Extracting Social Networks from Literary Fiction

David K. Elson
Dept. of Computer Science
Columbia University
delson@cs.columbia.edu

Nicholas Dames
English Department
Columbia University
nd122@columbia.edu

Kathleen R. McKeown
Dept. of Computer Science
Columbia University
kathy@cs.columbia.edu

Abstract

We present a method for extracting social networks from literature, namely, nineteenth-century British novels and serials. We derive the networks from dialogue interactions, and thus our method depends on the ability to determine when two characters are in conversation. Our approach involves character name chunking, quoted speech attribution and conversation detection given the set of quotes. We extract features from the social networks and examine their correlation with one another, as well as with metadata such as the novel's setting. Our results provide evidence that the majority of novels in this time period do not fit two characterizations provided by literacy scholars. Instead, our results suggest an alternative explanation for differences in social networks.

1 Introduction

Literary studies about the nineteenth-century British novel are often concerned with the nature of the community that surrounds the protagonist. Some theorists have suggested a relationship between the size of a community and the amount of dialogue that occurs, positing that "face to face time" diminishes as the number of characters in the novel grows. Others suggest that as the social setting becomes more urbanized, the quality of dialogue also changes, with more interactions occurring in rural communities than urban communities. Such claims have typically been made, however, on the basis of a few novels that are studied in depth. In this paper, we aim to determine whether an automated study of a much larger sample of nineteenth century novels supports these claims.

The research presented here is concerned with the extraction of social networks from literature. We present a method to automatically construct a network based on dialogue interactions between characters in a novel. Our approach includes components for finding instances of quoted speech, attributing each quote to a character, and identifying when certain characters are in conversation. We then construct a network where characters are vertices and edges signify an amount of bilateral conversation between those characters, with edge weights corresponding to the frequency and length of their exchanges. In contrast to previous approaches to social network construction, ours relies on a novel combination of pattern-based detection, statistical methods, and adaptation of standard natural language tools for the literary genre. We carried out this work on a corpus of 60 nineteenth-century novels and serials, including 31 authors such as Dickens, Austen and Conan Doyle.

In order to evaluate the literary claims in question, we compute various characteristics of the dialogue-based social network and stratify these results by categories such as the novel's setting. For example, the density of the network provides evidence about the cohesion of a large or small community, and cliques may indicate a social fragmentation. Our results surprisingly provide evidence that the majority of novels in this time period do not fit the suggestions provided by literary scholars, and we suggest an alternative explanation for our observations of differences across novels.

In the following sections, we survey related work on social networks as well as computational studies of literature. We then present the literary hypotheses in more detail. We describe the methods we use to extract dialogue and construct conversational networks, along with our approach to analyzing their characteristics. After we present the statistical results, we analyze their significance from a literary perspective.

2 Related Work

Computer-assisted literary analysis has typically occurred at the word level. This level of granularity lends itself to studies of authorial style based on patterns of word use (Burrows, 2004), and researchers have successfully "outed" the writers of anonymous texts by comparing their style to that of a corpus of known authors (Mostellar and Wallace, 1984). Determining instances of "text reuse," a type of paraphrasing, is also a form of analysis at the lexical level, and it has recently been used to validate theories about the lineage of ancient texts (Lee, 2007).

Analysis of literature using more semantically-oriented techniques has been rare, most likely because of the difficulty in automatically determining meaningful interpretations. Some exceptions include recent work on learning common event sequences in news stories (Chambers and Jurafsky, 2008), an approach based on statistical methods, and the development of an event calculus for characterizing stories written by children (Halpin et al., 2004), a knowledge-based strategy. On the other hand, literary theorists, linguists and others have long developed symbolic but non-computational models for novels. For example, Moretti (2005) has graphically mapped out texts according to geography, social connections and other variables.

While researchers have not attempted the automatic construction of social networks representing connections between characters in a corpus of novels, the ACE program has involved entity and relation extraction in unstructured text (Doddington et al., 2004). Other recent work in social network construction has explored the use of structured data such as email headers (McCallum et al., 2007) and U.S. Senate bill cosponsorship (Cho and Fowler, 2010). In an analysis of discussion forums, Gruzd and Haythornthwaite (2008) explored the use of message text as well as posting data to infer who is talking to whom. In this paper, we also explore how to build a network based on conversational interaction, but we analyze the reported dialogue found in novels to determine the links. The kinds of language that is used to signal such information is quite different in the two media. In discussion forums, people tend to use addresses such as "Hi Tom," while in novels, a system must determine both the speaker of a quotation and then the intended recipient of the dialogue act. This is a significantly different problem.

