IE 532-Analysis of Network Data

Fall 2023

Relationship Extraction and Network Analysis of the Mahabharata epic

Prathmesh Rathod, Sumeet Jagtap, Pradyumna Bada

Introduction

For our IE532 Final Project, we set out to find a research topic harmonious with our personal passion project, based on Natural Language Processing(NLP) and textual analysis. Thus to combine Network Analysis and the theories we have discussed in class with NLP, we have selected the project, "Relationship Extraction and Network Analysis of Mahabharata,". This project is aimed at dissecting intricate webs of interactions within a text, by conducting relationship analysis, to derive insights about the most central and important characters in the text, as well as the changes in their centrality and importance across various chapters.

Conducting network analysis on relationship extraction in a story primarily offers three significant insights. Firstly, it facilitates a deeper understanding of character dynamics and interactions, enabling the identification of central characters who are crucial to the progression of the narrative. This insight is particularly valuable in complex stories such, where numerous characters influence the plot in various ways. Secondly, such analysis sheds light on the narrative structure, revealing the intricate interplay of different story arcs and character interactions that collectively shape the overall plot. This helps in understanding not just the storyline, but also the structural craftsmanship of the narrative. Lastly, network analysis aids in thematic exploration by examining the nature of relationships between characters. This approach can uncover recurring themes and motifs, offering a nuanced understanding of the underlying narrative elements that might be less apparent through conventional reading. Thus with relationship analysis we can obtain a more comprehensive view of the complexities within a storybook.

The text in question, the Mahabharata, is an ancient Sanskrit epic of profound complexity and narrative depth, which we thought would pose an intriguing and challenging subject for our application of Network Analysis and NLP. It contains a dense narrative structure, involving a vast array of characters and multifaceted storylines. The general outline of this project is that first we will use NLP techniques to conduct textual analysis and obtain character name data, which will then be cleaned. This data will then be analyzed through network analysis methods, particularly focusing on aspects such as centrality, to identify pivotal characters and their roles within the narrative framework.

Tools and Libraries

For the NLP tools we decided to use Spacy, an open-source software library for advanced NLP. Spacy is designed to work with and process large volumes of text. It contains a host of NLP features, including some important ones such as lemmatization and Named Entity Recognition(NER), which easily picks up entities such as locations and names, which is useful for our context, and does so in a manner that is not too complex.

For the network analysis portion we decided to stick with Networkx, which we have familiarized with through the duration of the course. We also played around with Pyvis for visualization portions, together with Pandas for creating dataframes which provides interactive visual network graphs.

Methodology

First we sourced a set of Mahabharata chapters in a text format. Then we opened the first chapter and called Spacy to read through it, using the NER feature to pick out important entities.

Following the entities picked up, we conduct data cleaning and methods such as fixing a “window” or a set of sentences, and then picking up the entities found within each window. We will then compare the entities picked up to a list of characters that we input, to extract only character names from the entities identified. We will create relationships between these names picked up then conduct our network analysis on it.

The application of centrality metrics is crucial in revealing the key figures who drive the story's progression and their influence within the network, to understand the dynamics and significance of these relationships in the context of the narrative.The anticipated outcome is a comprehensive understanding of the Mahabharata's narrative structure.

