Discovering Structure in Social Networks of 19th Century Fiction

Siobhán Grayson
School of Computer Science, UCD, Ireland
siobhan.grayson@ucdconnect.ie

Karen Wade, Gerardine Meaney
Humanities Institute, UCD, Ireland
{karen.wade, gerardine.meaney}@ucd.ie

Jennie Rothwell, Maria Mulvany
Humanities Institute, UCD, Ireland
{jennie.rothwell, maria.mulvany}@ucd.ie

Derek Greene
School of Computer Science, UCD, Ireland
derek.greene@ucd.ie

ABSTRACT

Inspired by the increasing availability of large text corpora online, digital humanities scholars are adopting computational approaches to explore questions in the field of literature from new perspectives. In this paper, we examine detailed social networks of characters, extracted from several works of 19th century fiction by Jane Austen and Charles Dickens. This allows us to apply methodologies from social network analysis, such as community detection, to explore the structure of these networks. By evaluating the results in collaboration with literary scholars, we find that the structure of the character networks can reveal underlying structural aspects within a novel, particularly in relation to plot and characterisation.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Arts and Humanities

Keywords

Digital humanities, social network analysis, literary analysis

1. INTRODUCTION

For years organisations like Project Gutenberg, Google Books, and Open Library have been digitising and archiving cultural works, transforming the literary landscape and the way researchers engage with it. At the forefront of this revolution are digital humanities scholars, dedicated to reconciling computational approaches with more traditional literary analysis in order to explore texts from new perspectives. One approach which has amassed considerable attention is that of 'distant reading' [6]. This is the practice of understanding literature from a macro-level viewpoint, as opposed to taking a traditional micro-level 'close reading' approach. This form of analysis offers an opportunity to test assumptions about genre, narrative and other aspects of the novel that are based on the close reading of a small corpus of texts.

Numerous approaches have been proposed, ranging from statistical topic models [4], character profiling [3], character frequency analysis [1, 7], and sentiment analysis [2]. However, a method which has picked up considerable traction is social network analysis (SNA). SNA provides researchers with a unique level of abstraction (*i.e.* a network of nodes and edges), while maintaining the structure of the character collectives within a plot and thus, the societies they depict. In this paper, we construct and analyse *character networks* which are extracted from nine popular 19th century novels, written by the British authors Jane Austen and Charles Dickens. These networks are constructed from digital texts available from Project Gutenberg, which are manually annotated by literary scholars with a "radically inclusive" methodology, in order to include as many character entities as possible, including minor and collective-presenting characters. We apply the overlapping community detection algorithm OSLOM [5] to explore each network's structure at different levels of granularity to discover *character communities*. The results discussed in Section 3 show how the literary technique of the two authors, particularly in relation to characterisation, is revealed by the resulting communities.

2. DATA AND METHODS

Character Networks. We consider a collection of novels from two 19th century British novelists – six by Jane Austen and three by Charles Dickens – where the original texts were sourced online from Project Gutenberg. To annotate each text, we firstly construct a *character dictionary* which includes a single entry for each unique character in the novel and the corresponding *aliases* for that character which appear in that novel. Once the dictionary has been compiled, all instances of a character's aliases in the novel text are replaced with their definitive name. A node is created for each character in the novel's character dictionary, with that character's attributes attached. Then we tokenise each chapter in the previously-annotated text of the novel, and identify all *co-occurrences* of pairs of character definitive names that appear within a sliding window of length w words. For each chapter, we count the number of co-occurrences for every pair of characters, using a *collinear strategy* that looks at consecutive pairs of characters. Note that we use all types of co-occurrences rather than considering direct conversations alone, as this allows us to capture a wide variety of types of interactions and associations between characters. We then create a weighted character network for the chapter, where an edge is weighted such that it corresponds to the number of co-occurrences between the pair. Finally, we construct an overall character network for the novel by aggregating the networks from all chapters.

Novel	$\#N$	$\#E$	l	d	R1	R2
Northanger Abbey	94	255	2.4	0.06	7	1
Pride and Prejudice	117	485	2.4	0.07	10	1
Persuasion	136	437	2.8	0.05	16	2
Sense and Sensibility	158	508	2.6	0.04	12	2
Emma	193	620	2.7	0.05	17	5
Mansfield Park	218	649	2.9	0.03	16	7
Oliver Twist	286	696	3.2	0.02	16	10
Great Expectations	288	741	3.1	0.02	19	8
Bleak House	516	1526	3.4	0.01	44	16

Table 1: Summary of network results, top 6 - Austen, last 3 - Dickens. $\#N$ is number of nodes, $\#E$ is number of edges, l is the average path length, d is the network edge density. The numbers of communities identified at resolutions R1 and R2 are also reported.