3 Hypotheses

It is commonly held that the novel is a literary form which tries to produce an accurate representation of the social world. Within literary studies, the recurring problem is how that representation is achieved. Theories about the relation between novelistic form (the workings of plot, characters, and dialogue, to take the most basic categories) and changes to real-world social milieux abound. Many of these theories center on nineteenth-century European fiction; innovations in novelistic form during this period, as well as the rapid social changes brought about by revolution, industrialization, and transport development, have traditionally been linked. These theories, however, have used only a select few representative novels as proof. By using statistical methods of analysis, it is possible to move beyond this small corpus of proof texts. We believe these methods are essential to testing the validity of some core theories about social interaction and their representation in literary genres like the novel.

Major versions of the theories about the social worlds of nineteenth-century fiction tend to center on characters, in two specific ways: how many characters novels tend to have, and how those characters interact with one another. These two "formal" facts about novels are usually explained with reference to a novel's setting. From the influential work of the Russian critic Mikhail Bakhtin to the present, a consensus emerged that as novels are increasingly set in urban areas, the number of characters and the quality of their interaction change to suit the setting. Bakhtin's term for this causal relationship was *chronotope*: the "intrinsic interconnectedness of temporal and spatial relationships that are artistically expressed in literature," in which "space becomes charged and responsive to movements of time, plot, and history" (Bakhtin, 1981, 84). In Bakhtin's analysis, different spaces have different social and emotional potentialities, which in turn affect the most basic aspects of a novel's aesthetic technique.

After Bakhtin's invention of the chronotope, much literary criticism and theory devoted itself to filling in, or describing, the qualities of specific chronotopes, particularly those of the village or rural environment and the city or urban environment. Following a suggestion of Bakhtin's that the population of village or rural fictions is modeled on the world of the family, made up of

Author/Title/Year	Persp.	Setting	Author/Title/Year	Persp.	Setting
Ainsworth, *Jack Sheppard* (1839)	3rd	urban	Gaskell, *North and South* (1854)	3rd	urban
Austen, *Emma* (1815)	3rd	rural	Gissing, *In the Year of Jubilee* (1894)	3rd	urban
Austen, *Mansfield Park* (1814)	3rd	rural	Gissing, *New Grub Street* (1891)	3rd	urban
Austen, *Persuasion* (1817)	3rd	rural	Hardy, *Jude the Obscure* (1894)	3rd	mixed
Austen, *Pride and Prejudice* (1813)	3rd	rural	Hardy, *The Return of the Native* (1878)	3rd	rural
Braddon, *Lady Audley's Secret* (1862)	3rd	mixed	Hardy, *Tess of the d'Ubervilles* (1891)	3rd	rural
Braddon, *Aurora Floyd* (1863)	3rd	rural	Hughes, *Tom Brown's School Days* (1857)	3rd	rural
Brontë, Anne, *The Tenant of Wildfell Hall* (1848)	1st	rural	James, *The Portrait of a Lady* (1881)	3rd	urban
Brontë, Charlotte, *Jane Eyre* (1847)	1st	rural	James, *The Ambassadors* (1903)	3rd	urban
Brontë, Charlotte, *Villette* (1853)	1st	mixed	James, *The Wings of the Dove* (1902)	3rd	urban
Brontë, Emily, *Wuthering Heights* (1847)	1st	rural	Kingsley, *Alton Locke* (1860)	1st	mixed
Bulwer-Lytton, *Paul Clifford* (1830)	3rd	urban	Martineau, *Deerbrook* (1839)	3rd	rural
Collins, *The Moonstone* (1868)	1st	urban	Meredith, *The Egoist* (1879)	3rd	rural
Collins, *The Woman in White* (1859)	1st	urban	Meredith, *The Ordeal of Richard Feverel* (1859)	3rd	rural
Conan Doyle, *The Sign of the Four* (1890)	1st	urban	Mitford, *Our Village* (1824)	1st	rural
Conan Doyle, *A Study in Scarlet* (1887)	1st	urban	Reade, *Hard Cash* (1863)	3rd	urban
Dickens, *Bleak House* (1852)	mixed	urban	Scott, *The Bride of Lammermoor* (1819)	3rd	rural
Dickens, *David Copperfield* (1849)	1st	mixed	Scott, *The Heart of Mid-Lothian* (1818)	3rd	rural
Dickens, *Little Dorrit* (1855)	3rd	urban	Scott, *Waverley* (1814)	3rd	rural
Dickens, *Oliver Twist* (1837)	3rd	urban	Stevenson, *The Strange Case of Dr. Jekyll and Mr. Hyde* (1886)	1st	urban
Dickens, *The Pickwick Papers* (1836)	3rd	mixed	Stoker, *Dracula* (1897)	1st	urban
Disraeli, *Sybil, or the Two Nations* (1845)	3rd	mixed	Thackeray, *History of Henry Esmond* (1852)	1st	urban
Edgeworth, *Belinda* (1801)	3rd	rural	Thackeray, *History of Pendennis* (1848)	1st	urban
Edgeworth, *Castle Rackrent* (1800)	3rd	rural	Thackeray, *Vanity Fair* (1847)	3rd	urban
Eliot, *Adam Bede* (1859)	3rd	rural	Trollope, *Barchester Towers* (1857)	3rd	rural
Eliot, *Daniel Deronda* (1876)	3rd	urban	Trollope, *Doctor Thorne* (1858)	3rd	rural
Eliot, *Middlemarch* (1871)	3rd	rural	Trollope, *Phineas Finn* (1867)	3rd	urban
Eliot, *The Mill on the Floss* (1860)	3rd	rural	Trollope, *The Way We Live Now* (1874)	3rd	urban
Galt, *Annals of the Parish* (1821)	1st	rural	Wilde, *The Picture of Dorian Gray* (1890)	3rd	urban
Gaskell, *Mary Barton* (1848)	3rd	urban	Wood, *East Lynne* (1860)	3rd	mixed

