Community formation in Harry Potter

Mette Mortensen[a,1] **and Peter Nielsen**[a,1]

[a] Technical University of Denmark

This manuscript was compiled on December 14, 2022

Characters in the Harry Potter (HP) universe form a social network based on appearance proximity in the HP books. Communities in the network can be established based on text surrounding characters and their descriptions from HP Fandom. Given the groupings within the HP universe it is intuitive to assume similar communities would form for the social network. The graph communities and whether they align with the known groupings within the HP universe are unknown. The notion that Slytherins should inherently be evil has been around since the start, arguments for or against this are based on proofs of contradiction or founded in extreme character examples. Hence an analysis of whether or not there is a correlation between Slytherin and being evil based on their appearences in the HP books is required. Here we show that the communities found in the HP books in few cases align with the established groups within the universe, but in general the communities are formed around the narrative. We show that based on the text surrounding the characters, that Slytherins tend to have a lower sentiment than the rest of the characters, and also among the other Hogwarts houses. It was expected that the established groups would appear, what was found was one or two big communities for each book, which was more indicative of the main plot. Smaller communities were also found. It was expected that Slytherins would have an overall negative sentiment, their sentiment was found to be in the positive.

GitHub link: https://github.com/pete414n/02805SG

Social networks are used to describe how people interact or relate to each other. For this project the focus has been on the social network formed by the characters of the HP universe. This network may be formed by different approaches by linking between character pages on fandom sites, or as in the present case, by linking based on character appearances in the books. A challenge with the latter approach is that depending on the granularity that is chosen for character appearances the graph may become very dense or disconnected. The extreme examples being linking characters based on appearance in the entire series, resulting in a complete graph where every character is linked with each other, and the opposite linking based on two words, in which case the social network is very likely to be disconnected. There are many different groupings within the Harry Potter universe such as Hogwarts houses, wizard families, etc. From the narrative of the books it is made explicit which of these most of the characters that are encountered belong to. It is however not clear if the communities of the social network match these established groupings. As such this project aims to investigate if the established groupings will be reflected in the communities formed. Furthermore these communities may change throughout the books depending on which part of the text has been included, while the groupings remain the same by narrative, they may not appear in the communities. To this day it is still being discussed if Slytherins are where the evil wizards originate (1)(2)(3). At the time of this writing there seems to be no objective answer to this question. In this present work we want to investigate if there is some correlation between the sentiment of the Harry Potter houses. On expection of single characters such as Voldemort or Bellatrix Lestrange it is obvious that they are evil and also Slytherins. But there are also other characters such as Peter Pettigrew and Sirius Black who are at certain times considered good and at others evil, both of these are from Gryffindor. However looking at these isolated characters cannot tell us anything about Slytherins in general, even if they may be the source for the original notion.

Materials and Methods

For this project we have gathered information about the characters by using the Harry Potter fandom wiki page(4). We have also used Wikipedia(5) and Buzzfeed(6) to create a list of all the characters in the HP universe, although we have gone through this list manually and seen in the fandom pages whether a character actually appeared in the books, as many of them only appeared in a video game or a was mentioned on Pottermore (a webpage created by J. K. Rowling, not existing anymore). This was to make sure we only had characters in our network that were actually in the books.

We extracted the text from the fandom wiki page belonging to each character and from this we could find which aliases each character have, which we needed to find them mentioned in the books. We also found which parentage/blood-status and occupation each character has and which Hogwarts house they belong to, if any and who they are loyal to (Deatheaters, Dumbledore's Army or Order of the Phoenix). We have used all seven books to do our research in. The books were available online on OceanofPDF(7).

We have replaced all aliases in the books with the full name of the characters in order to easier look them up for our research. To do this we first created a dictionary with the aliases as the key and the full name as the value. Then for each book we iterated through the dictionary and asked to find all aliases alike and replace them with the corresponding full name.

Networks and communities. For each book we have created a network with the characters as vertices and edges for connection between characters. They have a connection if they appear in the same text piece. The edges have weights corresponding to how many times they appear together. The text piece we have chosen to check for connection is 5 sentences. We have tried to take different sizes of text piece, and found that this gives us the best network without

Significance Statement

Groupings in the Harry Potter universe and whether or not certain groupings of characters are good or bad have been a central discussion for the series throughout its existence. Through graph analysis of text pulled from HP-Fandom and the Harry Potter books, we have determined which communities form based on character appearances in the books separately and the series as a whole. Wordclouds have been used to see which groupings, if any, were represented by each community. Sentiment analysis has been used to investigate if characters from Slytherin are more evil than other characters.

