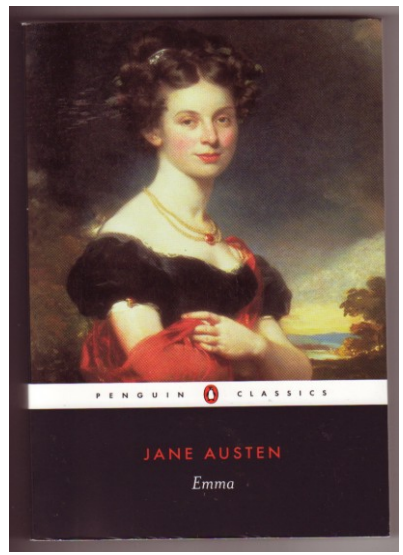
Extracting Social Networks from Literary Fiction

David K. Elson, Nicholas Dames, Kathleen R. McKeown

Presented by Audrey Lawrence and Kathryn Lingel

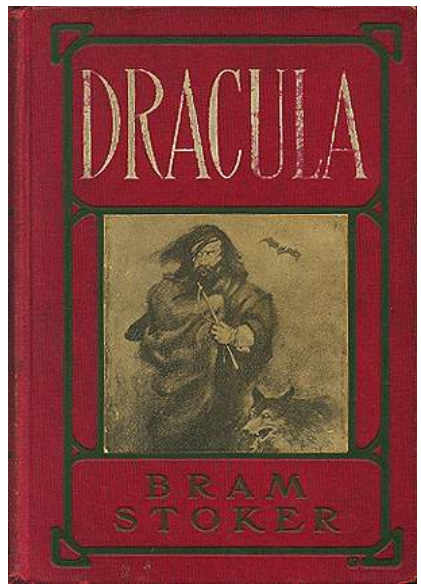
Introduction

- Network of 19th century novel's social structures
- Previous hypotheses
- No automated work on many novels
- Construct network based on dialogue
- Evaluate based on network



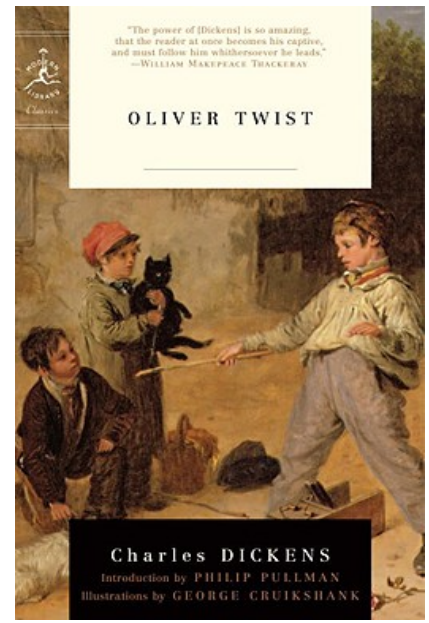
Related Work

- With computer, word based
  - Identifying author
  - Writing style
  - Lineage of ancient text
- Semantically oriented is rare
  - Sequences in news stories
- Models for novels without computation
- Computation based models:
  - ACE: unstructured text
  - Other structured



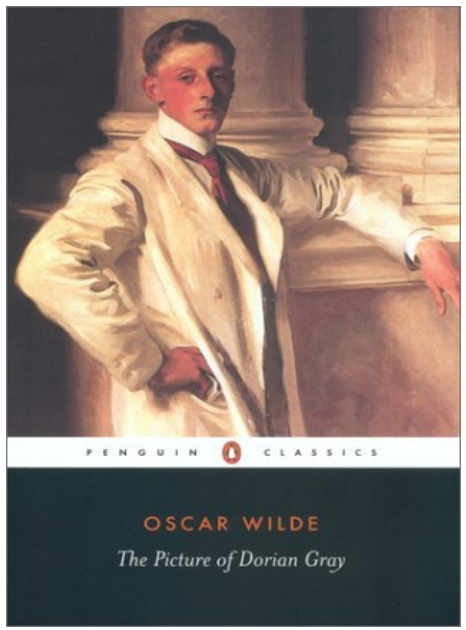
Why 19th Century Novels?