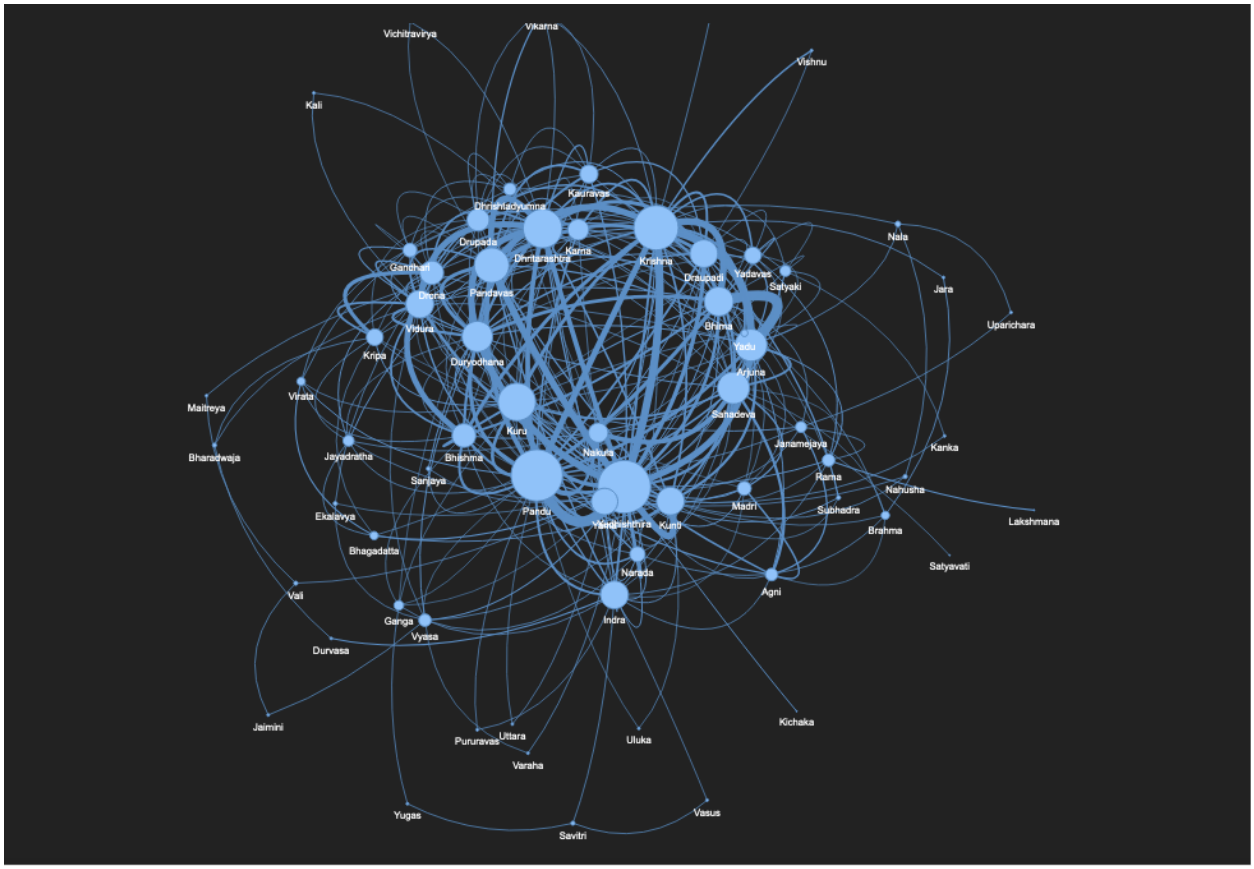

Fig.1: Using Pyvis to visualize degree of characters

We conduct centrality analysis to help us judge the importance of characters. If we define character importance by a. Involvement and b. Influence, we can use the Degree Centrality, Betweenness Centrality and Closeness Centrality theorems to help us visualize it.

Degree centrality(Fig.3) is a measure of the number of connections a character has in the network. In our context, it would represent how many other characters each character interacts with. A high degree centrality indicates characters who are central to the story, involved in many relationships or interactions.

Betweenness centrality(Fig.4) measures the extent to which a character lies on the shortest paths between other characters in the network, and reflects the importance of a node in connecting different parts of the network. In this context a high betweenness centrality indicates characters who act as bridges between different groups of characters, influencing the flow of information or

events. These characters play key roles in connecting different story arcs. This could be useful information to gain more insights on the leanings or motivations of characters and their actions.

Closeness centrality(Fig.5) measures how close a character is to others. Characters with high closeness centrality are close to many others in terms of relationships. This could suggest characters who are influential or have a quick impact on the narrative due to their proximity to other characters.

