Introduction

The ***character universe*** has emerged as an essential driver of Hollywood cinema. In place of either the actor-star system or the series model, the network of interrelated characters is now a dominant way that large-budget content is produced. Unlike the series, the character universe relies on a more fluid and non-linear relationship between individual movies. Our goal in this project is to study the character universe model: how it functions and how it potentially distinguishes itself from more traditional series-structures as well as past pre-cinematic forms like novels. We use social network analysis as a way of modeling the relationships between characters across movies and books to better understand the forms that shape the social worlds of different kinds of media.

We base our analysis on four notable "franchises": the Marvel Cinematic Universe, the DC Extended Universe, the Harry Potter series, and Balzac's *La Comédie humaine*, one of the greatest novelistic character universes ever created. All of our data can be found here.

Studying these four worlds, we find that the character universe generates two unique features in terms of its social network. First, character centrality behaves differently—the characters with the most amount of connections are not always the most important characters to the plot. Second, a weaker network is created as individual characters cross from one story to another. Rather than follow each other as a unit, characters in character universes develop weaker ties by leaving their main plot and joining other stories. The resulting networks are less dense than serial models, allowing for more social complexity.

Data

	Marvel	**DC**	***La Comédie humaine***	***Harry Potter***
Works	20	9	87	8
Dates	2008-2018	2005-2018	1827-1847	2001-2011
Characters	347	231	1262	181
Men:Women	2.8:1	2.5:1	2:1	2:1
Total Edges	7,382	4,417	27,211	12,294
Weighted Edges	6,535	4,260	26,253	7,772

Table 1: Dataset information.

Methodology

Networks

The social networks on the following page reveal the structure of each universe. Nodes represent characters from the different franchises, and edges connect characters that appear in the same work. Each colour indicates a separate work and the connections ("edges") that are formed within it. Larger nodes point to more central characters (more on centrality, later). In order to generate our networks, we scraped the character lists for each film in the Marvel, DC, and Harry Potter franchises from their IMDb pages. After removing unnamed characters (characters with official titles, like "The Bloody Baron," were okay—"Barman #2" was not) and resolving name differences, the "nodes" of each universe's networks were defined as each of the remaining characters. (Cont'd...)

Figure 1: Marvel network

Figure 2: DC Network

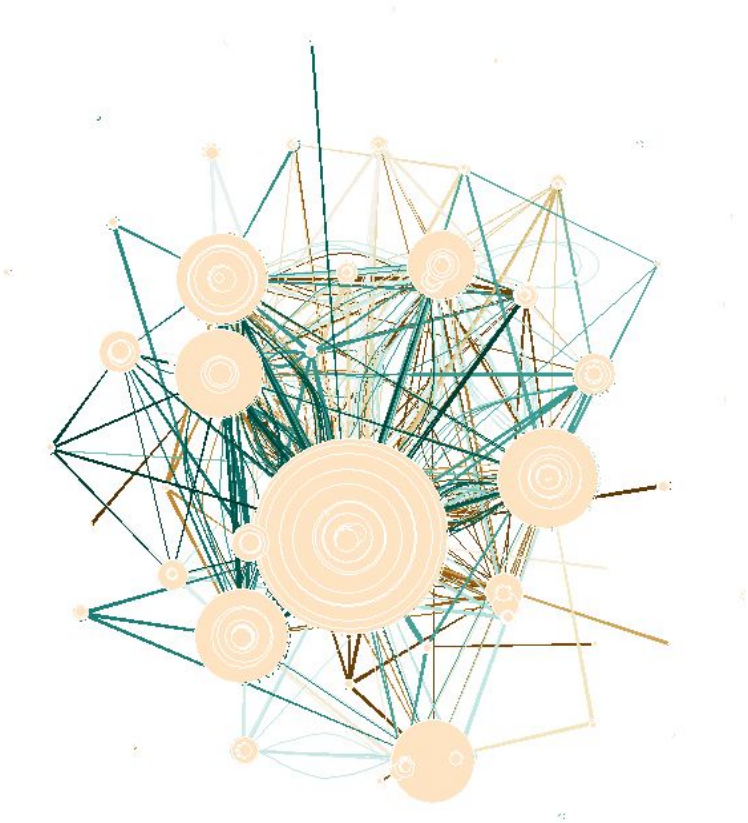


Figure 3: *La Comédie humaine* network

Figure 4: *Harry Potter* network

For *La Comédie humaine*, character lists were generated using a Named Entity Recognition (NER) tagger that pulled names directly from the source. While Balzac's overall work contains just under 100 finished and about 40 unfinished units, our corpus includes 87 texts after eliminating any work with fewer than 3 fictional characters.