- Novelistic innovations
- Actual social changes
  - Revolutions
  - Industry
  - Transportation
- Many theorists, yet no use of many novels



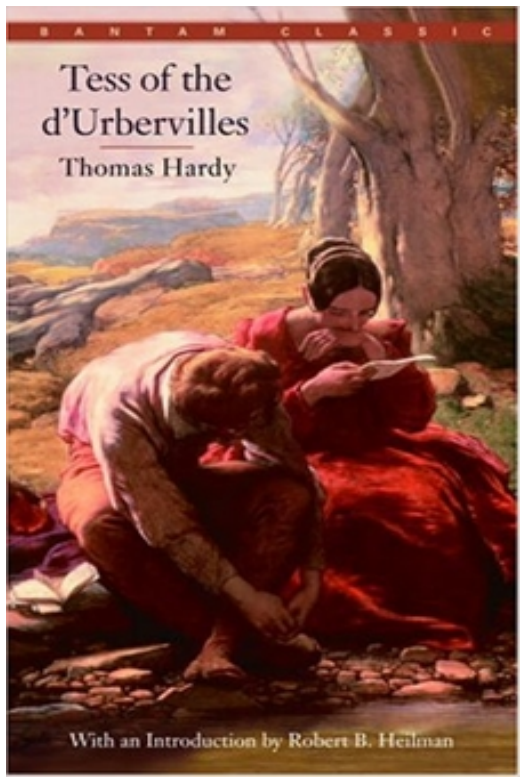
Past Theories

- Bakhtin: "chronotope", quality of interactions change by setting
- Williams: "knowable communities", rural is more connected with less characters but more dialogue
- Moretti: urban communities are more complex and larger and have more interactions without dialogue



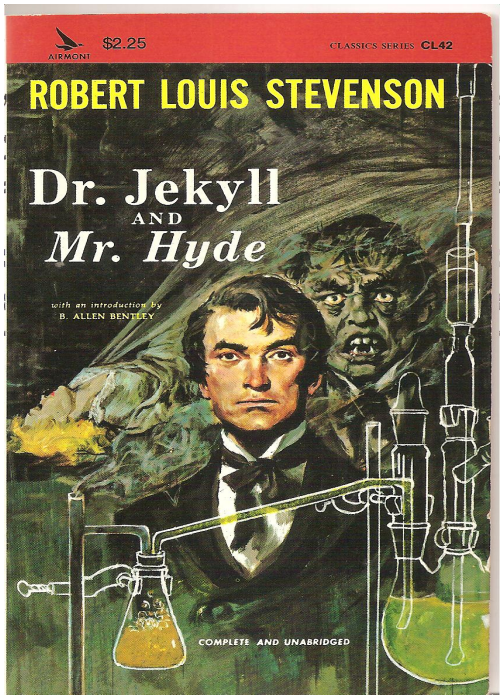
Novels

- 60
- By: authorial, historical, generic, sociological, technical
- Over 10 million words
- Urban vs Rural
- 1st person vs 3rd person



Hypotheses

- Inverse correlation between number of characters and amount of dialogue
- Differences are based upon geographical setting
  - Urban: more loose with more characters and less conversation
  - Rural: more tightly bound



Extracting Networks

- Create graphs
  - Characters as vertices
  - Dialogues as edges
  - Weights as amount of dialogue
- Conversation if:
  - Same place and time
  - Turns speaking
  - Mutually aware of one another
- Preprocess text first

