Extracting Social Networks from Literary Fiction

David K. Elson
Nicholas Dames
Kathleen R. McKeown
Columbia University
Overview

- Social network extraction from literature
  - Large corpus of unstructured text
    - 19th Century British literature
  - Conversational network

- Hypotheses to prove or disprove from literary theory:
  1. Larger conversational networks tend to fragment
  2. Less face time (conversation) in cities than rural settings

- Method based on quoted speech
  - Identify who talks to whom
  - Extract graph features that evaluate hypotheses
Contributions

- We demonstrate a high-precision method for conversation network extraction

- Features of the networks provide evidence against both hypotheses
  - High-volume empirical evidence complements traditional literary theory
Table of contents

- Corpus and Literary Hypotheses
- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method
- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Table of contents

- Corpus and Literary Hypotheses
- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method
- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Corpus

- 60 titles, 1800-1903, >10 million words
- Ainsworth, Austen, Braddon, Bronte sisters, Bulwer-Lytton, Collins, Conan Doyle, Dickens, Disraeli, Edgeworth, Eliot, Galt, Gaskell, Gissing, Hardy, Hughes, James, Kingsley, Martineau, Meredith, Mitford, Reade, Scott, Stevenson, Stoker, Thackeray, Trollope, Wilde, Wood
- Mix of social settings (urban, rural), genres, formal properties (1st/3rd person telling)
Hypothesis #1

● Larger conversational networks (with more people) tend to fragment
  ● Franco Moretti: number of characters has a “qualitative, morphological” impact: at 10 or 20 characters, possible to include “distant and openly hostile groups”
  ● Terry Eagleton: in a large community, “most of our encounters consist of seeing rather than speaking, glimpsing each other as objects rather than conversing as fellow subjects” (Introduction to the English Novel)

● Can we show empirically that conversational networks with fewer people are more closely connected?
Hypothesis #2

- Less “face time” in urban novels than rural ones
  - Raymond Williams’s *The Country and The City*: Authors “offer to show people and their relationships in essentially knowable and communicable ways”
  - In the country, face-to-face relations of a restricted set of characters are the primary mode of social interaction
  - Cities promote a “social isolation” from “competitive indifference”

- Can we empirically show more rural conversation and connectedness?
An initial surprise

- The two hypotheses collapse into one if urban novels usually have more characters

- We did not find this to be the case. Urban and rural novels had the same number of characters
  - Counting named character entities with a nontrivial share (>2%) of mentions (non-incidental characters)

- We evaluate hypotheses separately
Table of contents

- Corpus and Literary Hypotheses

- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method

- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Names and Nominals

- **Named entities** chunked with Stanford NER
- **Nominals** chunked with pattern matching
  - Determiner + modifier(s) + organism noun head
    - *Emma, her father, Mr. Knightley, the Governor’s daughter, some one*
- **Coreference for named entities**
  - Variations on long names projected by removing key words (such as titles)
    - *Mr. Ebenezer Scrooge → Scrooge*
- Gender assignment based on titles, first names
  - *Emma* in gender dictionary, *Mr. Knightley*
Table of contents

- Corpus and Literary Hypotheses
- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method
- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Example (Austin)

“Take it,” said Emma, smiling, and pushing the paper towards Harriet– “it is for you. Take your own.”

“Quoted Speech Attribution” problem (QSA)
## Separate QSA Corpus

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th># Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Austen</td>
<td><em>Emma</em></td>
<td>1815</td>
<td>549</td>
</tr>
<tr>
<td>Charles Dickens</td>
<td><em>A Christmas Carol</em></td>
<td>1843</td>
<td>495</td>
</tr>
<tr>
<td>Gustave Flaubert</td>
<td><em>Madame Bovary</em></td>
<td>1856</td>
<td>514</td>
</tr>
<tr>
<td>Mark Twain</td>
<td><em>The Adventures of Tom Sawyer</em></td>
<td>1876</td>
<td>539</td>
</tr>
<tr>
<td>Sir Arthur Conan Doyle</td>
<td>“The Red-Headed League”</td>
<td>1890</td>
<td>524</td>
</tr>
<tr>
<td></td>
<td>“A Case of Identity”</td>
<td>1888</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“The Boscombe Valley Mystery”</td>
<td>1888</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“A Scandal in Bohemia”</td>
<td>1888</td>
<td></td>
</tr>
<tr>
<td>Anton Chekhov</td>
<td>“The Steppe”</td>
<td>1888</td>
<td>555</td>
</tr>
<tr>
<td></td>
<td>“The Lady with the Dog”</td>
<td>1899</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“The Black Monk”</td>
<td>1894</td>
<td></td>
</tr>
</tbody>
</table>