We then generated edges for our graphs by connecting characters that appeared in the same movie or text. Finally, we collected data on several aspects of characters from Marvel and the DC regarding three categories: morality, superpower, and physical attributes (like facial hair). This allows us to study different kinds of power dynamics within social networks. We look at the relationship between gender and goodness, who vanquishes the most characters and physical objects, and whether hair is an indicator of moral direction (more on that later).

Limitations

Character extraction in French

While literary theorists have counted over 2,000 characters across *La Comédie humaine*, our numbers are lower for both character counts and the amount of works that major characters appear in. Named entity recognizers perform worse in French than they do in English, both missing instances of names and falsely tagging non-proper noun words as names. Our estimates are thus based on a smaller subset of detected characters.

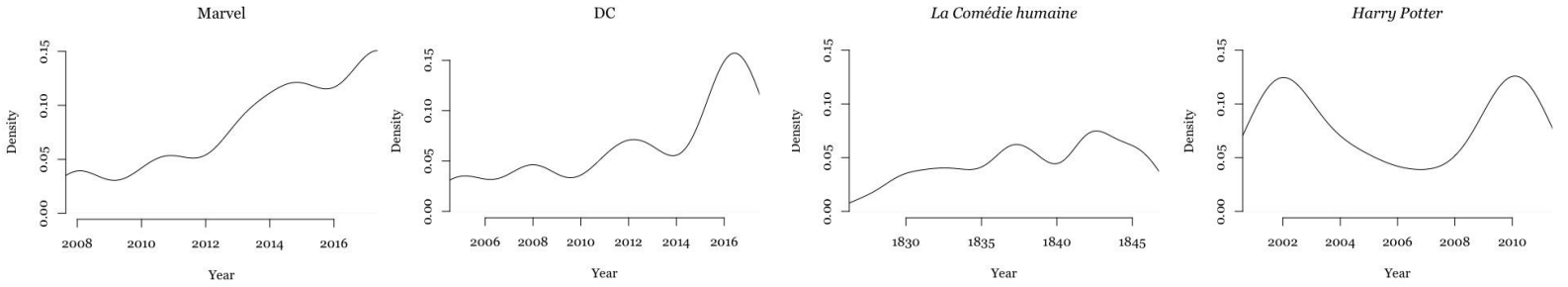
Comparability of Networks

Not only are there differing amounts of characters per work across the four worlds, but the amount of works that make up each universe also vary. In addition, Marvel and DC are not finished bodies of work and both have more films already announced for the coming several years. This is not the case for *La Comédie humaine* nor for *Harry Potter*. To control for universe size, we use sliding 20-work windows for *La Comédie humaine*.

Is the universe expanding?

Our four universes add new characters at different rates. Marvel adds more characters to its universe as time passes, while DC and *La Comédie humaine* show relatively stable growth (aside from DC's *Suicide Squad*).

On the other hand, *Harry Potter* introduces most characters at the beginning of the story, pointing to the importance of core characters to the series structure. The second peak, due to the 2010 release of *Harry Potter and the Deathly Hallows, Part 1*, perhaps indicates that the series may undergo renovation as they evolve before settling back into core components.



Figures 5-8: Characters added over time, per world.

The Stan Lee Effect

Centrality is a concept in social network analysis that makes it possible to understand the behaviour of the most important characters. *Degree centrality* depends on how many connections a node has—the more connections, the more central the node.

Across the three character universes, there is no significant difference in the average degree centrality. On the other hand, *Harry Potter* characters have a much higher average, meaning **characters in *Harry Potter* are more strongly connected to each other.**

> **In our three character universes, we also see an important overarching pattern: the most central characters are rarely the most central to the plot.** In Marvel, a character like this is Stan Lee (who cameos in every film); in DC, it is Alfred, Batman's butler.

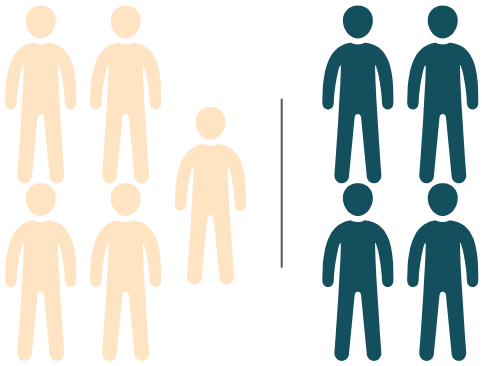
This is also a key feature of Balzac's universe. He consistently reintroduces minor characters to solidify the existence of the societal network and to create the impression of an intertwined world. Reappearing characters are often signaled as representatives of a given social group to which they belong[1]. By comparison, in the serialized model of *Harry Potter*, the most connected characters are Harry, Hermione, and Ron. Surprisingly, some minor characters manage to be equally ranked (Gregory Goyle), but still do not surpass the Golden Trio in terms of centrality.