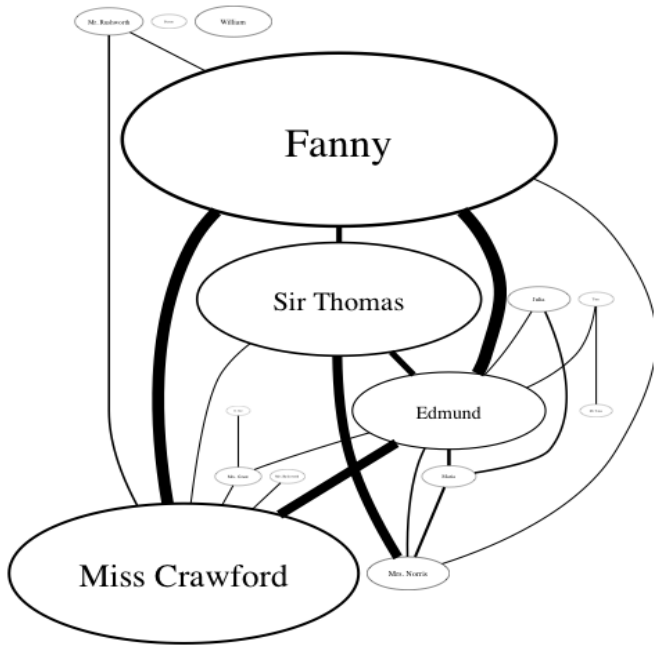
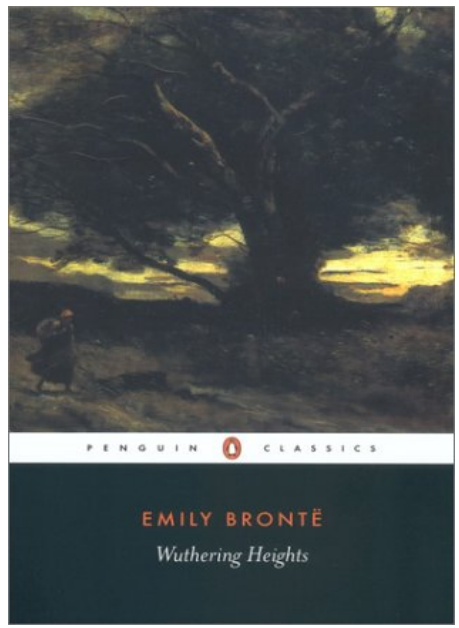
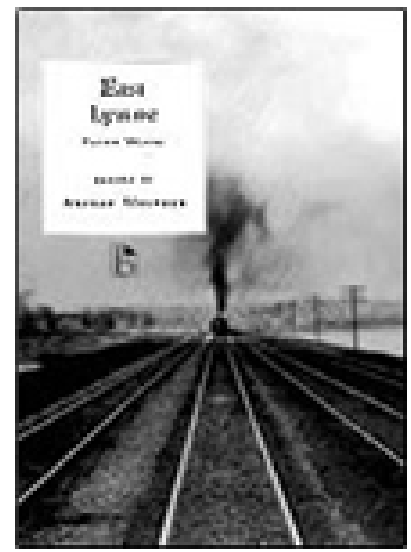


Figure 1: Automatically extracted conversation network for Jane Austen's *Mansfield Park*.



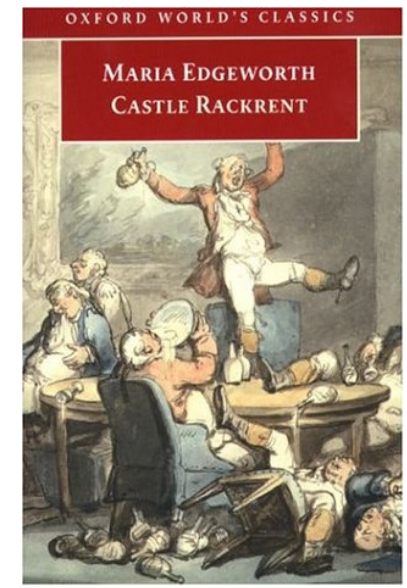
Character Identification

- Chunk names from text
- Stanford Ner tagger to identify noun phrases as people or organizations
- Cluster names
  - generate name variants for each
    - I.e. "Audrey", "Audrey Lawrence", "Ms. Lawrence"
    - or "Kathryn", "Kathryn Lingel", "Ms. Lingel"
  - try to find matches from entity list



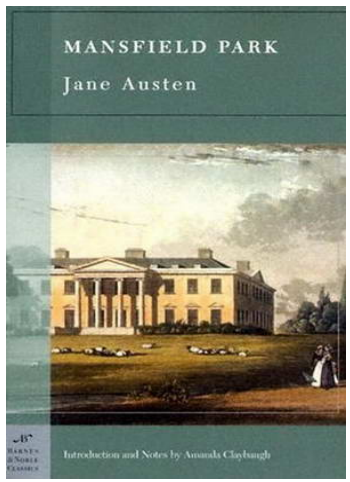
Quoted Speech Attribution

- Creation of training and test sets
  - 111,000 words and 3,176 quotes
- 3 annotators for each quote
- Trained to develop a categorizer
  - 5 categories
  - For example, "character trigram" is one with 99% accuracy
  - 5th category encompasses rest
- 57% recall
- 96% accuracy
- Low recall is ok because we are concerned with conversations, not single quotes
- Precision is necessary
- This setup tilts in favor of first hypothesis



