Modeling the Spread of Information within Novels

David Bamman
School of Information, UC Berkeley
dbamman@berkeley.edu

In collaboration with Matt Sims, Olivia Lewke, Anya Mansoor, Sejal Popat and Sheng Shen

Fiction as data

- BookCorpus (self-publishing)
  Zhu et al. 2015
- NarrativeQA
  Kočiský et al. 2017
- Commonsense stories
  Mostafazadeh et al. 2016
- Google Books
  Michel et al. 2010; Goldberg and Orwant 2013

Modeling literary phenomena

- Character types
  Bamman et al. 2013, 2014
- Relationships
  Iyyer et al. 2016, Chaturvedi et al. 2017
- Sentiment/plot
  Elsner 2012, Mohammad 2011, Jockers 2015, Reagan et al. 2018
- Character psychology
  Rashkin et al. 2018

Computational Humanities

Ted Underwood (2018), “Why Literary Time is Measured in Minutes”

Algee-Hewitt et al. (2016), “Canon/Archive: Large-Scale Dynamics in the Information Field”

Richard Jean So and Hoyt Long (2015), “Literary Pattern Recognition”

Ted Underwood, David Bamman and Sabrina Lee, The Transformation of Gender in English-Language Fiction (2018)

Holst Katsma (2014), Loudness in the Novel

So et al (2014), “Cents and Sensibility”

Matt Wilkens (2013), “The Geographic Imagination of Civil War Era American Fiction”

Jockers and Mimno (2013), “Significant Themes in 19th-Century Literature,”

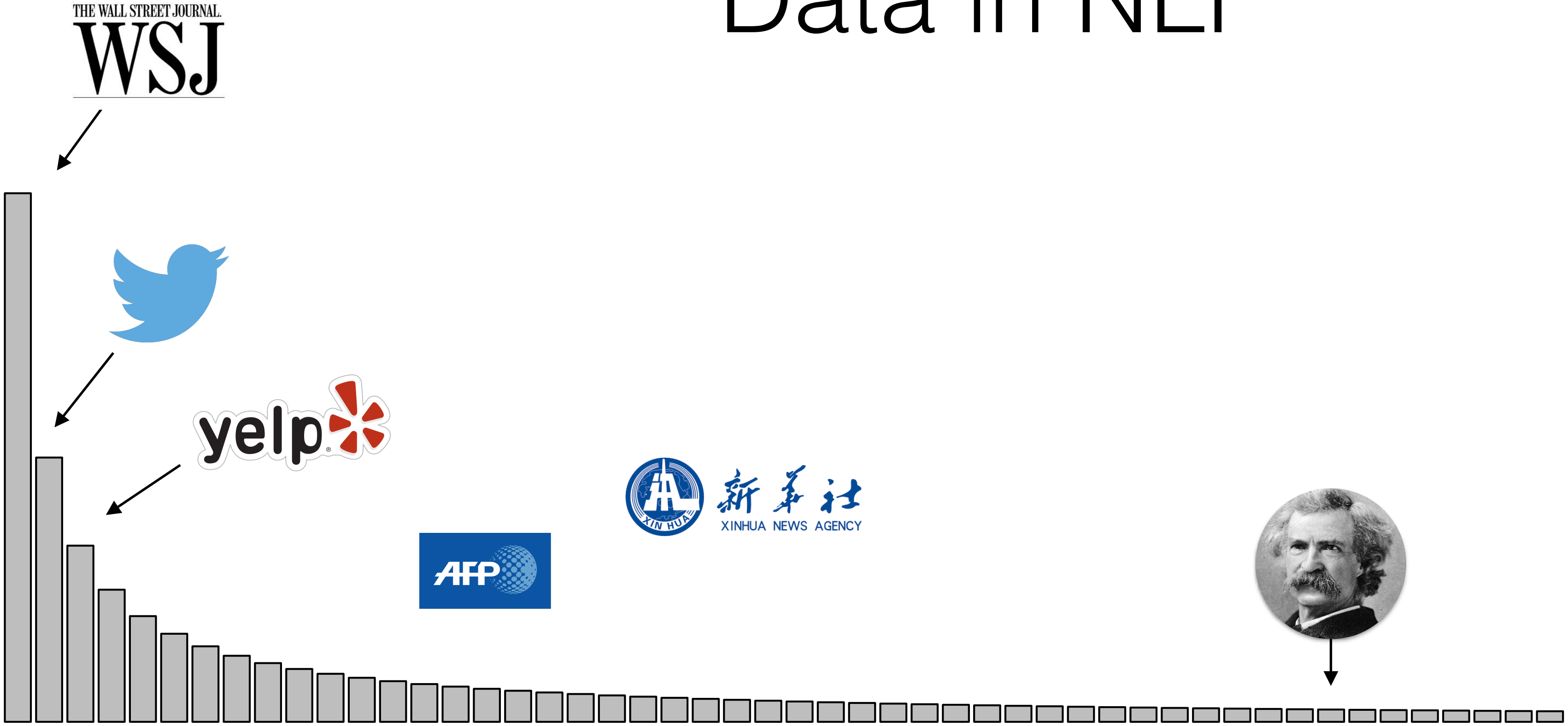
Ted Underwood and Jordan Sellers (2012). “The Emergence of Literary Diction.” JDH

