A Methodology for Character Networks at the Macroanalytical Level

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

Introduction and Background

We present a methodology that combines different techniques borrowed from computer science and the humanities in order to analyze character networks in a corpus of 1,800 British and American novels dating from 1849-1899 from a statistical standpoint.[1] This methodology consists of three stages: preprocessing, locating characters, and associating characters. Using this technique, interactive visualizations of character networks are also produced.

Franco Moretti, Alexander G. Sack, G. M. Park, and other humanists have presented different approaches to network analysis. While these studies increase our understanding of character networks, we noticed these analyses tend to focus on smaller corpora or individual novels. Using their studies as our starting point, we try to understand characters at a macroanalytical level. After testing our data and drawing some conclusions regarding authors' gender and authors' nationalities at a broad level, we have created a reliable methodology that can produce character data and be applied to future in-depth studies of character networks.

Methodology

Our process consists of three stages: preprocessing, locating characters, and associating characters. With a corpus of 1,800, third-person, 19th century novels, we began our project by interpreting the *character space* (Woloch 2009) within a novel based on the mention count of character names. We utilize the Stanford Core Natural Language Processing (NLP) with Named-Entity Recognition (NER) to create an XML document in which characters are identified by the NER value as "PERSON." As the XML is processed, each unique character name is recorded along with the sentence it occurs in, which allows us to identify each character as a node in the resulting network.

To associate characters, we take into account the *minimum mention count* (a percentage of the most mentioned character's count). 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 these characters have a statement distance of 2. A *maximum statement distance* is used to determine if the co-occurrences should be considered.

Once the statement distance has been determined, the weight function can be applied (see **Figure 1**).

[1] This research began in a class taught by Dr. Matthew Jockers and has continued as a project in the Nebraska Literary Lab at the University of Nebraska-Lincoln.

$$w(i,j) = \alpha^{dist(i,j)}$$

Figure 1: As in Park et. al., we use a power function to devalue a relationship as the statement distance increases. Here, *w* represents the weight, *dist(i,j)* is the statement distance, and alpha is a constant between 0 and 1, where 0 gives more detailed networks.

The *total link weight* is then the sum of these individual weights among all occurrences of characters *i* and *j*. The *minimum link weight* is defined as a percentage of the maximally weighted link. If a link is less than this minimum it is thrown out of the data.

The association of all the mentioned weights results in the construction of a fairly accurate network of characters in a given novel. **Figure 2**, for instance, shows the character network of Charles Dickens' *Oliver Twist*. In this case, Oliver is the most mentioned name, but it is not entirely a central node, because smaller characters do not share a link to it.

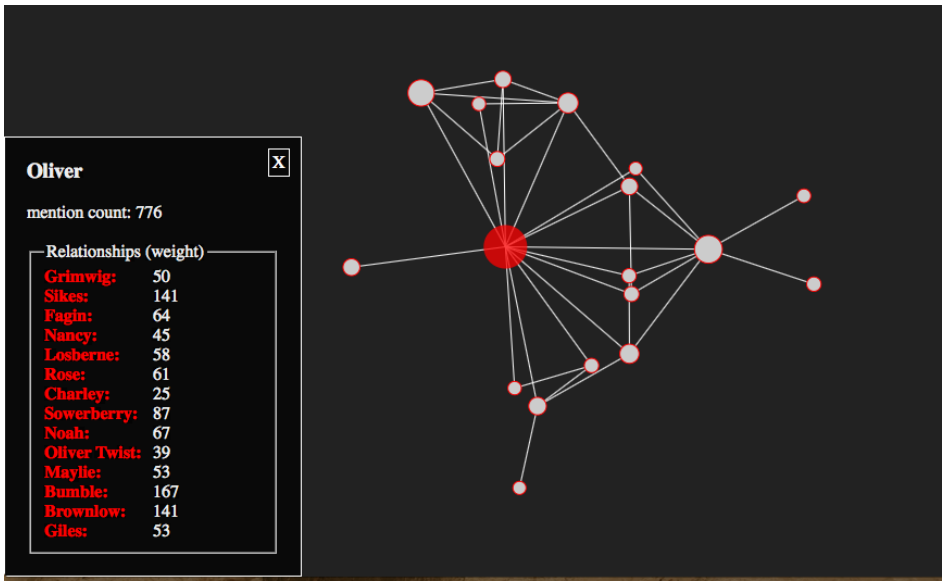

Figure 2. This is the character network we extracted for Charles Dickens' *Oliver Twist*.

Observations

Because data about authors' gender and nationality was readily available to us, we investigated these two binary questions:

1) Would it be possible to find a significant difference between British and American novels in terms of character network (**Figure 3**)?
2) Is there a noticeable division between works written by female and male authors (**Figure 4**)?