Table 1: Properties of the nineteenth-century British novels and serials included in our study.

an intimately related set of characters, many critics analyzed the formal expression of this world as constituted by a small set of characters who express themselves conversationally. Raymond Williams used the term "knowable communities" to describe this world, in which face-to-face relations of a restricted set of characters are the primary mode of social interaction (Williams, 1975, 166).

By contrast, the urban world, in this traditional account, is both larger and more complex. To describe the social-psychological impact of the city, Franco Moretti argues, protagonists of urban novels "change overnight from 'sons' into 'young men': their affective ties are no longer vertical ones (between successive generations), but horizontal, within the same generation. They are drawn towards those unknown yet congenial faces seen in gardens, or at the theater; future friends, or rivals, or both" (Moretti, 1999, 65). The result is two-fold: more characters, indeed a mass of characters, and more interactions, although less actual conversation; as literary critic Terry Eagleton argues, the city is where "most of our encounters consist of seeing rather than speaking, glimpsing each other as objects rather than conversing as fellow subjects" (Eagleton, 2005, 145). Moretti argues in similar terms. For him, the difference in number of characters is "not just a matter of quantity... it's a qualitative, morphological one" (Moretti, 1999, 68). As the number of characters increases, Moretti argues (following Bakhtin in his logic), social interactions of different kinds and durations multiply, displacing the family-centered and conversational logic of village or rural fictions. "The narrative system becomes complicated, unstable: the city turns into a gigantic roulette table, where helpers and antagonists mix in unpredictable combinations" (Moretti, 1999, 68). This argument about how novelistic setting produces different forms of social interaction is precisely what our method seeks to evaluate.

Our corpus of 60 novels was selected for its representativeness, particularly in the following categories: authorial (novels from the major canoni-

cal authors of the period), historical (novels from each decade), generic (from the major sub-genres of nineteenth-century fiction), sociological (set in rural, urban, and mixed locales), and technical (narrated in first-person and third-person form). The novels, as well as important metadata we assigned to them (the perspective and setting), are shown in Table 1. We define *urban* to mean set in a metropolitan zone, characterized by multiple forms of labor (not just agricultural). Here, social relations are largely financial or commercial in character. We conversely define *rural* to describe texts that are set in a country or village zone, where agriculture is the primary activity, and where land-owning, non-productive, rent-collecting gentry are socially predominant. Social relations here are still modeled on feudalism (relations of peasant-lord loyalty and family tie) rather than the commercial cash nexus. We also explored other properties of the texts, such as literary genre, but focus on the results found with setting and perspective. We obtained electronic encodings of the texts from Project Gutenberg. All told, these texts total more than 10 million words.