Network Construction

- Remove entities mentioned < 3 times or in less than 1% of mentions
- Adjacent if within 300 words and no attributed quotes in between
- Weight is the length of the quote, normalized to length of novel



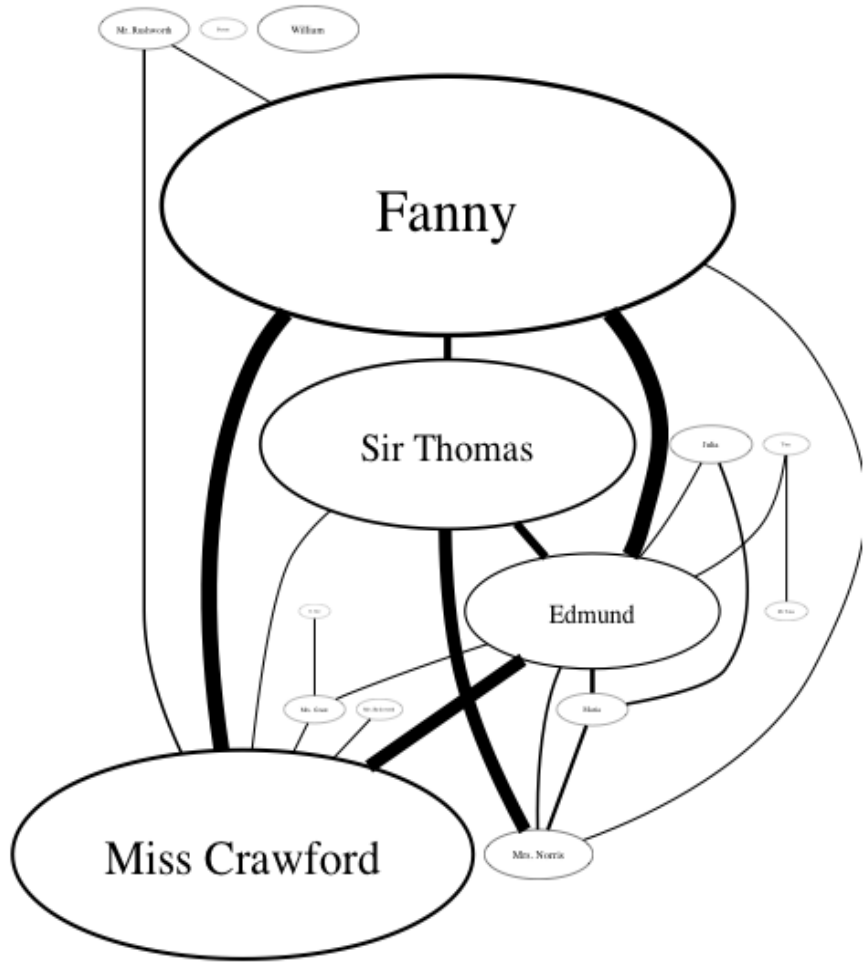
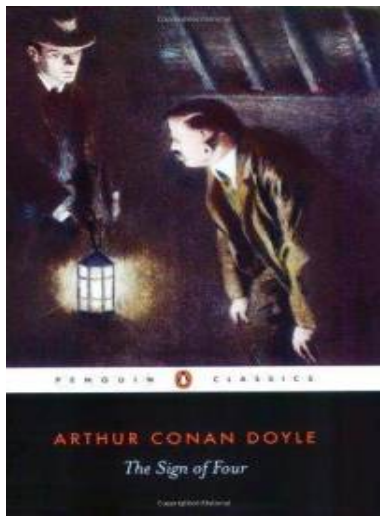


Figure 1: Automatically extracted conversation network for Jane Austen's *Mansfield Park*.