[1] A.O.(Author One) contributed equally to this work with A.T. (Author Two).

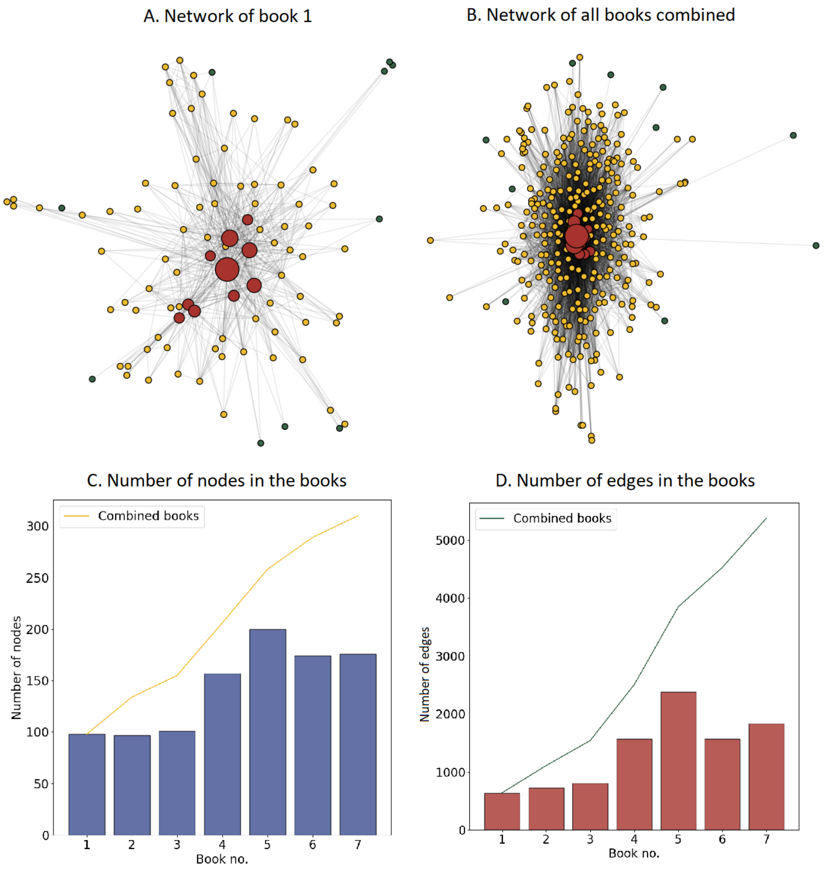

Fig. 1. A. shows the network of book 1. Each node color represents a size going between 50 and 800. We have then for each node found the weight sum and placed it in the interval by the equation explained in materials and methods. The red nodes are the 10 nodes with the biggest scaled weight sums, the green nodes are 10 nodes with the smallest scaled weight sums and the yellow nodes are those with scaled weight sums in between. B. shows the network of all the books combined, so the full network for the HP universe in the books. The node colors are as described for A. As can be seen there are many more nodes and edges in the network in B, it is much denser than the network in A. C. shows a bar plot with how many nodes there are in each book and the yellow line indicates how many nodes there are when the books are combined. This indicates that there aren't many new characters that are added in the story, and those that are replace some that only appears in one book, which we can also see by the line for the combined books, it increases a little for book 1-3, so new characters are added in the different books. D. shows a bar plot of how many edges there are in each book and the green line indicates how many edges there are in the books combined. As explained for C. we can also see that the amount of edges stays almost the same the same for book 1-3 but there's still an increase when combined, so new connections are added, while some don't stay in the books. Book 4 and 5 has the biggest increase in both amount of nodes and edges, which corresponds well with those books being some of the biggest in the series.

it being too dense. When drawing the network we first used the weight sum for each node as node size but this gave us some nodes that where too big. Instead we have used a scaled weight as the node size, so we have chosen a minimum and maximum size of the nodes, ensuring they would neither get too small nor too big. Thus, all nodes' weight sums have been scaled so they fit within those limits. We have also created networks where we add the network for one book at a time, so first we have network for book 1, then book 1 and 2, then book 1, 2, 3 and so on till all books are combined in the same network. This was too see how the network expanded over the timeline in the book series. We have also taken the 10 main characters and seen how their personal network evolve throughout the book series and plotted how many edges they have for the combined networks. We have created communities by using the Louvain algorithm(8). This have been done for the networks for each book separately and for the big network of all books combined. For each network we have taken the characters and found the text pieces in the book where they are mentioned and stored all the sentences. Then we have found the most frequent words and used TD-IDF to find the importancy of each word. The words have then been imported in wordclouds to better visualize the found words belonging to the communities.