We assembled this representative corpus in order to test two hypotheses, which are derived from the aforementioned theories:

1. That there is an inverse correlation between the amount of dialogue in a novel and the number of characters in that novel. One basic, shared assumption of these theorists is that as the network of characters expands– as, in Moretti's words, a quantitative change becomes qualitative– the importance, and in fact amount, of dialogue decreases. With a method for extracting conversation from a large corpus of texts, it is possible to test this hypothesis against a wide range of data.

2. That a significant difference in the nineteenth-century novel's representation of social interaction is geographical: novels set in urban environments depict a complex but loose social network, in which numerous characters share little conversational interaction, while novels set in rural environments inhabit more tightly bound social networks, with fewer characters sharing much more conversational interaction. This hypothesis is based on the contrast between Williams's rural "knowable communities" and the sprawling, populous, less conversational urban fictions or Moretti's and Eagleton's analyses. If true, it would suggest that the inverse relationship of hypothesis #1 (more characters means less conversation) can be correlated to, and perhaps even caused by, the geography of a novel's setting. The claims about novelistic geography and social interaction have usually been based on comparisons of a selected few novelists (Jane Austen and Charles Dickens preeminently). Do they remain valid when tested against a larger corpus?

4 Extracting Conversational Networks from Literature

In order to test these hypotheses, we developed a novel approach to extracting social networks from literary texts themselves, building on existing analysis tools. We defined "social network" as "conversational network" for purposes of evaluating these literary theories. In a conversational network, vertices represent characters (assumed to be named entities) and edges indicate at least one instance of dialogue interaction between two characters over the course of the novel. The weight of each edge is proportional to the amount of interaction. We define a conversation as a continuous span of narrative time featuring a set of characters in which the following conditions are met:

1. The characters are in the same place at the same time;

2. The characters take turns speaking; and

3. The characters are mutually aware of each other and each character's speech is mutually intended for the other to hear.

In the following subsections, we discuss the methods we devised for the three problems in text processing invoked by this approach: identifying the characters present in a literary text, assigning a "speaker" (if any) to each instance of quoted speech from among those characters, and constructing a social network by detecting conversations from the set of dialogue acts.

4.1 Character Identification

The first challenge was to identify the candidate speakers by "chunking" names (such as *Mr. Holmes*) from the text. We processed each novel

with the Stanford NER tagger (Finkel et al., 2005) and extracted noun phrases that were categorized as persons or organizations. We then clustered the noun phrases into coreferents for the same entity (person or organization). The clustering process is as follows:

1. For each named entity, we generate variations on the name that we would expect to see in a coreferent. Each variation omits certain parts of multi-word names, respecting titles and first/last name distinctions, similar to work by Davis et al. (2003). For example, *Mr. Sherlock Holmes* may refer to the same character as *Mr. Holmes*, *Sherlock Holmes*, *Sherlock* and *Holmes*.

2. For each named entity, we compile a list of other named entities that may be coreferents, either because they are identical or because one is an expected variation on the other.

3. We then match each named entity to the most recent of its possible coreferents. In aggregate, this creates a cluster of mentions for each character.

We also pre-processed the texts to normalize formatting, detect headings and chapter breaks, remove metadata, and identify likely instances of quoted speech (that is, mark up spans of text that fall between quotation marks, assumed to be a superset of the quoted speech present in the text).

4.2 Quoted Speech Attribution

In order to programmatically assign a speaker to each instance of quoted speech, we applied a high-precision subset of a general approach we describe elsewhere (Elson and McKeown, 2010). The first step of this approach was to compile a separate training and testing corpus of literary texts from British, American and Russian authors of the nineteenth and twentieth centuries. The training corpus consisted of about 111,000 words including 3,176 instances of quoted speech. To obtain gold-standard annotations, we conducted an online survey via Amazon's Mechanical Turk program. For each quote, we asked three annotators to independently choose a speaker from the list of contextual candidates– or, choose "spoken by an unlisted character" if the answer was not available, or "not spoken by any character" for non-dialogue cases such as sneer quotes.