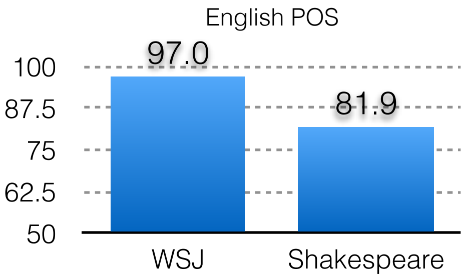
How do we design NLP to drive insight into literary texts?

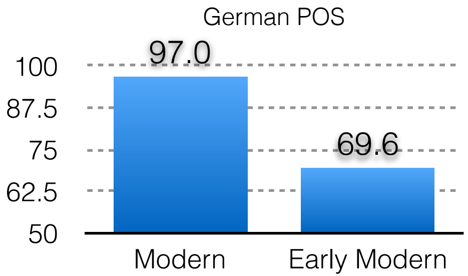
NLP Pipeline

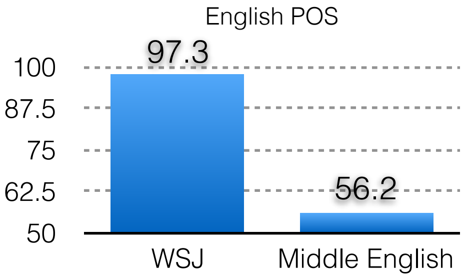
NLP Task	**Accuracy**
Tokenization	100%
Part-of-speech tagging	98.0% [Bohnet et al. 2018]
Named entity recognition	93.1 [Akbik et al. 2018]
Syntactic parsing	95.1 F [Kitaev and Klein 2018]
Coreference resolution	73.0 F [Lee et al. 2018]