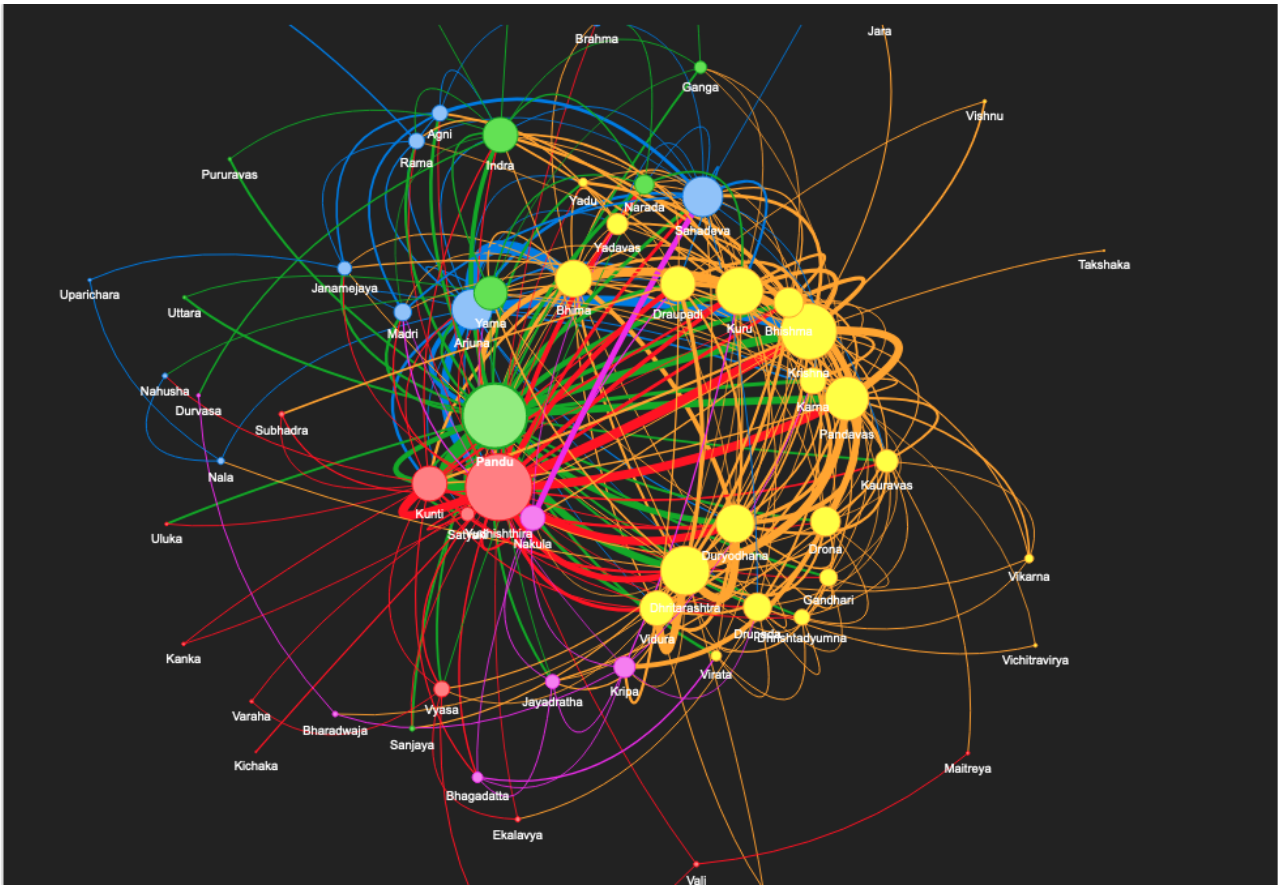

Fig.2: Using Pyvis to visualize Community partition of characters

We also make use of the Community Detection technique(Fig 2) to see how characters can be fit into different communities and how these interact with each other, using the louvain detection algorithm. Different communities are marked by different colors. In the above community graph, Yellow color represent one community, blue color represent another community and so on.

Finally we run the centrality results and iterate them over all the chapters of the text to observe changes as the plot progresses.