We divided this corpus into training and testing sets, and used the training set to develop a categorizer that assigned one of five syntactic categories to each quote. For example, if a quote is followed by a verb that indicates verbal expression (such as "said"), and then a character mention, a category called *Character trigram* is assigned to the quote. The fifth category is a catch-all for quotes that do not fall into the other four. In many cases, the answer can be reliably determined based solely on its syntactic category. For instance, in the *Character trigram* category, the mentioned character is the quote's speaker in 99% of both the training and testing sets.

In all, we were able to determine the speaker of 57% of the testing set with 96% accuracy just on the basis of syntactic categorization. This is the technique we used to construct our conversational networks. In another study, we applied machine learning tools to the data (one model for each syntactic category) and achieved an overall accuracy of 83% over the entire test set (Elson and McKeown, 2010). The other 43% of quotes are left here as "unknown" speakers; however, in the present study, we are interested in *conversations* rather than individual quotes. Each conversation is likely to consist of multiple quotes by each speaker, increasing the chances of detecting the interaction. Moreover, this design decision emphasizes the precision of the social networks over their recall. This tilts "in favor" of hypothesis #1 (that there are fewer social interactions in larger communities); however, we shall see that despite the emphasis of precision over recall, we identify a sufficient mass of interactions in the texts to constitute evidence against this hypothesis.

4.3 Constructing social networks

We then applied the results from our character identification and quoted speech attribution methods toward the construction of conversational networks from literature. We derived one network from each text in our corpus.

We first assigned vertices to character entities that are mentioned repeatedly throughout the novel. Coreferents for the same name (such as *Mr. Darcy* and *Darcy*) were grouped into the same vertex. We found that a network that included incidental or single-mention named entities became too noisy to function effectively, so we filtered out the entities that are mentioned fewer than three

times in the novel or are responsible for less than 1% of the named entity mentions in the novel.

We assigned undirected edges between vertices that represent *adjacency* in quoted speech fragments. Specifically, we set the weight of each undirected edge between two character vertices to the total length, in words, of all quotes that either character speaks from among all pairs of adjacent quotes in which they both speak– implying face to face conversation. We empirically determined that the most accurate definition of "adjacency" is one where the two characters' quotes fall within 300 words of one another with no attributed quotes in between. When such an adjacency is found, the length of the quote is added to the edge weight, under the hypothesis that the significance of the relationship between two individuals is proportional to the length of the dialogue that they exchange. Finally, we normalized each edge's weight by the length of the novel.

An example network, automatically constructed in this manner from Jane Austen's *Mansfield Park*, is shown in Figure 1. The width of each vertex is drawn to be proportional to the character's share of all the named entity mentions in the book (so that protagonists, who are mentioned frequently, appear in larger ovals). The width of each edge is drawn to be proportional to its weight (total conversation length).

We also experimented with two alternate methods for identifying edges, for purposes of a baseline:

1. The "correlation" method divides the text into 10-paragraph segments and counts the number of mentions of each character in each segment (excluding mentions inside quoted speech). It then computes the Pearson product-moment correlation coefficient for the distributions of mentions for each pair of characters. These coefficients are used for the edge weights. Characters that tend to appear together in the same areas of the novel are taken to be more socially connected, and have a higher edge weight.

2. The "spoken mention" method counts occurrences when one character refers to another in his or her quoted speech. These counts, normalized by the length of the text, are used as edge weights. The intuition is that characters who refer to one another are likely to be in conversation.

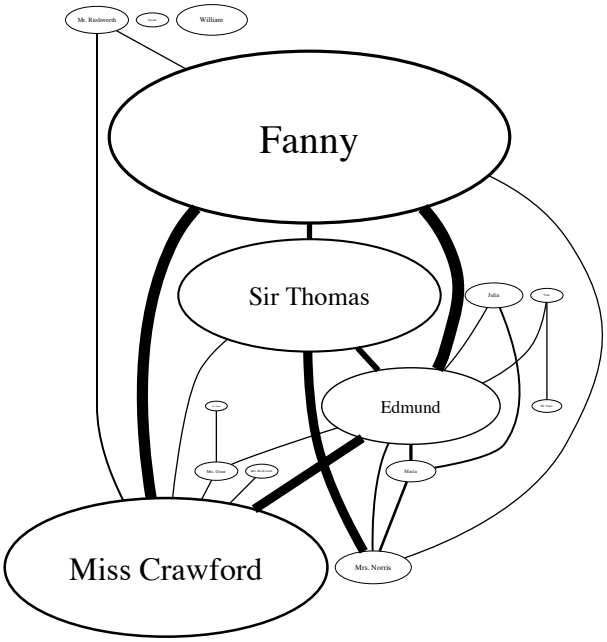

Figure 1: Automatically extracted conversation network for Jane Austen's *Mansfield Park*.