Sentiment analysis. Sentiment analysis has been performed using VADER sentiment analyzer(9). Preliminary tests indicated that calculating based on LabMT1.0 (10) or using VADER to analyse the sentiment produced results with similar trends. The results found from LabMT1.0 was around a neutral value of 5 for the tested inputs, while the same text produced clearly positive and negative sentiments using VADER. VADER was chosen over LabMT1.0 because of this, and because it has a clearly defined neutral region for sentiment. The sentiment analysis has been performed by taking the average of sentiment analyzed on the concordance with a width of approximately 3 sentences for a character name. By doing this we on average ensure that the sentence that the character appeared in is included. We determined that 3 sentences was a good number of sentences experimentally.

Results

The number of nodes and edges for the network increases throughout the books, as can be seen in fig. 1. This coincides with the length and the complexity of the books. The first three books follow a similar relatively simple story pattern that is focused around Hogwarts, this is reflected in the number of nodes and edges, in that these numbers remain similar. The reason for the similar number of nodes, is that most of the characters remain the same e.g. students and professors. And those that change are replaced by someone else, one clear example is the Defense Against the Dark Arts professor, which is changed and replaced by a new node. If we look at the line showing the number of nodes and edges we can also see that the slope is biggest for book 4 and 5 which also are two of the biggest books in the series. Book 7 is also big, but as it is the last book most of the characters in this book have already been in the previous books, therefore this book don't add many new nodes or edges to the combined network. If we look at the network drawn for just book 1 combined with the network for all the combined books we can see that it gets much denser, tying all characters more together.

We have also investigated for what we consider as the ten main characters how their networks evolve. This is shown in fig 2. We can see that especially Harry Potter, Ronanld Weasley and Hermione Granger gets more and more connections throughtout the books. Harry Potter starts with being connected with 93 characters and end up being connected with 299 characters. Ronald and Hermione both starts with around 60 connections and end up with around 230 connections. The character with the biggest increase in connections percentwise is Sirius Black. He starts with having 2 connections in the first book and end up with 133 connections, this is an increase in percentage of 6550, which means he has gotten a network that is 65 times bigger than in the beginning of the whole story.

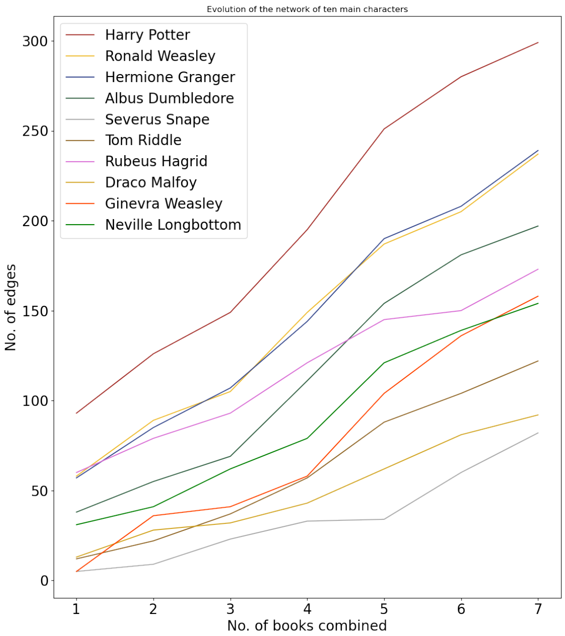

Fig. 2. Plot showing the evolution of ten main characters from the HP universe. All characters gets a bigger network throughout the book serie and we can see that Harry Potter, Ronald Weasley and Hermione Granger are the ones with the biggest networks.

For each book we can see that the communities forms after the story of the book, so those characters that appears much together are in the same community. The communities thereby changes a little for each book, as the stories changes, but for all the books the main characters always ends up in the biggest community. For most of the books we also have a community which have all the Dursley family members and other muggles and we often have a community with many of the Weasley family members. We have investigated the communities formed in book 5, Order of the Phoenix a little further as this is one of the books where the sentiment score was most interesting. For this book we have created a list of all the words belonging to each community and, as described in Materials and Methods[1], used TF-IDF to weigh them properly so we could put them in word clouds. The word clouds for the communities can be seen in fig. 3.