Figure 9: Stan Lee in cameo roles across various Marvel films (ew.com)

[1] Laforgue, Pierre. *La Fabrique de La Comédie humaine*. Presses Univ. De Franche-Comté, 2013.

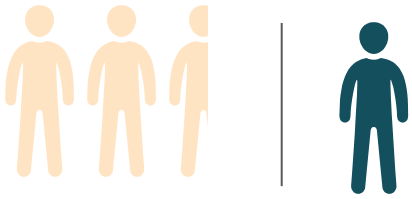
Is the bald, bearded guy always the villain?

In Marvel, not necessarily—there are...

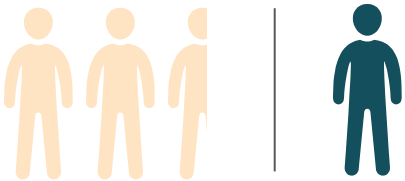


5 bearded and bald **good** guys for every 4 bearded and bald **bad** guys

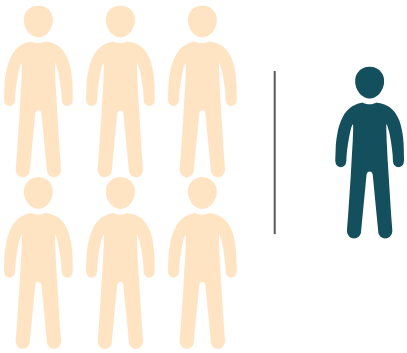
Similarly, out of bearded men, there are five good guys for every four bad guys. The classic trope does hold for the DC universe, though. Bad guys are...



2.5x more likely to be bald than good guys,



2.5x more likely to be bearded, and



6x more likely to be both.

How interconnected is the universe?

The contradiction of the most "central" characters that we saw in the previous section can be confirmed using a measure called ***graph density***, which tells us how many connections are made out of the total possible connections that could exist. In addition, ***vulnerability*** indicates how easy it is to break the network as a percentage of the nodes that need to be selective removed, starting from the most central, to break the network.

In *Harry Potter*, the majority of characters reoccur in multiple, if not all, films. By contrast, the most commonly occuring characters in the character universes only occur in a fraction of them—like Frédéric de Nucingen, who appears in 35 out of the 100 texts from *La Comédie humaine*. Consequently, **the *Harry Potter* network is almost five times denser than the densest character universe, DC**. In the series, almost everyone knows everyone else.

The Marvel network is the least vulnerable of the three character universes: twice as high of a proportion of characters need to be removed before the network breaks than in DC. The *Comédie humaine* network is broken from the start, but when zeroing in on only the biggest cluster of nodes in the network, only 1 character, Jean-Jacques Bixiou, needs to be removed to split the network. On the other hand, three times as many Harry Potter characters, compared to Marvel, would need to be removed before the network splits, revealing a much more strongly connected network than our three character universes.

These measures also reveal the power of *Infinity War*, often dubbed "the most ambitious crossover of all time." This Marvel film brings onto one screen multiple characters who would otherwise not have had the chance to meet. Ambitious or not, it certainly is effective.

> **When *Infinity War* is removed from the network, Marvel's vulnerability score drops to almost half of what it was**, and its density decreases as well, highlighting that the network gains strength from the film's presence and relies on the connections established in the film to strengthen its interconnectedness.

Are there subcommunities in the universe?

Louvain modularity is a measure used to determine communities that exist within a social network. A higher modularity score indicates that nodes separate off into groups, with denser connections within clusters than across them, forming "communities". **Marvel characters have significantly less of a tendency than DC and *La Comédie humaine* characters to split up into communities, but *Harry Potter* has the lowest modularity score of the four.**

This does not necessarily mean that the characters in the wizarding storyline are less likely to break up into communities socially (you can think of some off the top of your head—Hogwarts students, Death Eaters, etc.); rather, this is more an effect of the structure of the films. In Marvel and DC, communities approximate individual movies or subseries. *Harry Potter*'s lower modularity demonstrates how all characters occur together and interact, regardless of film.

Power Dynamics in Marvel

Can we **rank the strength** of superheroes quantitatively? By tracking how many objects and people a character defeated (on screen), and adding a defeated foe's power to the vanquisher's power level, we can determine who does the most damage in the Marvel universe. So when Tony Stark defeats Whiplash, we add Whiplash's strength to his.