4.4 Evaluation

To check the accuracy of our method for extracting conversational networks, we conducted an evaluation involving four of the novels (*The Sign of the Four*, *Emma*, *David Copperfield* and *The Portrait of a Lady*). We did not use these texts when developing our method for identifying conversations. For each book, we randomly selected 4-5 chapters from among those with significant amounts of quoted speech, so that all excerpts from each novel amounted to at least 10,000 words. We then asked three annotators to identity all the conversations that occur in all 44,000 words. We requested that the annotators include both direct and indirect (unquoted) speech, and define "conversation" as in the beginning of Section 4, but exclude "retold" conversations (those that occur within other dialogue).

We processed the annotation results by breaking down each multi-way conversation into all of its unique two-character interactions (for example, a conversation between four people indicates six bilateral interactions). To calculate inter-annotator agreement, we first compiled a list of all possible interactions between all characters in each text. In this model, each annotator contributed a set of "yes" or "no" decisions, one for every character pair. We then applied the kappa measurement for agreement in a binary classification problem (Co-

Method	Precision	Recall	F
Speech adjacency	.95	.51	.67
Correlation	.21	.65	.31
Spoken-mention	.45	.49	.47

Table 2: Precision, recall, and F-measure of three methods for detecting bilateral conversations in literary texts.

hen, 1960). In 95% of character pairs, annotators were unanimous, which is a high agreement of $k = .82$.

The precision and recall of our method for detecting conversations is shown in Table 2. Precision was .95; this indicates that we can be confident in the specificity of the conversational networks that we automatically construct. Recall was .51, indicating a sensitivity of slightly more than half. There were several reasons that we did not detect the missing links, including indirect speech, quotes attributed to anaphoras or coreferents, and "diffuse" conversations in which the characters do not speak in turn with one another.

To calculate precision and recall for the two baseline social networks, we set a threshold t to derive a binary prediction from the continuous edge weights. The precision and recall values shown for the baselines in Table 2 represent the highest performance we achieved by varying t between 0 and 1 (maximizing F-measure over t). Both baselines performed significantly worse in precision and F-measure than our quoted speech adjacency method for detecting conversations.

5 Data Analysis

5.1 Feature extraction

We extracted features from the conversational networks that emphasize the complexity of the social interactions found in each novel:

1. The number of characters and the number of speaking characters

2. The variance of the distribution of quoted speech (specifically, the proportion of quotes spoken by the n most frequent speakers, for $1 \leq n \leq 5$)

3. The number of quotes, and proportion of words in the novel that are quoted speech

4. The number of 3-cliques and 4-cliques in the social network

5. The *average degree* of the graph, defined as

$$\frac{\sum_{v \in V} |E_v|}{|V|} = \frac{2|E|}{|V|} \quad (1)$$

where $|E_v|$ is the number of edges incident on a vertex v, and $|V|$ is the number of vertices. In other words, this determines the average number of characters connected to each character in the conversational network ("with how many people on average does a character converse?").

6. A variation on *graph density* that normalizes the average degree feature by the number of characters:

$$\frac{\sum_{v \in V} |E_v|}{|V|(|V|-1)} = \frac{2|E|}{|V|(|V|-1)} \quad (2)$$

By dividing again by $|V| - 1$, we use this as a metric for the overall connectedness of the graph: "with what *percent* of the entire network (besides herself) does each character converse, on average?" The weight of the edge, as long as it is greater than 0, does not affect either the network's average degree or graph density.

5.2 Results

We derived results from the data in two ways. First, we examined the strengths of the correlations between the features that we extracted (for example, between number of character vertices and the average degree of each vertex). We used Pearson's product-moment correlation coefficient in these calculations. Second, we compared the extracted features to the metadata we previously assigned to each text (e.g., urban vs. rural).

Hypothesis #1, which we described in Section 3, claims that there is an inverse correlation between the amount of dialogue in a nineteenth-century novel and the number of characters in that novel. We did not find this to be the case. Rather, we found a weak but positive correlation (r=.16) between the number of quotes in a novel and the number of characters (normalizing the quote count for text length). There was a stronger positive correlation (r=.50) between the number of unique speakers (those characters who speak at least once) and the normalized number of quotes, suggesting that larger networks have more conversations than smaller ones. But because the first

correlation is weak, we investigated whether further analysis could identify other evidence that confirms or contradicts the hypothesis.

