Long paper submission

Roel Smeets
Radboud University Nijmegen, Department of Literary and Cultural Studies

Community Detection in Present-Day Dutch Literary Fiction

Introduction

Recent years have increasingly witnessed the use of social network analysis for the study of fictional storyworlds (Alberich et al 2002; Stiller et al 2003; Elson et al. 2010; Lee & Yeung 2012; Agarwal et al 2013; Jayannavar et al 2015; Karsdorp et al 2015; Lee & Wong 2016: Grayson et al 2016). By means of social network analysis, such studies have attempted to lay bare social dynamics between characters to gain insight into, for instance, the centrality of certain characters as opposed to others (e.g. Van der Deijl & Smeets 2018; Smeets et al 2019). A key aspect of these social dynamics is community formation. As the clustering of characters into distinct communities indicates how individual characters are represented as being part of a collective identity, detecting these communities can enable a deeper understanding of the extent of integration or segregation between characters in fictional storyworlds. Breaking down social networks of characters into separate communities, furthermore, makes it possible to quantitatively operationalize a seminal literary theory by Mikhail Bakhtin (1929), who coined the term 'polyphony' to assess how homo- or heterogeneous literary narratives are. In this presentation, I will argue that the extent of polyphony (or multivoicedness) of fictional storyworlds is partly the result of the communities characters function in.

As part of a larger research project on the social dynamics between characters in present-day Dutch literary (e.g. Van der Deijl et al 2016, Smeets et al 2019, Volker & Smeets 2019), two data-driven models were developed to analyze communities in fictional storyworlds. Applying these models on a dataset of 2137 characters in a corpus of 170 present-day Dutch novels, this paper demonstrates the technical challenges of detecting communities in narrative fiction. As such, it aims to contribute to the elaboration of network analytic techniques for the study of narrative fiction, as well as to the integration of these techniques with literary theory.

Methodological approach

In order to detect communities of characters in each single novel within the research corpus, two alternative approaches were explored. Both models were applied to the same dataset, which consists of extensive metadata (gender, descent, education, age) on 2137 characters in a sample of 170 novels from 2012 written in the Dutch language,[1] which constitutes 36,9% of all original Dutch language fiction published in that year). For each of the texts in the corpus, social networks of characters were extracted with the co-occurrence approach described in

[1] All data and code for this paper can be accessed through https://github.com/roelsmeets/character-networks.

Smeets et al 2019. Python's Networkx library was used in both of the models described below.[2]

1. Community detection algorithms

Community detection in social networks can be done with a range of state-of-the-art algorithms that group nodes in separate clusters based on statistically significant cut-off points. As nones of these algorithms are built specifically for analyzing character networks, it is a challenge to find the most suitable algorithm for detecting communities in literary texts. Experiments with two of the most popular algorithms – the Clauset-Newman-Moore greedy modularity maximization algorithm (Clauset et al 2004) and the Girvan-Newman algorithm (Girvan & Newman 2002) – yielded negative results. There are least two explanations for this: 1) the character networks in the corpus are too small (on average, 12.6 characters were identified), 2) the characters networks are too dense (with a mean density of 0.46). With relatively few nodes in relatively dense networks, it appeared to be impossible to find meaningful clusters of characters.

In order to find a solution to this problems, I distinguished between two clusters, e.g. of equal size, by optimizing a separation criterion. The Kernighan-Lin algorithm (Kernighan & Lin 1970) bisects a network into two clusters by 'iteratively swapping pairs of nodes to reduce the edge cut between the two sets'.[3] Although this results in only two communities for each novel, it seems to be the only feasible way to group the character networks into separate clusters. As such, two communities of equal size were thus detected for each of the 170 novels in the corpus. After that, repeated measures ANOVAs were conducted to statistically test the extent to which these communities are segregated by gender, descent, education, or age, in order to get a sense of how heterogeneous, polyphone, or multivoiced these communities are.

2. Homophily scores

Whereas the first model breaks down the character networks into communities and then computes gender, descent, education, and age diversity within these communities, the second model departs from the similarities between any two nodes. A central observation in sociological research on social networks is that communities tend to consist of members with a similar background (Marsden 2011: 599). This mechanism is studied through the concept of homophily, 'a principle of social organizing defined as people sharing similarities tending to have more social interaction' (Seidel 2011: 382). So-called 'assortativity coefficients' (or homophily coefficients) were computed for gender, descent, education, and age for each novel in the corpus, each of which indicates the extent to which characters from a certain demographic category tend to flock together. In order to determine how normal or peculiar

[2] https://networkx.github.io/documentation/stable/index.html, last accessed 5-3-2020.

[3] The documentation of this algorithm can be found here: https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.kernighan_lin.kernighan_lin_bisection.html#networkx.algorithms.community.kernighan_lin.kernighan_lin_bisection, last accessed 14-2-2020

the mean assortativity coefficients (or mean homophily scores) for the corpus are, a permutation test was conducted to create a baseline against which these homophily scores were compared.

