. Using both of these datasets (characters and comics) for Marvel and DC, we will be able to get a significant amount of information to construct our graph.

We can expect the parsing and the cleaning of the data to be a bit tedious as the data is coming from a Wiki Fandom: everyone can edit it and thus the writing format might not always be the same. Moreover, we will have to parse the comics one by one, as for the characters, which will take a lot of time since there are more than 100 000 comics and 80 000 characters in total. We've already started the parsing process and we are confident with the usability of the obtained data as long as we continue in the same way.