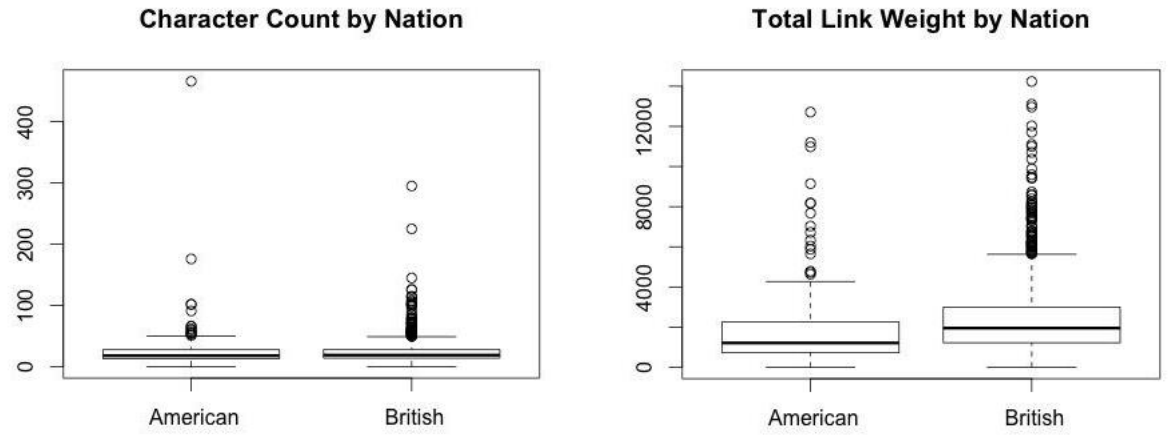

Figure 3. Novels by British writers have a visibly higher total link weight while having a very similar overall character count.

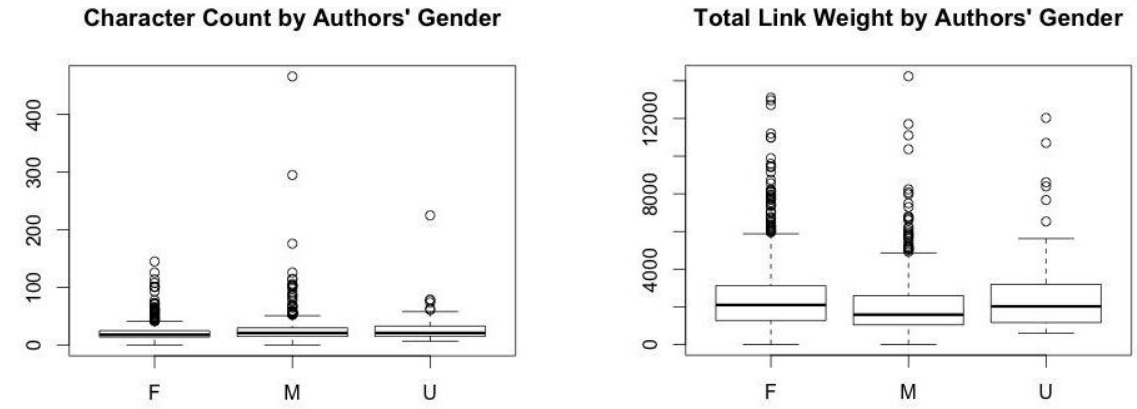

Figure 4. Novels authored by female writers have marginally fewer characters and larger total network link weights.

Conclusions and Further Research

It is difficult to determine the success of our technique without an objective measure. However, a subjective viewing of a random sampling of networks not only revealed the problems described above, but also seemed to capture the relationships appropriately. We were able to extract significant data from novels and create visual networks of the worlds within a given book. For that reason, we believe that the methodology presented here could be repurposed to carry out further research such as: individual authors, evolution of networks over the timespan of a narrative, the significance of an author's choice of using more pronouns than names, a study beyond social networks within the novels to the social structure of a particular nation, and more.

Bibliography

Elson, D.K., Dames N., McKeown, K.R. (2010). "Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138-147.

Hochman, B. (1985). *Character in Literature*. New York: Cornell University Press.

Jockers, M. (2014, Apr. 6). "Simple Point of View Detection." Retrieved from http://www.matthewjockers.net.

Moretti, F. (2011). "Network Theory, Plot Analysis." From *Pamphlets of the Stanford Literary Lab*.

Moretti, F. (2013). "Operationalizing, or, the Function of Measurement in Modern Literary Study." From *New Left Review*, (84), 103-119.

Park, G.M., Kim, S.H., Cho, H.G. (2013). "Structural Analysis on Social Network Constructed from Characters in Literary Texts." From *Journal of Computers, 8(9),* 2442-2447.

Pierson, E. (2014). "Parsing is Such Sweet Sorrow." Retrieved from http://fivethirtyeight.com/features/parsing-is-such-sweet-sorrow/

Sack, G.A. (2011, Nov.). "Simulating Plot: Towards a Generative Model of Narrative Structure." Paper presented at *2011 Complex Adaptive Systems: Energy, Information, and Intelligence Conference,* Arlington, Virginia.

Sack, G.A. (2013). "Character Networks for a Narrative Generation: Structural Balance Theory and the Emergence of Proto-Narratives." Proceedings from *2013 Workshop on Computational Models of Narrative*, M.A. Finlayson, B. Fisseni, B. Lowe, J.C., Meister (Eds.).

Shurkin, J.N. (2012). "Using Social Networks to Analyze the Classics." Retrieved from http://www.insidescience.org/content/using-social-networks-analyze-classics/747

Woloch, A. (2009). *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press.