Malazan Network Analysis

Christopher Peralta

Friday, Apr 26, 2019

Abstract

In this project, I attempted to visualize the co-occurrence data of characters in the series "Malazan Book of the Fallen". This series is notable in that it is one of the most complex and long fantasy series with a continuous single plotline. There are 3,252,031 words in the series and somewhere between 1,186 and 1,479 characters in the series with 457 unique points of view. Additionally, many of the characters have multiple aliases and nicknames adding another layer of complexity. A character might go by completely different names in different novels.

To mine the character co-occurrence data, I started by converting the novels to `.txt` file and reading them into R. Then I added book numbers, chapter numbers, and stripped the front and back matter. From there, I turned the text into a series of ngrams of length 20 split by book and chapter. At this point, I wrote a function that extracts the characters' names from the ngrams by row and then puts the co-occurrence data into a workable format. Cleaning and formatting the co-occurrence data was the next step. Finally, I used a variety of methods to visualize the text and the network data.

The base network graph was quite convuluted due to the sheer number of characters, so I use a few different methods to make it into a readable graph. **ADD CONCLUSIONS**

Introduction

The goal of this study was to analyze the network data of the Malazan series, and to see if there were any interesting conclusions. I used numerous datasets from several sources in this project. The co-occurrence data was mined from the books by me. The books were converted from `.epub` format to `.txt` format. I made most of the `alias` data manually and crowdsourced some of aliases on Reddit. The `name` data was manually extracted from the *Dramatis Personae* sections at the start of each book and manually extracted from the Malazan Wiki.

Method and results

Assumption

- The partial names were shorter in length than the full names. ## Method I'll begin this section with a detailed description of the methods and assumptions I used in mining the character names from the novels as that was the most difficult part of the project.

I began by joining the character name data with the alias data into a single dataset. I then split all of this data by spaces in order to get variations of the names and rejoined the partial name data back to the full name data to get a comprehensive dataset of full and partial names. I then filtered out stop words, formal titles, and commonly capitalized words that aren't names from this list. Then, I arranged the list by character length.

At this point, I went back to the book data and turned the text into ngrams of length 20 by book and chapter. I then wrote a function that tries to extract the first name in the name list from the string and tries with every name in the list for every name and every ngram. This method was incredibly computationally intensive as there were 3,080 names in my name list and 3,250,530 ngrams. I changed my code a bit, added some code for parallelization, and broke my data into 263 chunks; all of this managed to get my code to run in around 40 hours on my laptop.

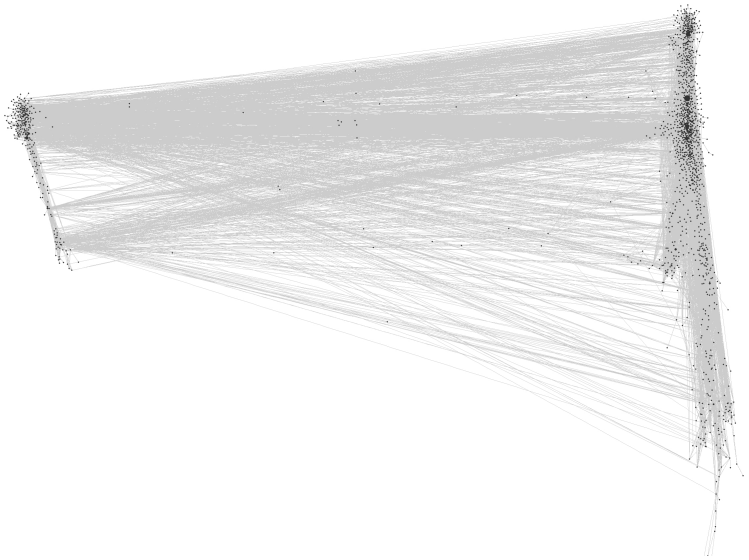
Once I got the co-occurrence data into a usable format, I had a lot of data cleaning to do. I formatted and removed variations of all names with over 100 appearances in the network data. I used regular expressions to achieve this. I made two important assumptions at this point:

- The names with over 100 occurrences in the co-occurrence data are representative of all co-occurrence relationships.
- The most common variations of names accurately represent the co-occurrence relationships of their specific character.
- Any extremely uncommon name variations will be filtered out by

After cleaning the data, I began to analyze the co-occurrence data.

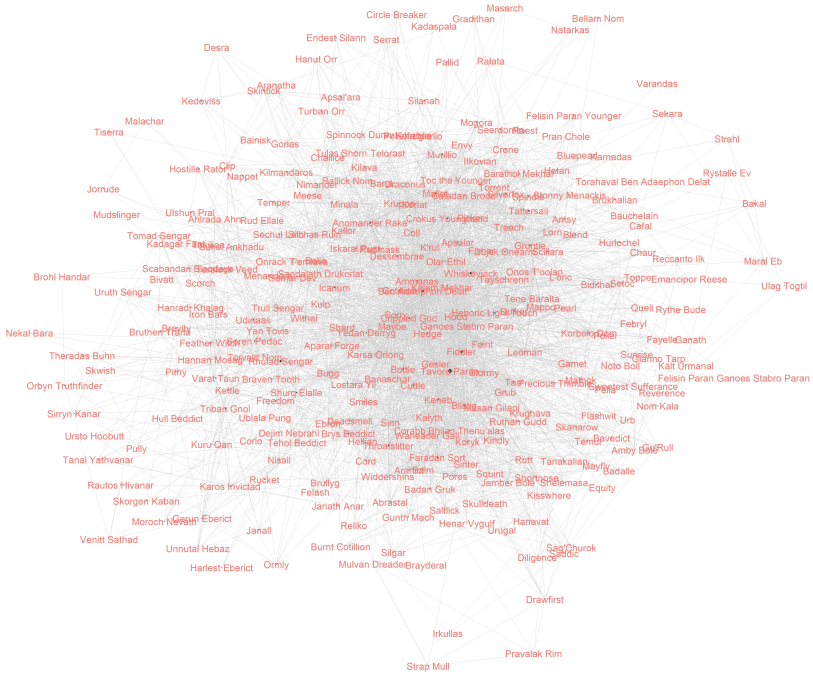
Results

I'll start with showing the complete network graph:



The only interesting observation I can make from this graph is that there is a clear distinction the two main continents from the 4 first books in the series and the third main continent introduced in book 5.

Let's look at a graph with only characters who are given points of view in the series. I'll also remove any nodes with under 10 edges.



This is much better and

Bibliography

`https://www.reddit.com/r/Malazan/comments/a1ukxk/main_series_character_pov_data/`