Another way to interpret hypothesis #1 is that social networks with more characters tend to break apart and be less connected. However, we found the opposite to be true. The correlation between the number of characters in each graph and the average degree (number of conversation partners) for each character was a positive, moderately strong r=.42. This is not a given; a network can easily, for example, break into minimally connected or mutually exclusive subnetworks when more characters are involved. Instead, we found that networks tend to stay close-knit regardless of their size: even the density of the graph (the percentage of the community that each character talks to) grows with the total population size at r=.30. Moreover, as the population of *speakers* grows, the density is likely to increase at r=.49. A higher number of characters (speaking or non-speaking) is also correlated with a higher *rate* of 3-cliques per character (r=.38), as well as with a more balanced distribution of dialogue (the share of dialogue spoken by the top three speakers decreases at r=−.61). This evidence suggests that in nineteenth-century British literature, it is the small communities, rather than the large ones, that tend to be disconnected.

Hypothesis #2, meanwhile, posited that a novel's setting (urban or rural) would have an effect on the structure of its social network. After defining "social network" as a conversational network, we did not find this to be the case. Surprisingly, the numbers of characters and speakers found in the urban novel were *not* significantly greater than those found in the rural novel. Moreover, each of the features we extracted, such as the rate of cliques, average degree, density, and rate of characters' mentions of other characters, did not change in a statistically significant manner between the two genres. For example, Figure 2 shows the mean over all texts of each network's average degree, with confidence intervals, separated by setting into urban and rural. The increase in degree seen in urban texts is not significant.

Rather, the only type of metadata variable that *did* impact the average degree with any significance was the text's perspective. Figure 2 also separates texts into first- and third-person tellings and shows the means and confidence intervals for the

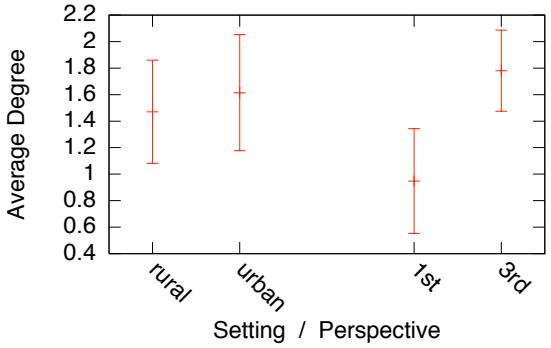

Figure 2: The average degree for each character as a function of the novel's setting and its perspective.

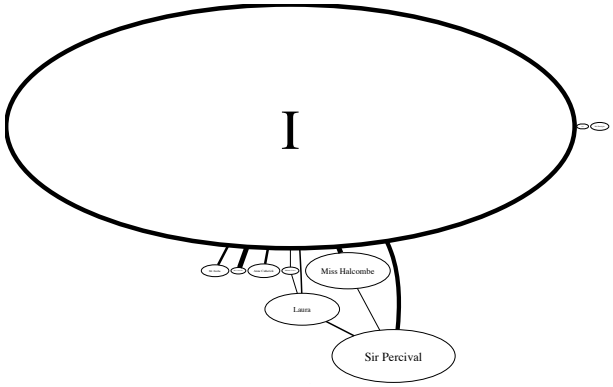

Figure 3: Conversational networks for first-person novels like Collins's *The Woman in White* are less connected due to the structure imposed by the perspective.

average degree measure. Stories told in the third person had much more connected networks than stories told in the first person: not only did the average degree increase with statistical significance (by the homoscedastic t-test to $p < .005$), so too did the graph density ($p < .05$) and the rate of 3-cliques per character ($p < .05$).

We believe the reason for this can be intuited with a visual inspection of a first-person graph. Figure 3 shows the conversational network extracted for Collins's *The Woman in White*, which is told in the first person. Not surprisingly, the most oft-repeated named entity in the text is *I*, referring to the narrator. More surprising is the *lack* of conversation connections between the auxiliary characters. The story's structure revolves around the narrator and each character is understood in terms of his or her relationship to the narrator. Private conversations between auxiliary characters would not include the narrator, and thus do not appear in a

first-hand account. An "omniscient" third person narrator, by contrast, can eavesdrop on any pair of characters conversing. This highlights the importance of detecting reported and indirect speech in future work, as a first-person narrator may hear about other connections without witnessing them.