Community Finding. For the purpose of community detection, we employ the order statistics local optimisation method (OSLOM) [5]. This algorithm works by measuring the quality of potential communities using a fitness function based on their statistical significance. This is derived by estimating the probability of finding a similar community with an identical degree sequence in a null model devoid of community structure. The final output of the process is a set of one or more communities which potentially overlap. When applying OSLOM, two key parameter values govern the nature of the communities that the algorithm detects: *coverage parameter* (cp): determines whether communities should be merged or not. *p-threshold* (t): alters the threshold at which a community is deemed to be significant. We jointly refer to these parameter pairs as the *resolution*.

3. RESULTS

Details of the resulting overall character networks and communities are provided in Table 1. We immediately observe that the networks of Austen contain fewer characters, have a shorter average path length, and have a higher edge density than the networks of Dickens. This suggests that the fictional societies constructed by Austen are generally more compact and closely-knit than those of Dickens. We focus on results for two distinct resolutions: R1 ($cp = 0.1$ and $t = 1.0$), R2 ($cp = 0.1$ and $t = 0.1$). R1 resulted in the highest number of communities returned for both authors, while R2 returns fewer communities which are more granular. Next, we provide a case study of the results obtained for the text *Oliver Twist*.

Case Study: Oliver Twist. An interesting aspect of the communities found in Oliver Twist is that in many instances, they can be convincingly linked to "micronarratives" – *i.e.* miniature subplots or anecdotes within the novel which do not contribute to large-scale plot trajectories. These communities can often be found within an extremely limited textual space. Fig. 1 visualises a subgraph of the overall character network for the novel, where only those nodes assigned to communities of size ≥ 2 at resolution R2 are shown. The highlighted community **A** consists of three characters ("a man who was hung in Jamaica", "his murdered master", Mr. Grimwig), where the first two only appear within a single anecdote, which serves primarily to illustrate the character of the speaker rather than to impact upon the novel's plot. Such communities are apparent for each of the Dickens novels in our study, reflecting the frequent use of the micronarrative device by the author. This is in contrast to the networks from Austen's novels, where communities appear more likely to centre on the social circles of specific individuals.

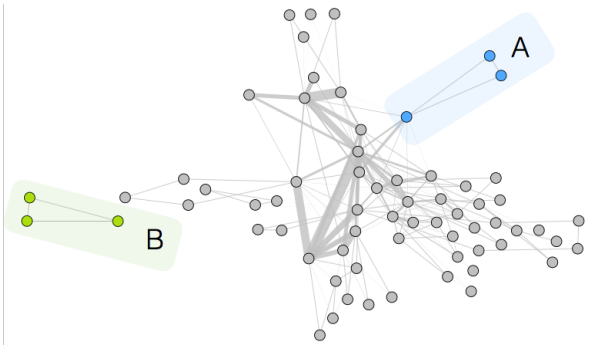

Figure 1: Subgraph of the overall character network for *Oliver Twist*, showing communities returned at resolution R2. Highlighted communities A and B correspond to "micronarratives".

4. CONCLUSIONS AND FUTURE WORK

We have examined highly detailed social networks of characters extracted from several works of 19th century fiction by Jane Austen and Charles Dickens. We applied the community detection algorithm OSLOM to each network using a number of different 'resolution' values to investigate the diversity in the communities returned and the insights which they provide. The differing levels of granularity allowed us to identify and observe micro-plots or anecdotes within the novels. Additionally, we observed a number of overlapping characters and found protagonists to be members of the most communities. We plan to extend our study to include novels by other authors, covering different time periods and different genres, in order to assess more complex literary hypotheses. Additionally, we suggest that our radically inclusive approach to the construction of detailed social networks provides insights into the construction of the novel that are elided when the focus is solely on the main characters.

Acknowledgments. This research was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, in collaboration with the Nation, Genre and Gender project funded by the Irish Research Council.

5. REFERENCES

[1] M. Elsner. Character-based kernels for novelistic plot structure. In *European Chapter of the Association for Computational Linguistics*, pages 634–644, 2012.

[2] M. Elsner. Abstract Representations of Plot Struture. *LiLT (Linguistic Issues in Language Technology)*, 12(5), 2015.

[3] L. Flekova and I. Gurevych. Personality Profiling of Fictional Characters using Sense-Level Links between Lexical Resources. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, 2015.

[4] M. L. Jockers and D. Mimno. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769, 2013.

[5] A. Lancichinetti, F. Radicchi, J. Ramasco, S. Fortunato, and E. Ben-Jacob. Finding statistically significant communities in networks. *PLoS ONE*, 6(4):e18961, 2011.

[6] F. Moretti. Network Theory, Plot Analysis. *New Left Review*, 68:80–102, 2011.

[7] G. Sack. Simulating plot: Towards a generative model of narrative structure. In *2011 AAAI Fall Symposium Series*, 2011.