A Character Network Construction for Macroanalysis

Morgan Condello, Ross Harrison, Jennifer Isasi, Alex Kinnaman, Ashanka Kumari
University of Nebraska-Lincoln

Abstract

We present a methodology that can be repurposed for future character network experimentation at a macro or microanalytical scale. This methodology combines and extends previously established character network studies. Our method locates, labels, and associates characters simultaneously to build individual networks for novels within a corpus. This study is applied to a macroanalytical survey of third-person nineteenth century novels to determine what data can be extracted about character and character networks using a combined method.

I. Introduction

Until recently, the primary research on “character” focuses on the characteristics of specific characters in a novel at an analytical level. Characters are generally analyzed through close reading and measured against each other from novel to novel, defining certain common characteristics (i.e., genre, author nationality, context). Recent studies, however, have been analyzing “character” from a statistical standpoint (e.g., Moretti, Sack, etc.). Upon reading and researching these scholars, we noticed the focus continues to be at a microanalytical level, and we became interested in trying to understand characters on a macroanalytical level by examining character networks.

Our initial interest was to focus on distinguishing unique characteristics of nineteenth century character networks, particularly the author’s nationality and gender, as well as how the character networks of these novels might evolve over time. The nineteenth century corpus we had access to contained specific information on author gender, author nationality, date of publication, and point of view, thus we began to focus on how to retrieve more data from this corpus. Combining previous methodologies, we created a method for extracting data. Specifically, our method employs preprocessing to extract network data. Though we originally wanted to observe a single facet (i.e., author gender, nationality, or genre) of nineteenth century novels, the true significance of our work is our process rather than what we have observed from these findings. Our work only scratches the surface of what information could be extracted through our procedure.

II. Background

Network and character space analysis have been approached by several digital humanists on many different levels, such as determining narrative and textual space, measuring dialogue, defining characteristics, and other experiments to extract networks from novels. Much of the recent work has focused on measuring and extracting social networks through character relations and attention given to characters. Franco Moretti’s “Network Theory, Plot Analysis” provides much of the foundation of our project. He describes the Network Theory as connections between objects in order to quantify plot ([Moretti (2011)]). Objects can be people, places, and things and are referred to as nodes connected by interactions called edges. Moretti applied the idea of operationalizing character space and aimed to directly measure it by analyzing its physical aspects: words, sentences, lines, and pages ([Moretti (2013)]). He measured how much space was dedicated to each character, and then measured the weight...
of character relationships based on the frequency of co-occurrences, an idea that we have incorporated in our methodology. Moretti’s interest in character space derives from Alex Woloch’s work ([Woloch (2009)]) in the field, which is the second most important foundational work in our experiment. Woloch theorizes that each character is given a certain amount of space within the novel. Character space is the attention each character receives from the reader, and a character system refers to the arrangements of multiple character spaces within the novel. Woloch’s idea of character space has heavily influenced our process because we theorize that character space defines a network.

Other works that are significant in our ongoing research revolve around character interactions and how those interactions form into character networks. Elson et. al. initially theorized that average “face to face” time between any two characters diminishes as the number of characters in the novel grows ([Elson, et.al. (2010)]). Relationships are measured through dialogue between characters ([Elson, et.al. (2010)]), so Elson et.al. constructed a social network by chunking together character names and testing a separate corpus for quoted speech attribution. This was significant to our experiment because interactions between characters form the character networks, and while we did not focus on dialogue, this is another method for representing character space.

Sack’s perspective on character relations is more theoretical and examines the positive and negative properties of character relationships rather than just characters’ text space, i.e. the amount of attention given to each character in a novel. He proposed that character relations within a novel’s social network greatly impacts the outcome of the narrative. In a triangular relationship, each node can represent characters or places, and each relationship can be either positive or negative ([Sack (2013)]). Character relationships define the network’s stability. This theory breaks down the social network into its relationships between character in order to establish how stable or unstable the overall network is and how that can affect the plot.

We were also interested in examining how narrative attention was divided up within nineteenth century novels, and Sack’s work became part of the foundation of ours. Sack also studied plot and characterization by examining generative models. Sack conducted “an inspection of the way narrative attention accumulates over the course of a novel” ([Sack (2011)]). Sack discovered that characters marked “narratively important” ([Sack (2011)]) in the beginning of the novel are still important by the end of the novel. He also found that characters change throughout the novel in that characters who appear to be minor in one part of the book may become major later on. Sack’s computer simulation model generates first, a character hierarchy for each strand consisting of main, supporting, and incidental characters, and second, a random plot sequence ([Sack (2011)]) to create a consistent model of how narrative attention is allocated in novels.