- **Object** - 50 lbs
- **Person** - 200 lbs
- **Creature** - 400 lbs
- **Vehicle** - 10,000 lbs
- **Building** - 100,000,000 lbs

1. **Thanos** - 3,001,068,500 lbs defeated *(Infinity War)*
2. **Dr. Strange** - 2,600,013,850 lbs *(Dr. Strange)*
3. **Kaecilius** - 2,300,002,300 lbs *(Dr. Strange)*
4. **Tony Stark** - 800,685,550 lbs *(Iron Man)*
5. **Thor** - 500,605,400 lbs *(Thor)*

Iron Man wrecks the most vehicles in the universe, destroying **54 cars, robots, and other machines**.

Kaecilius demolishes the most infrastructure, wrecking **22 buildings** in just one movie.

Thor takes down the most **people** on screen—**144**.

Scarlet Witch is the strongest woman, and the **15th most powerful overall.**

Gender and Superpowers

Does gender influence your likelihood of having superpowers? Although there are far fewer women than men in Marvel and in DC, women are just as likely as men to have superpowers.

Are women with superpowers more likely to be good? In both Marvel and DC, women who have superpowers are just slightly more likely to be good than men (1.4 times more likely in Marvel, 1.1 in DC). These results exclude women and men who switch between hero and villain during their respective series (an example of a character like this is Loki Odinson from the *Thor* and *Avengers* films).

In Marvel, men are 1.3x more likely than women to change sides, but in DC, women are almost twice as likely as men to change sides. Moral inconsistency is more strongly attached to women in DC, indicating an interesting difference in terms of the movies' larger narratives.

Conclusion

When it comes to the social worlds of fiction and movies, character universes lack the density and centrality of traditional series. They are vulnerable and subject to fractioning off. This is happening because they do not follow the same cast of characters from start to finish, across one, cohesive storyline. Marvel, however, stands out among the three due to strategic crossovers that are not seen in the other universes. When *Infinity War* is removed, Marvel's vulnerability increases, its modularity score drops, and its transitivity increases, becoming more comparable to the networks belonging to the DC and *La Comédie humaine*. Notably, with *Infinity War* removed, Marvel becomes more strongly different from *Harry Potter*. That is to say, higher rates of "crossover" strengthen the network and lead it to behave more like a traditional series.

Balzac's goal in the construction of the *Comédie humaine* universe was to create a realistic depiction of society, which is why it was so vital to ensure that characters, who each represented different facets of society, would continuously re-appear. Future applications of this research could look to determine which kind of universe most closely approximates real-world networks or whether more fine-grained relationships like those established through dialogue or shot- or sentence-structure provide different portraits of the social worlds imagined on-screen or on the page.

A table with all social network analysis results can be found in the ***Appendix***.

Appendix

Movie Selection

The selection of films that are members of the Marvel Universe was done based loosely on the Marvel Studios owned films. While previous films such as *The Amazing Spider-Man* feature popular Marvel comic book heroes, the creators are no longer Marvel Studios. Non-Marvel Studios owned franchise films were not considered part of the MCU, excepting *Spider-Man: Homecoming*, which was a Marvel Studios and Sony collaboration. Consequently, Sony's *Venom* (2018) was excluded, as well as the *X-Men* and *Fantastic Four* films, owned by Fox. The motivation is to model only the project of Marvel Studios within this dataset.

In our dataset, the DCEU consists of two studio bodies: the recently created DC Films (est. 2013), which first produced *Man of Steel* and all subsequent DCEU films, as well as the Warner Brothers *Batman* trilogy and *Green Lantern*. DC Films is a subset of Warner Bros, and although *Batman* was recast between the two studios, no origin film has been announced for Ben Affleck's *Batman*, resulting in the inclusion of the original three films. While, arguably, *Green Lantern* has not been included in further projects such as *Justice League*, the goal of the project is not to streamline narratives but represent accurately a model for collaborative narratives.

Fantastic Beasts and Where to Find Them (2016) as well as *Fantastic Beasts: The Crimes of Grindelwald* (2018) are purposefully excluded from our *Harry Potter* data set. They, along with the 8 original films, form the *Wizarding World* franchise, which is, effectively, a character universe. The exclusion of these two most recent installments allows for the most accurate representation of what we refer to as a typical film series.

Results

	Marvel	**DC**	***La Comédie humaine****	***Harry Potter***
Density	12.2%	15.9%	9.62%	73.7%
Degree Centrality	42	38	33	136
Vulnerability	10.4%	2.1%	0.08% (1 node)**	31.7%
Robustness	69.3%	28.9%	49.3%**	96.6%
Community Detection	0.516	0.604	0.612	0.153
Transitivity	52.0%	78.4%	88.9%	67.9%

Table 2: Social Network Analysis Results

* These results are based on the averages of sliding 20-text windows.
** Based on the largest cluster of connected nodes in the *Comédie humaine* universe.