- 111,000 words, 3,176 “quotes” *excerpt*
- Split into 1/3 development, 2/3 training/testing
Corpus Annotation

- **Candidate speakers** extracted for each quote
  - Each name or nominal within 10 paragraphs
- Mechanical Turk annotation
  - 3 annotators shown each quote and its candidates
  - They select speaker or thinker of the quote, if any
  - Can’t select pronouns, only their entities/nominals
- Agreement: Majority vote in 95% of cases
  - 93% of quotes “answerable”
    - Excerpt was long enough, and speaker was chunked
- Data released publicly
  - www.cs.columbia.edu/nlp/tools.cgi
Data “Abstraction”

- Pattern matching to reduce text to backoff symbols
  - Developed using 1/3 of corpus (development set)
  - Symbols are *quote- and candidate-specific*
    - Abstraction performed for every quote-candidate pair
- Formatting normalized
  - Multi-paragraph quotes condensed
  - Clauses with impertinent information removed

### Symbols

- `<TARGET_QUOTE>`
- `<OTHER_QUOTE>`
- `<TARGET_PERSON>`
- `<OTHER_PERSON>`
- `<PRONOUN>`
- `<EXPRESS_VERB>`
Example Abstraction

"Dear Emma bears everything so well," said her father. "But, Mr. Knightley, she is really very sorry to lose poor Miss Taylor, and I am sure she will miss her more than she thinks for."

Emma turned away her head, divided between tears and smiles. "It is impossible that Emma should not miss such a companion," said Mr. Knightley. "We should not like her so well as we do, sir, if we could suppose it; but she knows how much the marriage is to Miss Taylor’s advantage."

<OTHER_QUOTE> <EXPRESS_VERB> <OTHER_PERSON> .
<OTHER_QUOTE>

<OTHER_PERSON> TURNED AWAY HER HEAD .
<TARGET_QUOTE> <EXPRESS_VERB> <TARGET_PERSON> .
<OTHER_QUOTE>
Intuition

- Many quotes can be reliably solved through syntactic patterns.
  - `<TARGET_QUOTE> <EXPRESS_VERB> <TARGET_PERSON> ➔` The person said the quote in 99% of the corpus
- **Dialogue Chains** of context-dependent quotes
  - “Quote,” said person. “Added quote by same person.”

- All these authors notate quotes roughly the same
  - Goal: To not overfit patterns to corpus
  - Estimate: 80-20 rule for “standard quote notation”
# Syntactic Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Rate</th>
<th>Prediction</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backoff</td>
<td>n/a</td>
<td>.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Added Quote</td>
<td>&lt;QUOTE by Person1&gt; &lt;TARGET_QUOTE&gt;</td>
<td>.19</td>
<td>Person1</td>
<td>.95</td>
</tr>
<tr>
<td>Apparent Conversation</td>
<td>&lt;QUOTE by Person1&gt; &lt;QUOTE by Person2&gt; &lt;TARGET_QUOTE&gt;</td>
<td>.18</td>
<td>Person1</td>
<td>.96</td>
</tr>
<tr>
<td>Quote-Said-Person</td>
<td>&lt;TARGET_QUOTE&gt; &lt;VERB&gt; &lt;TARGET_PERSON Person1&gt;</td>
<td>.17</td>
<td>Person1</td>
<td>.99</td>
</tr>
<tr>
<td>Quote alone</td>
<td>Quote alone in paragraph.</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaphora</td>
<td>&lt;TARGET_QUOTE&gt; &lt;EXPRESS_VERB&gt; &lt;PRONOUN&gt;</td>
<td>.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quote-Person-Said</td>
<td>&lt;TARGET_QUOTE&gt; &lt;TARGET_PERSON Person1&gt; &lt;EXPRESS_VERB&gt;</td>
<td>.02</td>
<td>Person1</td>
<td>.92</td>
</tr>
</tbody>
</table>