In addition to scholarly work on the theory of character networks, we also researched previous on computational studies. Park et. al. computationally extracted a social network to analyze its dynamics by finding the frequency of each character, which was the basis of the creation of our character networks. Characters are always interacting between Park’s definition of “text space” (i.e. the words, sentences, paragraphs, etc.) so a network can be formed by observing the closeness of one character. The textual proximity of one character to another is based on whether the names appear in the same sentence. Combined, a multidimensional social network is formed showing not just who is connected to who but also how “close” they are textually. This is the foundation of the parameters we set for our networks, which are described in our methodology.

Originally we were interested in examining specific aspects of character networks to determine trends in the corpus. These included the evolution of character networks over time, difference in author genders and nationalities,
as well as the significance an author’s choice of using a pronouns versus the use of a character’s name. However, as we developed our methodology, we became more interested in the potential of our work rather than the end result. What we have done here is establish a strong methodology for future application.

III. Methodology

Our process consisted of three stages: preprocessing, locating characters, and associating characters.

I. Preprocessing

Park, et. al. relied on predefined lists of characters from each novel for identifying character occurrences. This is not an option for our investigation due to our corpus size. Obtaining and verifying character lists for hundreds or thousands of novels is intractable. Instead, we utilize the Stanford Core Natural Language Processing (NLP) with Name Entity Recognition (NER) to create an XML document. The document provides each sentence its own ID, tokenizes each sentence, and predicts the named entity value. Possible named entity values include: PERSON, LOCATION, and ORGANIZATION.

II. Character Location

After the text has been tagged, characters identified by the NER value as “PERSON” are associated with individual tokens. Multi-token names (e.g. “Tony Stark”) are identified by consecutive appearances of “PERSON” NER value. As the XML is processed, each unique character name is recorded along with the sentence they occur in. This method of character location gives rise to the “Bennet Problem” as described below.

III. Character Association

The final stage, associating characters, is the most important. First, the set of found characters is limited by a minimum mention count. The **minimum mention count** is defined as a percentage of the most mentioned character’s count. For example, if Paul Atreides is mentioned 500 times in *Dune Messiah* and the minimum mention count is set to 0.10, then only character mentioned more than 50 times will be considered. These characters are the nodes in the network. Next, the occurrences of every character are compared to every other character. Statement distance is used to determine the weight for these occurrences. For example, if John is at sentence index 1,375 and Jane is at sentence index 1,377 then they have a statement distance of 2. A maximum statement is used to determine if the co-occurrences should be considered.

Once the statement distance has been determined, the weight function can be applied. As in Park et. al., we are using a power function to devalue a relationship as the statement distance increases. The function is defined as

$$w(d_s) = \alpha^{d_s}$$

where $w$ is the weight, $d_s$ is the statement distance, and alpha is a constant between 0 and 1 that we define. If alpha is 1 then statement distance is not punished. If alpha is 0 then only statement distance 0 will count. More practically, if alpha is 0.5 then the weight of the relationship is halved with the statement distance: $w(0) = 1, w(1) = 0.5, w(2) = 0.025, w = .125$, and so on.

The total link weight is then the sum of these weights. In analyses where the existence, or non-existence, of a link is important then another filter can be applied. The **minimum link weight** is defined similarly to the minimum mention count, as a percentage of the maximally weighted link. If a link is less than this minimum it is thrown out of the data.

IV. Corpus

We began with a corpus of 3,515 first and third person nineteenth century novels. We then created a corpus of one hundred novels, choosing ten novels from each decade of the 1800’s to analyze. However, due to the results obtained from the different metrics established to
compare the networks and the lack of a stable variance throughout the years, we soon realize that one hundred novels was not enough data from which to retrieve accurate information. We therefore decided to expand our corpus to all of the third person novels in the original 3,515 novel corpus. Thus we filtered our corpus to 2,978 third person novels.

IV.1 The “Bennet” Problem

Throughout our experimental stage, we encountered several problems in our coding as well as in theory that challenged our initial plan. First, and most unexpectedly, using Stanford’s NER as our primary tool to recognize characters via their first or second names, we noticed that it was not going to give us a precise account of mentions through a text. The most notable example was found in the network for *Pride and Prejudice*. We realized that there was a node accounting for a total of 322 mentions: Bennet. We somehow expected this result because throughout the novel there are five sisters referred to as “Miss Bennet,” their mother is referred to as “Mrs. Bennet,” as well as their father who is referred to as “Mr. Bennet.” Without the context of their first name, the program would tag every “Bennet” as the same character and thus merge their character space into one. By close reading the novel it is evident which Bennet is being referred to. However, the resolution of this issue is beyond the scope of this project. Thus, we decided to use the Stanford NLP or part-of-speech tagger as it provides a better clustering of names and surnames. Though it did not solve the problem, for most of the references to the characters are done by their honorific rather than first name, we obtained a more accurate mention count, reducing the node for all Bennets to 301 hits. In spite of this family’s surname forming their own separated node, other surnames also formed different nodes: Bingley (reduced from 250 to 109), Wickham (192 to 183), Collins (180 to 179), etc.