Alternate Methods

- Correlation
  - Divide text into 10 paragraph sections
  - Count mentions
  - Compute Pearson product-moment correlation coefficient
- Spoken Method
  - Count when one refers to another within a quote

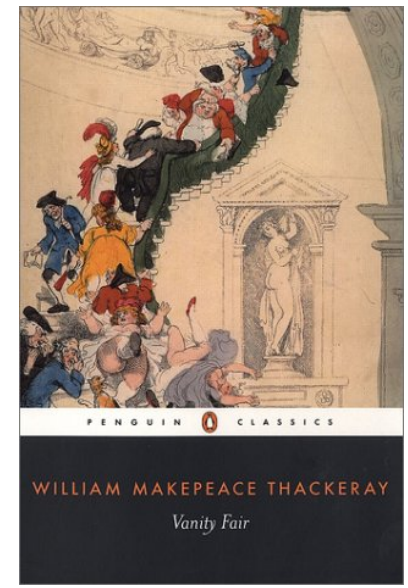


Evaluation

- Check accuracy of extraction
- Picked 4-5 random chapters from 4 novels
- Over 10,000 words/novel
- 3 annotators

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.



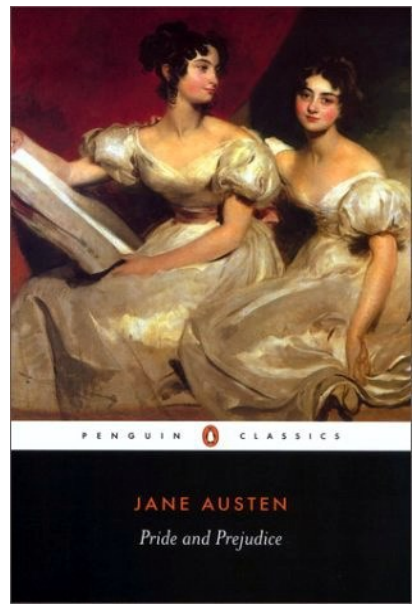
Data Analysis: Features

- Number of characters/speaking characters
- Variance of distribution of quoted speech
- Number of quotes given number of words
- Number of 3-cliques or 4-cliques
- Average Degree

$$\frac{\sum_{v \in V} |E_v|}{|V|} = \frac{2|E|}{|V|}$$

- Graph Density

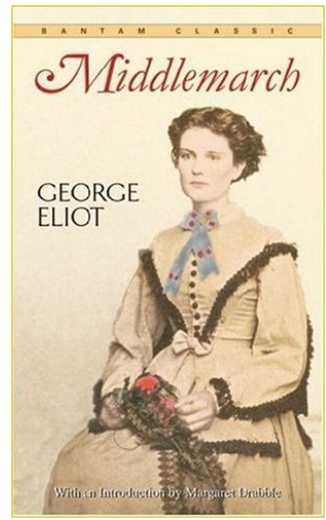
$$\frac{\sum_{v \in V} |E_v|}{|V|(|V|-1)} = \frac{2|E|}{|V|(|V|-1)}$$



Data Analysis: Hypothesis Results

- Hypothesis 1: inverse correlation between number of characters and amount of dialogue
  - Not supported
  - Positive correlations found instead
    - Number of characters vs number of quotes

- Hypothesis 2: setting (urban or rural) affects the network
  - Not supported
  - All features were statistically similar



Data Analysis: Results

- Perspective: 1st vs 3rd

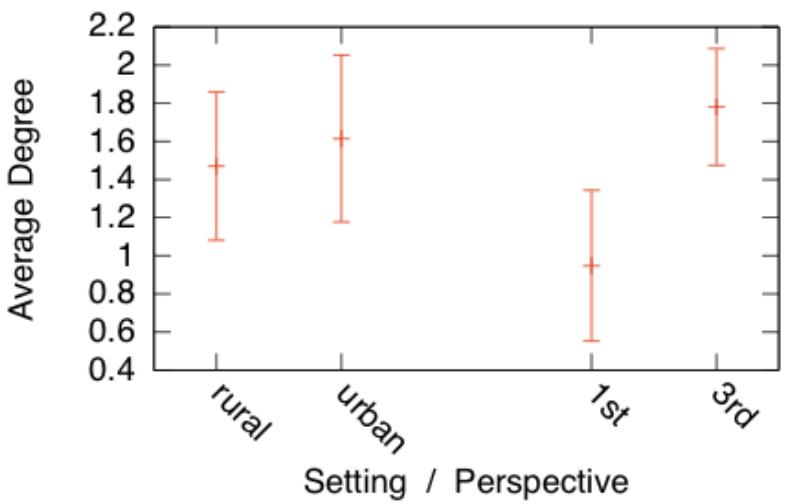


Figure 2: The average degree for each character as a function of the novel's setting and its perspective.