(Development set)
Initial Result: Notational Consistency

- 57% of the testing set is covered by the categories with predictions

- The category predictions solve these with 96% accuracy
Applying Learning Tools

- One model per category

- Features extracted are *quote- and candidate-specific*
  - Words and other characters between candidate and quote
  - Formatting, punctuation
  - Prevalence of other quotes by speaker
  - Prevalence of mentions of candidate
  - Number of names, quotes, words nearby
  - Type of word found immediately before and after quote
  - Length, other features about quote itself

- Binary classification into *speaker* or *non-speaker*
  - Each candidate a separate data point
Learning Parameters

- Multiple classifiers (from WEKA)
  - Logistic regression
  - J48
  - JRip

- Multiple ways to normalize feature vectors
  - None
  - Usual vector normalization (by unit vector)
  - Find average vector for all candidates considered for a quote; take distance between candidate and group norm

- Cross-validation on training/test set (2/3 of corpus)
Reconciliation methods

- From many binary decisions to one candidate selection:
  - Zero “candidate” $\rightarrow$ no speaker, one “candidate” $\rightarrow$ selection, 2+ “candidates” $\rightarrow$ “unknown”
  - Find probabilities of speaker classification; take top-scoring candidate (above a threshold)
  - Hybrid of the above
  - Combine probabilities from multiple classifiers; take top-scoring candidate (above a threshold)
### Results, testing set

<table>
<thead>
<tr>
<th>Syntactic Category</th>
<th>Rate</th>
<th>Top-performing solvers</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quote-Said-Person</td>
<td>.22</td>
<td><strong>Syntactic category’s prediction</strong> Logistic + J48 combined</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.96</td>
</tr>
<tr>
<td>Added quote</td>
<td>.19</td>
<td><strong>Syntactic category’s prediction</strong> J48</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.97</td>
</tr>
<tr>
<td>Backoff</td>
<td>.18</td>
<td>Logistic+J48+JRip</td>
<td>.64</td>
</tr>
<tr>
<td>Quote alone</td>
<td>.16</td>
<td>Logistic+J48+JRip</td>
<td>.63</td>
</tr>
<tr>
<td>Apparent conversation</td>
<td>.12</td>
<td>JRip Syntactic category’s prediction</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.91</td>
</tr>
<tr>
<td>Anaphora trigram</td>
<td>.09</td>
<td>Logistic</td>
<td>.63</td>
</tr>
<tr>
<td>Quote-person-said</td>
<td>.04</td>
<td>JRip Syntactic category’s prediction</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.93</td>
</tr>
<tr>
<td>Overall (using <strong>bold</strong> solvers)</td>
<td></td>
<td></td>
<td><strong>.83</strong></td>
</tr>
<tr>
<td>Baseline: Assume most recent character mention is speaker</td>
<td></td>
<td></td>
<td>.45</td>
</tr>
<tr>
<td>Baseline: Assume closest character mention is speaker</td>
<td></td>
<td></td>
<td>.52</td>
</tr>
</tbody>
</table>
Table of contents

- Corpus and Literary Hypotheses
- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method
- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Precision-based metric

- Precision matters more than recall
  - False positives will give a false sense of connectedness
  - This tilts “in favor” of proving hypotheses (smaller networks are more connected)
  - Nonetheless, our results provide negative evidence

- Can still find connections
  - Need at least one “hit” to find a relationship
Conversation detection

- *Quote adjacency* heuristic for detecting conversations (edges in social network)

- Face-to-face conversations involve sequential quotes, possibly separated by non-quotes
  - Parameterized search radius (to tolerate intermediate quotes by others)

- Edge weight set to share of detected conversations
  - Node width set to share of name mentions
Edge weight

- Interwoven dialogue indicates substantial conversation

- High word count inside quotes adds to evidence of relationship
Meredith, The Egoist

Oval width proportional to share of name mentions
Edge weight proportional to share of spoken dialogue
Austen, *Pride and Prejudice*
Table of contents

- Corpus and Literary Hypotheses
- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method
- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Accuracy-check design