Along with these two obstacles to our research, we found two other minor issues that we decided to disregard. First, we ignored the nicknames (Lizzy, Kitty, etc.) in *Pride and Prejudice* since trying to merge both nodes (first name + nick name) would account for a micro-analysis rather than the broad understanding of character networks that we are trying to study in this present research. Second, Miss. and Mr. and Chapter would be classified as characters. When we saw any of these nodes for Miss, Mr., or Chapter, we simply brushed them aside in order to maintain our attention on the broad spectrum of nodes and their connections.

V. First Person Text

With the assistance of Dr. Matthew Jockers at the University of Nebraska-Lincoln, we were able to create a point-of-view detector ([Jockers (2014)]) to narrow down our corpus, and manually test a sampling of this corpus to ensure the detector was identifying points of view correctly. We decided to focus specifically on third person novels because it was difficult to recognize the narrator in a first-person novels as single characters, because the pronoun “I” would generally refer to the main character while narrating the story, as well as this character speaking to any other character who would also use this pronoun in dialogue. This situation would result in a single node accounting for all the first person singular pronouns with inaccurate demonstration of the main and central character of a novel, such as in a Bildungsroman. We also chose not to measure pronoun use in our corpus, because we are creating networks on a macroanalytical scale, so it is not possible to attribute pronouns to specific characters with a corpus this size. Taking this into account, we interpret the character space within a novel based on the mention count of character names.
tinctions that we investigated: author nationality and author gender. The two metrics used to investigate these dichotomies were character count and total link weight.

Total character count was determined by counting all characters who occurred more than the minimum character count. As stated before, the minimum character count is actually a percentage of the maximally occurring character. For our analysis, we used the value 0.05 for the minimum character count.

Total link weight is then defined by the sum of all links between the characters in the resulting set. For this measurement three variables are important: alpha, maximum statement distance, and minimum link weight. Like Park, et. al., we decided to go with 0.7 for alpha, because the power function decreases at a subjectively pleasing rate. The maximum statement distance was chosen to be 10, again somewhat arbitrarily. The minimum link weight was set at 0, because the (non)existence of a link does not affect our analysis only the total weight of the entire graph.

I. Author Nationality

The first question we addressed was whether it would be possible to find a significant difference between British and American novels in terms of character network. As seen in figure 1, novels by British writers had a visibly higher total link weight while having a similar overall character count. Nonetheless, figure 2 shows the fact that the British outliers could point towards a difference between British and American authors in this corpus. However, while intriguing, it would be premature to jump to conclusions. Our corpus is not equally distributed between nationalities, and a more balance dataset would need to be obtained before making conclusions about novels from both sides of the Pond.

II. Author Gender

Gender also presents an interesting division of the novels. While the British-American division is lopsided, the gender occurrence is much more balanced. Looking at the same metrics from before we found an interesting pattern.

In figures 3 and 4, novels authored by female writers have marginally fewer characters, they also appear to have larger total network link weights. The differences are not substantial, but it is worth noting that female authors in this corpus appear to be creating smaller
networks (in terms of the number of nodes), but also networks with stronger connections. Further studies must be conducted in order to determine whether these results are due to the number of character mentions or character closeness.

VI. Future Work

Our methodology was designed in order to be repurposed and reused for further experimentation with character networks. We focused on third person novels, but similar research on other point of view as well as comparisons among points of view are also possible. For our research we focused on American authors versus British authors, but comparing other nationalities to one another and to our current work could lead to analysis and conclusions that could extend beyond social networks within the novels that parallel the social structure of a particular nation. In the future, this methodology can be used to study how author genders may affect the social networks of their novels. For example, examining if female authors write novels that are more connected in character networks while perhaps male authors write a more individualized novel. Using the same techniques to look at gender roles in later corpus, such as 1870-1970 to observe the changes would also be a valuable insight into changing gender roles.

I. Microanalysis

A microanalytic application of these techniques would also be valuable. Observing individual authors, sub-genres, and narrow time periods could provide interesting insights into those series.

A particularly compelling example is George R.R. Martin’s *A Song of Ice and Fire se-
ries. The narrative in this series involves a set of limited third person point of view characters. That is, the language is third person, but the information available to the reader is limited to the senses of the point of view character. This series is also interested because of the volume of text, and the complexity of relationships.

II. Time Dimension

Currently the networks are constructed with the novels as a whole in mind. This technique could also be used to look at snapshots of novels to determine the state of the network at any given time. Park et. al. discussed this as a goal for their work as well. This could be done by picking a statement location in the novel, and then using anormal distribution to weigh the weights.

III. Statistics

The statistics used are basic. The number of links weights, and number nodes are not particularly letting statistics. More interesting network statistics and manipulation could be possible. For example, the distribution of link weights, mention counts, and node centrality could be more descriptive of the character of the network.

REFERENCES