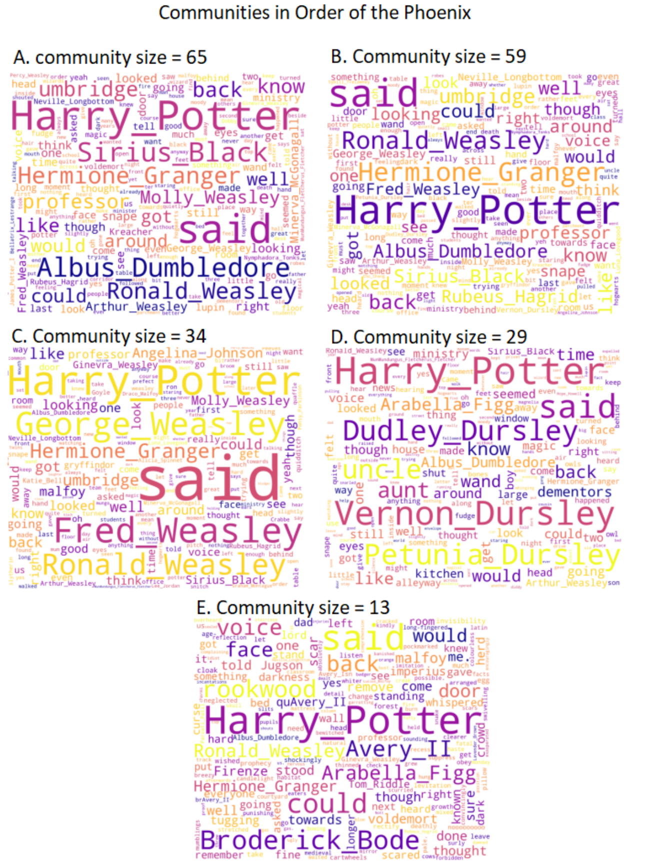

Fig. 3. Word clouds representing the communities formed in book 5, Order of the Phoenix. In A. and B. we see that Harry Potter is mentioned often as well as Hermione and Ron, which is the three main characters. Other wise there is not much that represents a pattern to why they are in the same community. In C we can see many Weasley members are mentioned, so this is a community with a lot of focus on the Weasley family. In D. we see the Dursley family are mentioned much, so they are a big part of this community. In E. we again have some very different characters mentioned which don't really seem to have an obvious connection.

We found that Slytherins had an overall lower sentiment than the remaining characters as can be seen on figure 4A. Similar analysis was made for the other houses as can been seen on figure 4B-C, which showed that the other houses had an overall higher sentiment when compared to the remaining characters. The average sentiment for the full text of all of the books was also calculated for Slytherin, Gryffindor, Ravenclaw, and Hufflepuff. Here we saw a similar trend with sentiment values of 0.016, 0.11, 0.23, and 0.079 for the four houses respectively. Hence all of the houses had a positive average. Upon investigating the five characters with the highest and lowest sentiments, we found that in the highest sentiment all were Gryffindors. In the lowest five sentiments there were two Slytherins, two Gryffindors, and the Slytherin house ghost.

Discussion. The network of the HP universe increases a lot throughout the books as we would have expected. In the first 3 books it does not evolve much, which fits well with those 3 books being smaller than the others and are books that mainly are to get us into the universe and learn about all the characters, Hogwarts and the whole magical world. Then in book 4 and 5 the network evolves much more, which goes well with the books, they are some of the biggest and there is a lot going on in the stories in those books. For book 7, there is still some evolution of the network, but not as much, but this is also the last book, so here all loose ends must be tied up and the whole story ends, so it makes sence that there