- Two chapters taken from four works for manual annotation and evaluation of accuracy
  - *The Sign of the Four, Emma, David Copperfield* and *The Portrait of a Lady*
  - Over 40K words, 3 annotators

- Annotators told to identify *all* pairwise conversation edges
  - Including described (unquoted) speech
  - Require characters to be in the same place at the same time, speak in turns, be mutually aware

- Cast as a binary classification problem
  - In N x N matrix of characters
Accuracy-check Baselines

- **Baseline: Correlation heuristic**
  - Divide text into bins, count character mentions in each bin
  - Correlations in distributions between characters indicate relationship

- **Baseline: Spoken mention heuristic**
  - Count one character’s mention of another
    - Parameterize threshold for relationship
Accuracy-check results

- High inter-annotator agreement
  - Unanimous agreement in 95% of cases, Kappa=.82
  - 9% of possible interactions occurred

- Adjacency algorithm outperforms baselines

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quoted speech adjacency</td>
<td>.95</td>
<td>.51</td>
<td>.67</td>
</tr>
<tr>
<td>Correlation baseline</td>
<td>.21</td>
<td>.65</td>
<td>.31</td>
</tr>
<tr>
<td>Spoken-mention baseline</td>
<td>.45</td>
<td>.49</td>
<td>.47</td>
</tr>
</tbody>
</table>

- Error analysis: Mix of reasons for misses
  - Indirect speech, anaphoric attributions, group conversations without adjacencies between all pairs
Table of contents

- Corpus and Literary Hypotheses
- Extracting Characters and Conversations
  - Chunking character names and nominals
  - Attributing quoted speech fragments to names
  - Detecting conversations between characters
  - Checking the accuracy of the extraction method
- Using Networks to Check Hypotheses
  - Extracting features
  - Results and Conclusions
Questions: Reminder

1) Do networks with more people fragment or stay connected?

2) Are rural networks more conversational (less isolating) than urban networks?
Variables for Correlation

- Number of characters, speaking or non-speaking
- Distribution of quotes among characters (even or lopsided)
- Amount of dialogue (normalized)
  - Cliques (number, rate)
- Average degree for each vertex
  - “How many conversation partners does each person have?”
- Graph density (normalized average degree)
  - “With what percent of the entire network does each person converse?”
Hypothesis 1: Impact of Network Size

- As the number of named characters increases, we expect:
  - Same or less total speech
    - Weak yes: Normalized number of quotes flat at $r=.16$
  - Less lopsided distribution of quotes among speakers
    - Yes: Share of quotes by top 3 speakers decreases at $r=-.61$
Hypothesis 1: Impact of Network Size

- As the number of named characters increases, we expect:
  - Lower density (fewer conversational partners as percentage of population)
    - **No**: Increases at $r=.30$. Larger networks are more connected
  - Same or fewer cliques
    - **No**: 3-clique rate increases at $r=.38$. Larger networks form cliques more often
Hypothesis 1: Impact of Network Size

- As the number of speakers increases, we expect:
  - Less overall dialogue ("glimpsing rather than speaking")
    - **No**: Increases at .50. Larger networks are more talkative
  - Lower density
    - **No**: Increases at .49. In larger networks, people know more of their neighbors
Hypothesis 2: Urban vs. Rural

- No variable changes significantly between urban and rural settings
  - Not even # characters and # speakers
  - Urban/rural setting works independently of large/small cast of characters
Alternate Explanation

- Surprisingly, text perspective dominates the shape of the network
  - 3rd person tellings: Significant increases in
    - Normalized number of quotes (p<.05)
    - Average degree (p<.005)
    - Graph density (p<.05)
    - Rate of 3-cliques (p<.005)
      - ...With no significant difference in number of characters or speakers
  - Hypothesis: First-person narrators not privy to other characters’ conversations with each other

- Caveat: Quoted-speech-adjacency is likely a sensitive metric for this effect
  - Extant relationships shift to reported, indirect speech
3rd Person Narrative

(Austen, *Persuasion*)

![Character Diagram]
“Close 3rd” Narrative
(Braddon, Lady Audley’s Secret)
1\textsuperscript{st} Person Narrative

(Dickens, \textit{David Copperfield})