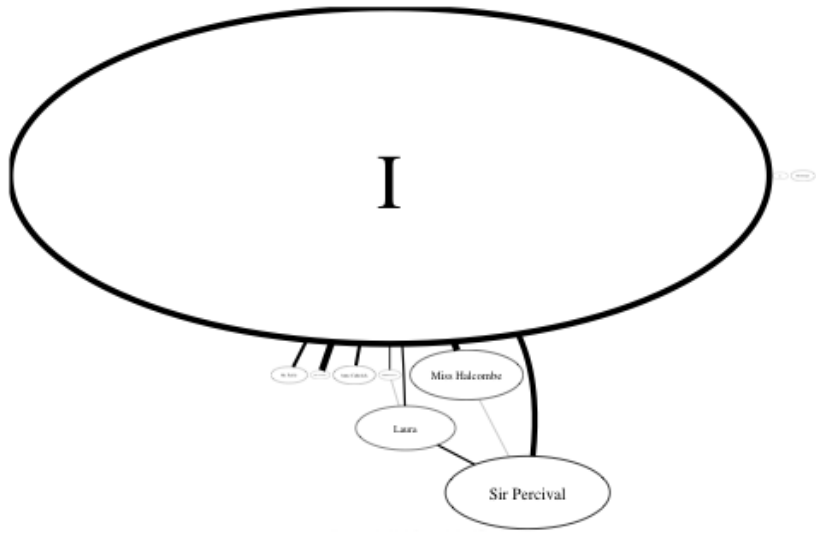
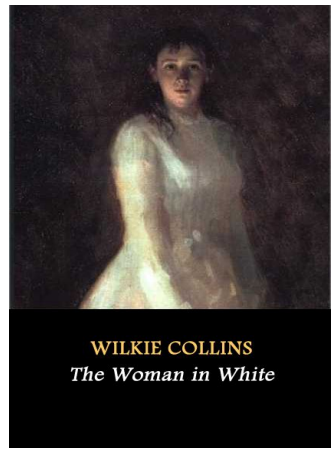
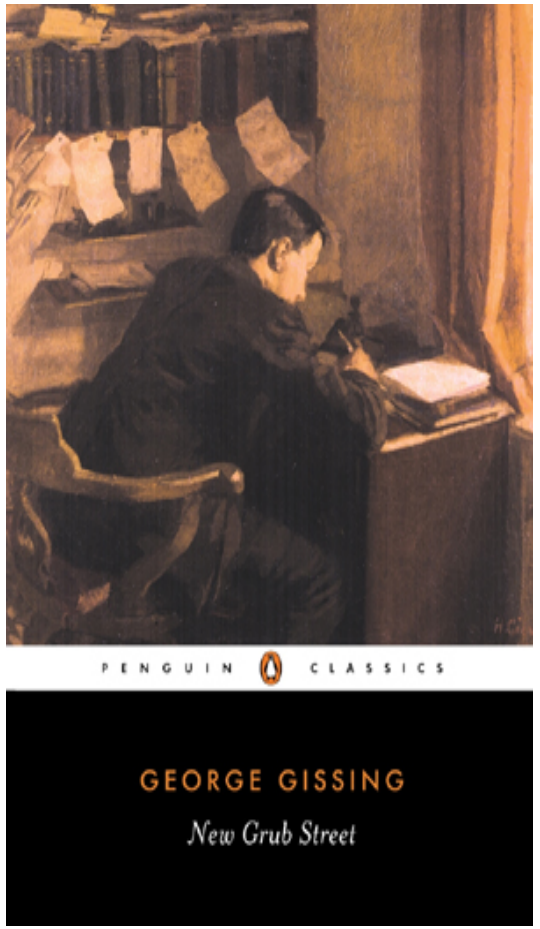


Figure 3: Conversational networks for first-person novels like Collins's *The Woman in White* are less connected due to the structure imposed by the perspective.