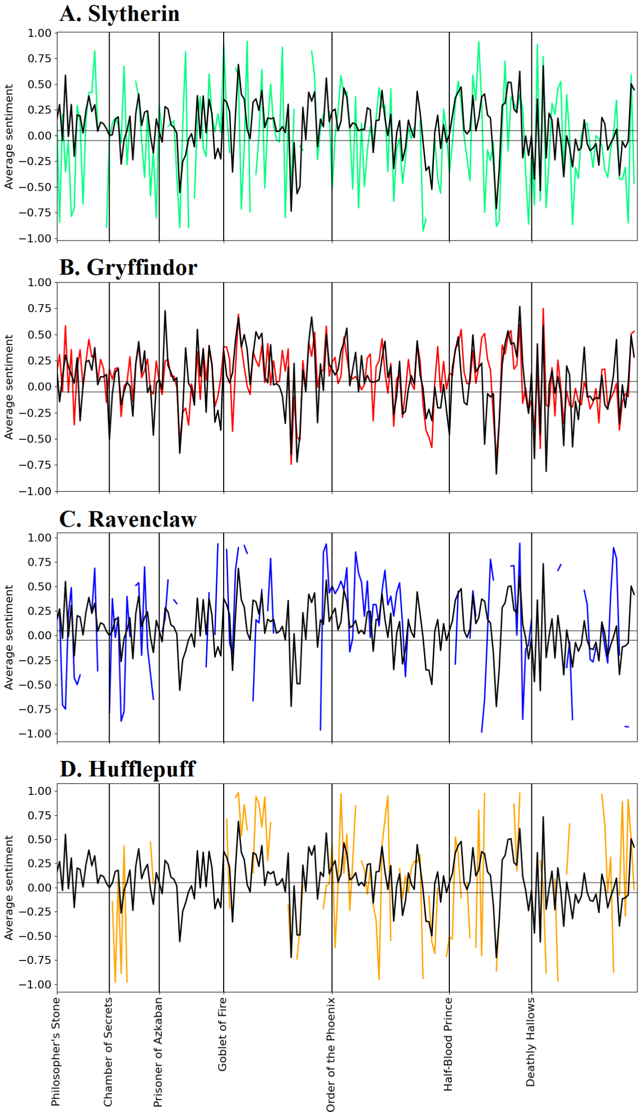

Fig. 4. Average sentiments for houses throughout the books compared to the average sentiment of the characters not in that house, the two horizontal lines in the plots denote the neutral region for sentiment, and the vertical lines denote the start of a new book. A) The average sentiment of Slytherins in green and the average sentiment of all other characters in black, the sentiments for Slytherins has a greater variation than the average, and more of the values fall below the neutral section B) The average sentiment of Gryffindors in red and the average sentiment of all other characters in black, Gryffindors have little sentiment below the neutral region in general. C) The average sentiment of Ravenclaws in blue and the average sentiment of all other characters in black. D) The average sentiment of Hufflepuffs in orange and the average sentiment of all other characters in black.

aren't a lot of new characters and connections presented in this. When looking at the communities they did not divide as we had expected. We had thought they might divide in the different houses the students belong to at Hogwarts and with the professors for themselves, or that there would be a community that consisted of the more evil wizards but this was not the case. The community corresponded much to the story in each book, and as mentioned the biggest community always had the main characters in it. For most of the times we also had a community with many of the Weasleys together and with the Dursleys together, this is what would be expected. But we can't conclude much from our communities and the corresponding word clouds. Slytherins did have a lower sentiment than the remaining characters, and the overall development in Slytherins sentiment throughout the books tends to be lower than neutral, it may not be sufficient to say that Slytherins are more likely to be evil wizards. Since the books are written from the perspective of Harry Potter the impressions we get of the other characters are seen through his eyes. Hence those that are his friends may get a higher sentiment and opposite for those that are not his friends. Throughout the books Draco Malfoy, Vincent Crabbe, Gregory Goyle, and Severus Snape are all Slytherins that have negative interactions with Harry. For the first three these interactions are not necessarily indicative of evilness. Furthermore Hogwarts has a House Cup, which creates negativity at various points. Besides these in-universe caveats, it is also evident from figure 4 that it is only Gryffindor that is close to being fully represented throughout the books. The other houses have many gaps, with Hufflepuff having close to no representation in both the first and the third book.

To mitigate the low representation of characters from Slytherin, and characters in general, one could try extending the text included. Including dialog and character description from various HP video games could give more text for the different characters, and include new characters.

1. Reddit: Are all slytherins evil? (2022).
2. Quora: Why do members of the slytherin house almost always turn out evil? (2022).
3. More than just evil? a study into slytherin (2022).
4. Various, Harry potter fandom wiki (year?).
5. Various, Harry potter wiki (2022).
6. Various, Buzzfeed harry potter (2022).
7. Harry potter on ocean of pdf (2022).
8. Louvain algorithm (2022).
9. Vader sentiment analyzer (2022).
10. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter (2011).