Results

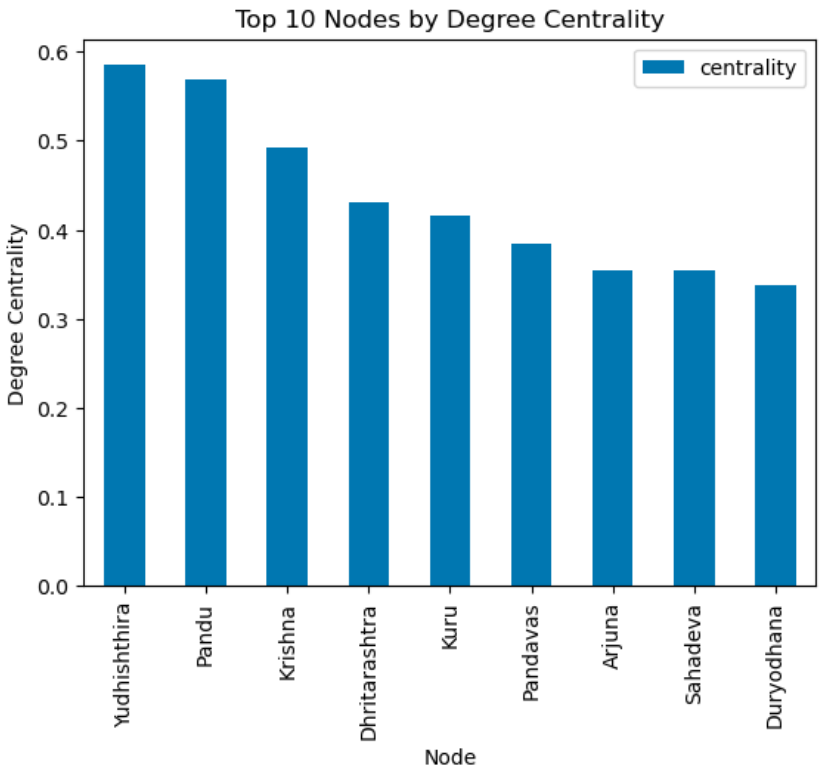

Fig.3: Degree Centrality of characters

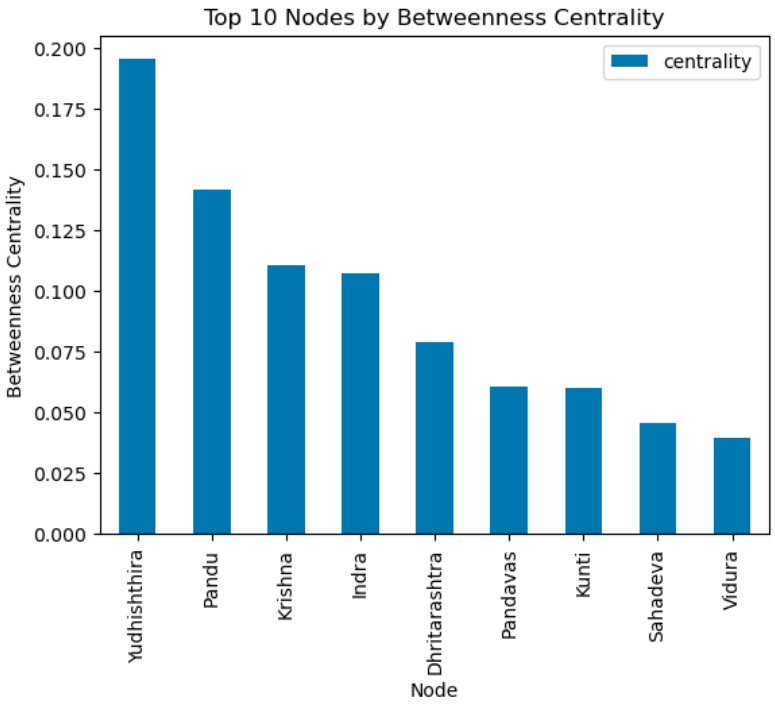

Fig.4: Betweenness Centrality of characters

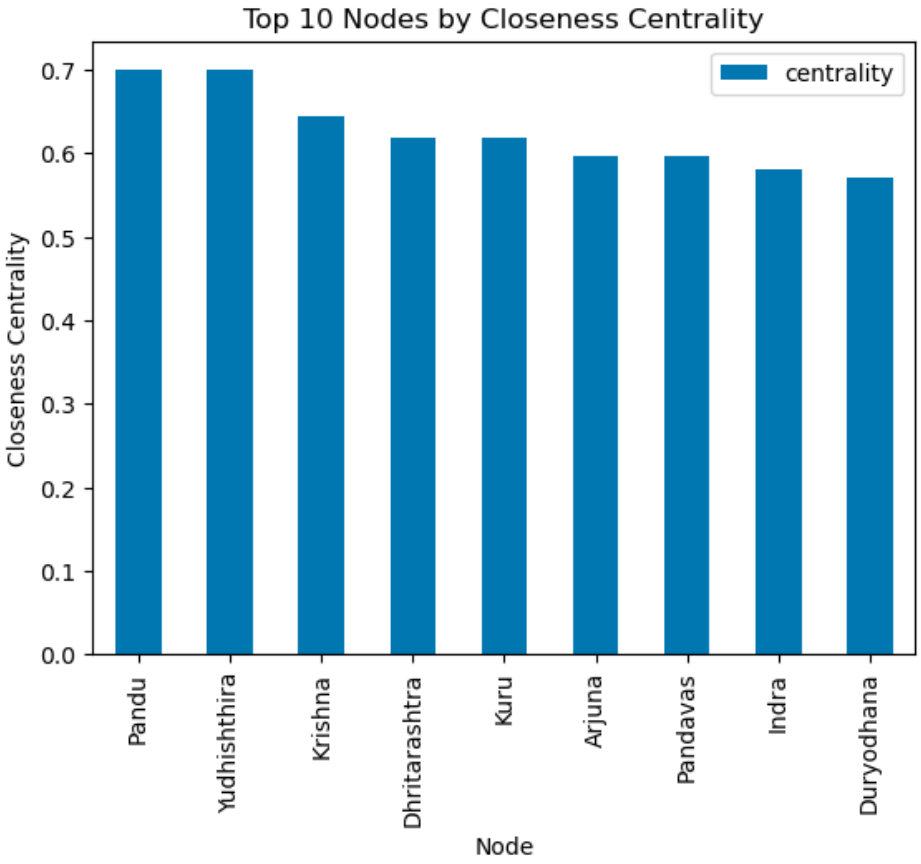

Fig.5: Closeness Centrality of characters

We observe from the Centrality analysis that Yudhishthira and his father Pandu, are the 2 most well-connected and influential characters. The high connectedness of Pandu is logical - he is the biological father of the Pandavas (Yudhishthira, Bhima, Arjuna, Nakula, and Sahadeva). His role as a father sets the stage for the central conflict in the Mahabharata, and he is significant in the birth of the Pandavas and the early events in the story. Yudhishthira's character is central to the narrative due to his actions that snowballed into the main conflict.

Thus as we analysed 10/15 chapters, we logically obtain the high degree characters that set the context for the story. This indicates that our analysis is on the right track.

We further run some Community detection steps, using the Louvain algorithm. Community analysis helps us identify different groups and alliances in the narrative, like the well-known division between the Pandavas and the Kauravas. It also uncovers power dynamics and hierarchies, highlighting influential characters within specific groups. The louvain algorithm is chosen because it is well-suited for large numbers of characters.