Literary Analysis

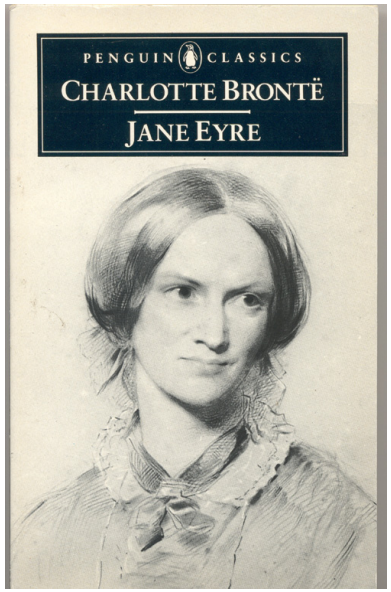
- Narrative voice trumps setting

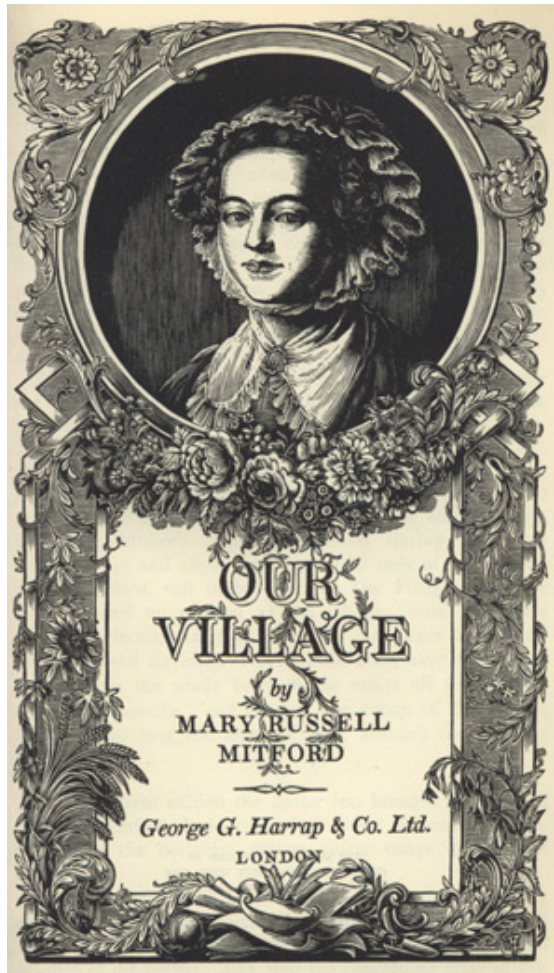


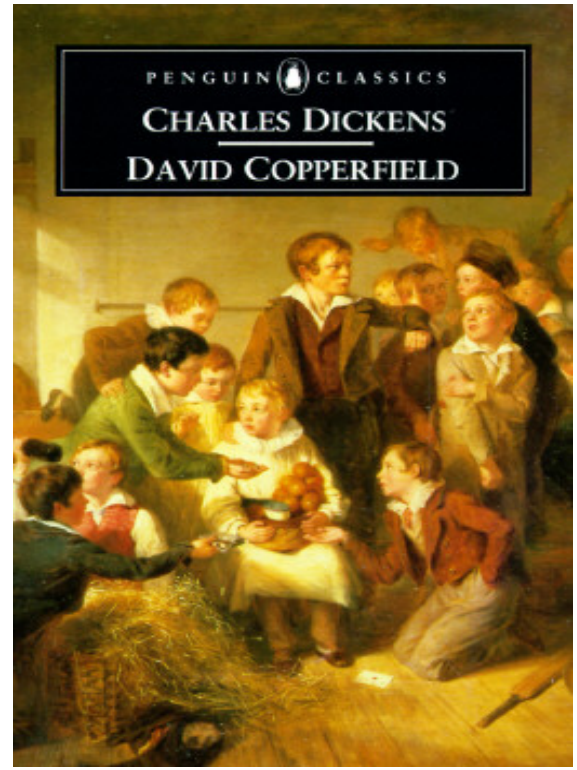
Conclusion

- Developed system to automatically create social networks from novels
- High precision, low recall
- Found hypotheses were not supported
- Yet correlation between narrative voice and network structure

Questions?


PENGUIN CLASSICS
CHARLOTTE BRONTË
JANE EYRE


OUR VILLAGE
by
MARY RUSSELL MITFORD
George G. Harrap & Co. Ltd.
LONDON


PENGUIN CLASSICS
CHARLES DICKENS
DAVID COPPERFIELD