Results

Gender was in neither of these models put forward as a cause for segregation between characters. In both models, however, descent and age were suggested as being a cause for divides. The second model based on homophily scores also suggests that education has a statistically significant effect on segregation. Following these results, it can be argued that present-day Dutch literature stages a divide between migrant and non-migrant characters, older and younger characters, and higher and lower educated characters, but *not* between male and female characters. In this presentation, these results will be contextualized in light of Bakhtin's concept of polyphony.

References

- Agarwal et al 2013 – A. Agarwal et al., 'Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland.' Online, 2013. http://www.cs.columbia.edu/~apoorv/Homepage/Publications_files/MAIN.pdf
- Alberich et al 2002 – R. Alberich et al., 'Marvel Universe looks almost like a real social network.' Online, 2002. https://arxiv.org/abs/cond-mat/0202174v1
- Bakhtin 1929 – Bakhtin, Mikhail, Problems of Dostoevsky's Poetics. Trans. C. Emerson, Minneapolis (University of Minnesota Press): 1984 (originally published in 1929). In: The Bakhtin Reader: Selected Writings of Bakhtin, Medvedev, Voloshinov (ed. Pam Morris), London (Arnold): 2003.
- Clauset et al 2004 – Clauset, A., Newman, M. E., & Moore, C. "Finding community structure in very large networks." *Physical Review E* 70(6), 2004.
- Elson et al. 2010 – D. K. Elson et al., 'Extracting Social Networks from Literary Fiction 2010.' In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (2010), pp. 138–147.
- Girvan & Newman 2002 – M. Girvan and M. E. J. Newman, 'Community structure in social and biological networks', *Proceedings of the National Academy of Sciences*, 99 (2002), 7821–7826 https://www.pnas.org/content/99/12/7821
- Grayson et al. 2016 – S. Grayson, K. Wade, G. Meaney & D. Greene, 'The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature.' In: *International Workshop on Computational History and Data-Driven Humanities* (2016), pp. 65-77.
- Jayannavar et al 2015 – P. Jayannavar et al, 'Validating literary theories using automatic social network extraction'. In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (2015), 32-41.
- Karsdorp et al. 2015. – F. Karsdorp, M. Kestemont, C. Schöch & A. van den Bosch. 'The 'Love Equation: Computational Modeling of Romantic Relationships in French

Classical Drama.' In: *Proceedings of the Sixth International Workshop on Computational Models of Narrative*, 2015, 98-107.

- Lee & Yeung 2012 – J. Lee & C.Y. Yeung, 'Extracting Networks of People and Places from Literary Texts'. In: *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation* 2012 (2012), 209-218.
- Lee & Wong 2016 – J. Lee & T.S. Wong, 'Hierarchy of Characters in the Chinese Buddhist Canon'. In: *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*,(2016), 531-534.
- Marsden 2011 – Peter V. Marsden, 'Network Clusters and Communities', In: *Encyclopedia of Social Networks*, edited by George A. Barnett. SAGE Knowledge: 2011 http://sk.sagepub.com/reference/socialnetworks
- Smeets et al 2019 – Smeets, Roel, Sanders, Eric & Antal van den Bosch. 'Character Centrality in Present-Day Dutch Literary Fiction'. In: *Digital Humanities Benelux Journal* 1:1 (2019)
- Seidel 2011 – Marc-David L. Seidel, 'Network Clusters and Communities', In: *Encyclopedia of Social Networks*, edited by George A. Barnett. SAGE Knowledge: 2011 http://sk.sagepub.com/reference/socialnetworks/n150.xml
- Stiller et al 2003 – J. Stiller et al, 'The Small World of Shakespeare's Plays.' In: *Hum Nat* 14 (2003) 4, p. 397-408.
- Van der Deijl et al 2017 – Deijl, Lucas van der, Pieterse, Saskia, Prinse, Marion & Smeets, Roel. 2016. 'Mapping the Demographic Landscape of Characters in Recent Dutch Prose: A Quantitative Approach to Literary Representation.'In: *Journal of Dutch Literature* 7 (2016) 1..
- Van der Deijl & Smeets 2018 – L. van der Deijl & R. Smeets, 'Tussen close en distant. Personage-hiërarchieën in Peter Buwalda's Bonita Avenue'. In: *Tijdschrift voor Nederlandse Taal – en Letterkunde* 134 (2018) 2, p. 123–145.
- Volker & Smeets 2019 – Volker, Beate, and Roel Smeets. "Imagined social structures: Mirrors or alternatives? A comparison between networks of characters in contemporary Dutch literature and networks of the population in the Netherlands." *Poetics* (2019): https://doi.org/10.1016/j.poetic.2019.101379