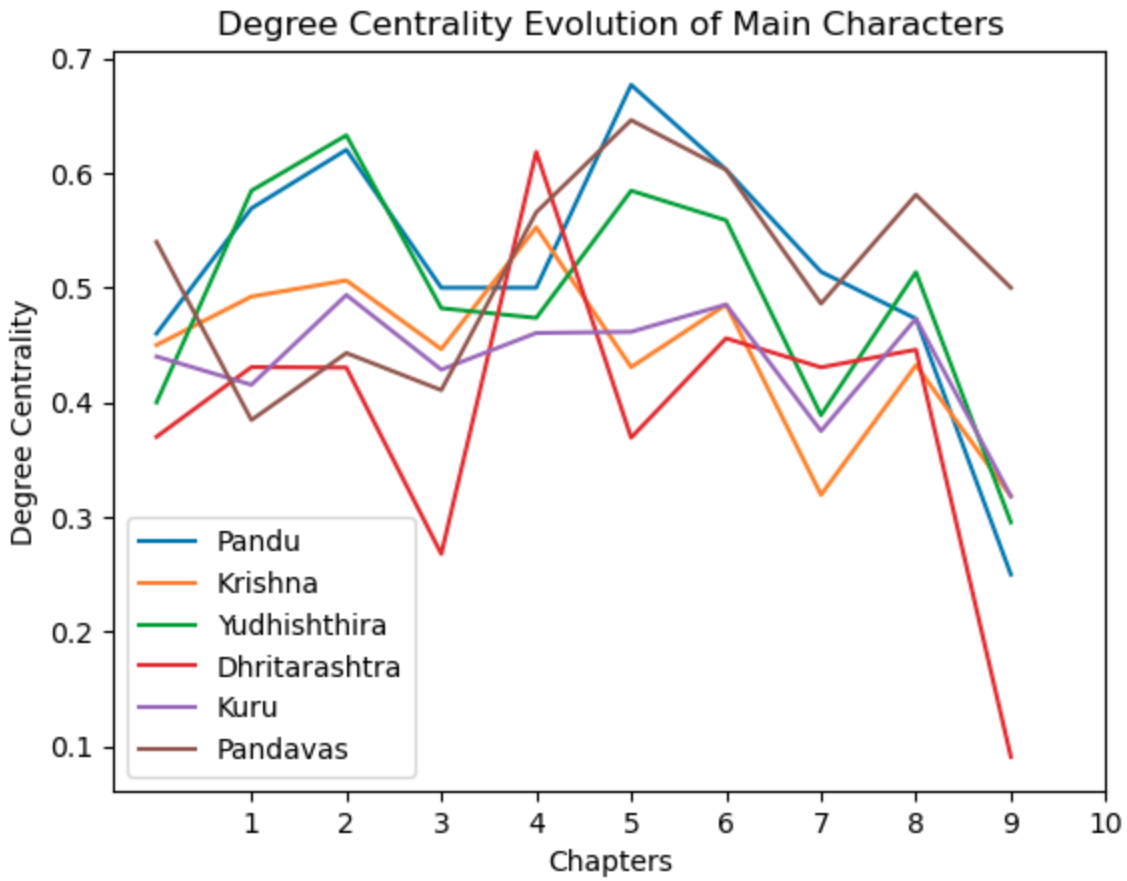

Fig.6: Centrality Evolution of main characters

Finally we ran a full iteration of degree centralities of the most important characters(based on the earlier centrality plots) across 10 chapters(reduced from original full 15 for computational ease) and analyzed their trends.

An interesting observation is that the trendline for Arjuna, one of the protagonists, closely matches the trendline of the Pandavas (Arjuna is part of Pandavas) at times, which highlighted a possible limitation of the study where we have considered groups, which may cause redundant relationships when looking at individual characters from these groups.

It is also commonly assumed that Arjuna is the protagonist in Mahabharata hence he should be one of the most important characters, however based on our current analysis we see that he ranks outside the top 5. All the character degrees of the top 5 also seem to drop drastically from chapter 8 onwards, indicating a canon event in the storyline.

Areas of Improvement and Future Enhancements

Some of the entities we included were groups, such as the Pandavas or Kauravas or other tribes. This causes redundancy when analyzing relationships of the individuals contained within the groups. Perhaps it would be beneficial to conduct a separate cluster analysis purely dedicated to groups, for instance to analyse the strengths of the 2 key groups in the Mahabharata, the Pandavas and Kauravas, across the epic.

Currently our methodology involves analysing character relationships based on a window of 5 sentences. However we believe that this could be relooked, as it is not necessary that mention of characters in close proximity necessarily indicates a relationship. Moreover, multiple accounts of a particular interaction could skew results.

We also would like to add sentiment analysis to our project, as it is another key feature of understanding relationships and analysing information flow networks. These improvements would involve a significantly deeper understanding of the workings of the text as well as more advanced NLP and Network analysis techniques on our part.

Due to computational restrictions, we could only run the analysis on 10 out of 15 chapters, which might have skewed the results. Running iterations across the full text with all the characters might help us better visualise the centrality trends of all the characters and their influence at each stage of the plotline.

References

mahabharata/books at master · kunjee17/mahabharata · GitHub
https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.louvain.louvain_communities.html
https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/