6 Literary Interpretation of Results

Our data, therefore, markedly do not confirm hypothesis #1. They also suggest, in relation to hypothesis #2 (also not confirmed by the data), a strong reason why.

One of the basic assumptions behind hypothesis #2– that urban novels contain more characters, mirroring the masses of nineteenth-century cities– is not borne out by our data. Our results do, however, strongly correlate a point of view (third-person narration) with more frequently connected characters, implying tighter and more talkative social networks.

We would propose that this suggests that the form of a given novel– the standpoint of the narrative voice, whether the voice is "omniscient" or not– is far more determinative of the kind of social network described in the novel than where it is set or even the number of characters involved. Whereas standard accounts of nineteenth-century fiction, following Bakhtin's notion of the "chronotope," emphasize the content of the novel as determinative (where it is set, whether the novel fits within a genre of "village" or "urban" fiction), we have found that content to be surprisingly irrelevant to the shape of social networks within. Bakhtin's influential theory, and its detailed reworkings by Williams, Moretti, and others, suggests that as the novel becomes more urban, more centered in (and interested in) populous urban settings, the novel's form changes to accommodate the looser, more populated, less conversational networks of city life. Our data suggests the opposite: that the "urban novel" is not as strongly distinctive a form as has been asserted, and that in fact it can look much like the village fictions of the century, as long as the same method of narration is used.

This conclusion leads to some further considerations. We are suggesting that the important element of social networks in nineteenth-century fiction is not where the networks are set, but from what standpoint they are imagined or narrated. Narrative voice, that is, trumps setting.

7 Conclusion

In this paper, we presented a method for characterizing a text of literary fiction by extracting the network of social conversations that occur between its characters. This allowed us to take a systematic and wide look at a large corpus of texts, an approach which complements the narrower and deeper analysis performed by literary scholars and can provide evidence for or against some of their claims. In particular, we described a high-precision method for detecting face-to-face conversations between two named characters in a novel, and showed that as the number of characters in a novel grows, so too do the cohesion, interconnectedness and balance of their social network. In addition, we showed that the form of the novel (first- or third-person) is a stronger predictor of these features than the setting (urban or rural). Our results thus far suggest further review of our methods, our corpus and our results for more insights into the social networks found in this and other genres of fiction.

8 Acknowledgments

This material is based on research supported in part by the U.S. National Science Foundation (NSF) under IIS-0935360. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Mikhail Bakhtin. 1981. Forms of time and of the chronotope in the novel. In Trans. Michael Holquist and Caryl Emerson, editors, *The Dialogic Imagination: Four Essays*, pages 84–258. University of Texas Press, Austin.

John Burrows. 2004. Textual analysis. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell, Oxford.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *In Proceedings of the 46th Annual Meeting of the Association of Com- putational Linguistics (ACL-08)*, pages 789–797, Columbus, Ohio.

Wendy K. Tam Cho and James H. Fowler. 2010. Legislative success in a small world: Social network analysis and the dynamics of congressional legislation. *The Journal of Politics*, 72(1):124–135.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Peter T. Davis, David K. Elson, and Judith L. Klavans. 2003. Methods for precise named entity matching in digital collections. In *Proceedings of the Third ACM/IEEE Joint Conference on Digital Libraries (JCDL '03)*, Houston, Texas.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon.

Terry Eagleton. 2005. *The English Novel: An Introduction*. Blackwell, Oxford.

David K. Elson and Kathleen R. McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*, Atlanta, Georgia.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Anatoliy Gruzd and Caroline Haythornthwaite. 2008. Automated discovery and analysis of social networks from threaded discussions. In *International Network of Social Network Analysis (INSNA) Conference*, St. Pete Beach, Florida.

Harry Halpin, Johanna D. Moore, and Judy Robertson. 2004. Automatic analysis of plot for story rewriting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, Barcelona.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 472–479, Prague.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanual. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272.

Franco Moretti. 1999. *Atlas of the European Novel, 1800-1900*. Verso, London.

Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, London.

Frederick Mostellar and David L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer, New York.

Raymond Williams. 1975. *The Country and The City*. Oxford University Press, Oxford.