Discovering Structure in Social Networks of 19th Century Fiction

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

Karen Wade, Gerardine Meaney
Humanities Institute, UCD, Ireland
{karen.wade, gerardine.meaney}@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 [Applied Computing]: 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.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WebSci ’16 May 22-25, 2016, Hannover, Germany
© 2016 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4208-7/16/05.
DOI: http://dx.doi.org/10.1145/2908131.2908196

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.
Table 1: Summary of network results, top 6 - Austen, last 3 - Dickens.

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

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