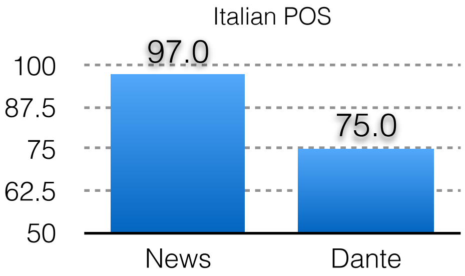
Data in NLP

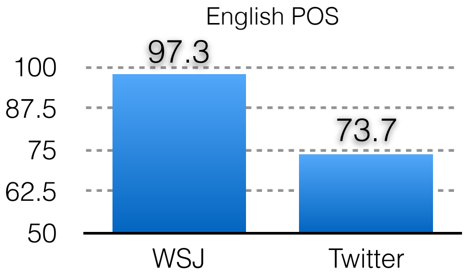

THE WALL STREET JOURNAL.
WSJ
yelp
AFP
XIN HUA
新華社
XINHUA NEWS AGENCY

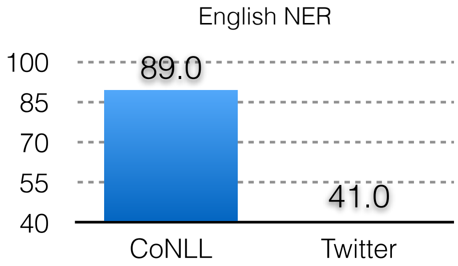

English POS
100
87.5
75
62.5
50
97.0
81.9
WSJ
Shakespeare

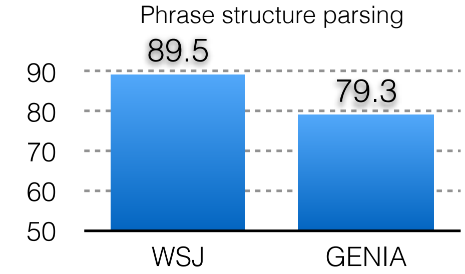

German POS
100
87.5
75
62.5
50
97.0
69.6
Modern
Early Modern

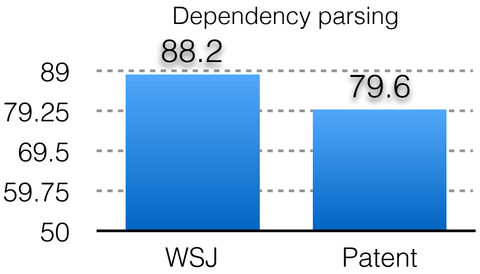

English POS
100
87.5
75
62.5
50
97.3
56.2
WSJ
Middle English

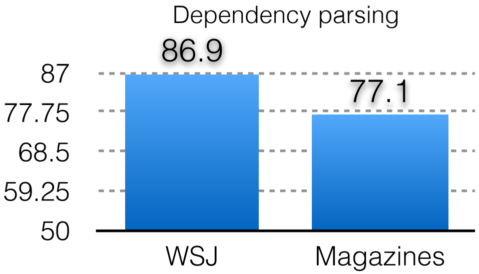

Italian POS
100
87.5
75
62.5
50
97.0
75.0
News
Dante


English POS
100
87.5
75
62.5
50
97.3
73.7
WSJ
Twitter


English NER
100
85
70
55
40
89.0
41.0
CoNLL
Twitter


Phrase structure parsing
90
80
70
60
50
89.5
79.3
WSJ
GENIA


Dependency parsing
89
79.25
69.5
59.75
50
88.2
79.6
WSJ
Patent


Dependency parsing
87
77.75
68.5
59.25
50
86.9
77.1
WSJ
Magazines

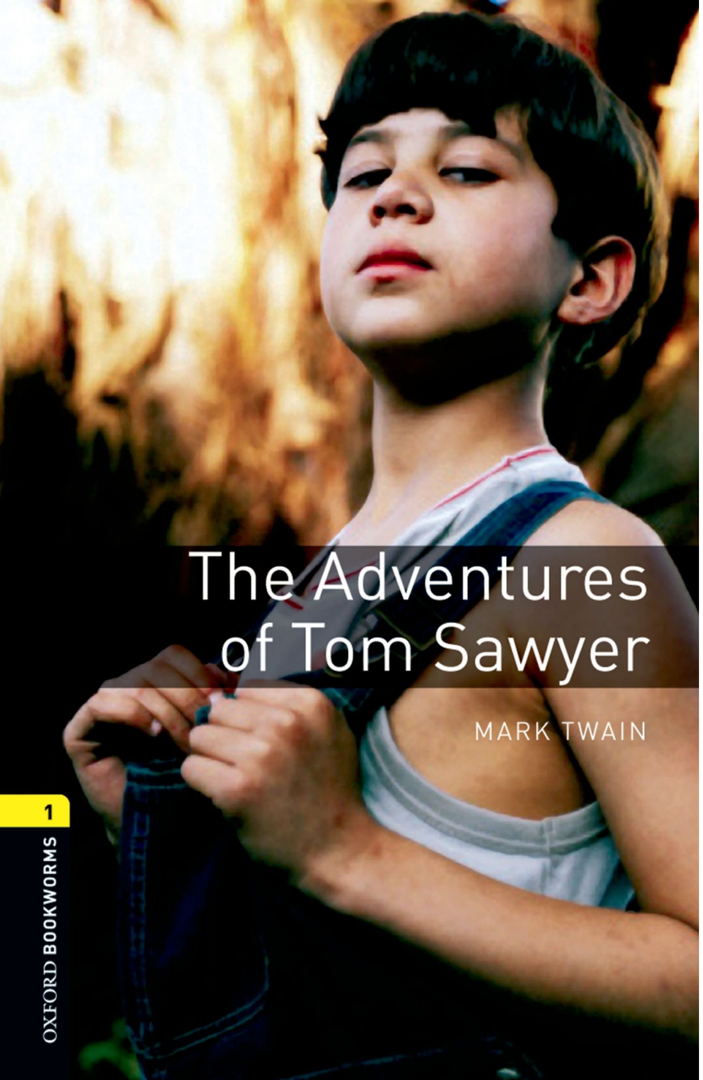
Active work

- Domain adaptation
  [Chelba and Acero, 2006; Daumé and Marcu, 2006; Daumé 2009; Duong et al. 2015; Glorot et al. 2011, Chen et al. 2012, Yang and Eisenstein 2014, Schnabel and Schütz 2014]

- Contextualized word representations
  [Peters et al. 2018; Devlin et al. 2018; Howard and Ruder 2018; Radford et al. 2019]

- Data annotation. 210,532 tokens from 100 different novels, annotated for:
  - Entities (person/place, etc.)
  - Events
  - Coreference
  - Quotation attribution

} available on Github now



Literary entities

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Literary entities

Most work in NLP focuses on *named* entity recognition — mentions of specific categories (person, place, organization) that are explicitly named.

> Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, *Emma*

Entity recognition

- Mr. Knightley
- a sensible man about seven or eight-and-thirty
- a very old and intimate friend of the family
- the family
- Isabella
- Isabella’s husband
- the elder brother of Isabella’s husband

> Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, *Emma*

Entity recognition

- Mr. Knightley
- a sensible man about seven or eight-and-thirty
- a very old and intimate friend of the family
- the elder brother of Isabella’s husband

- the family

- Isabella

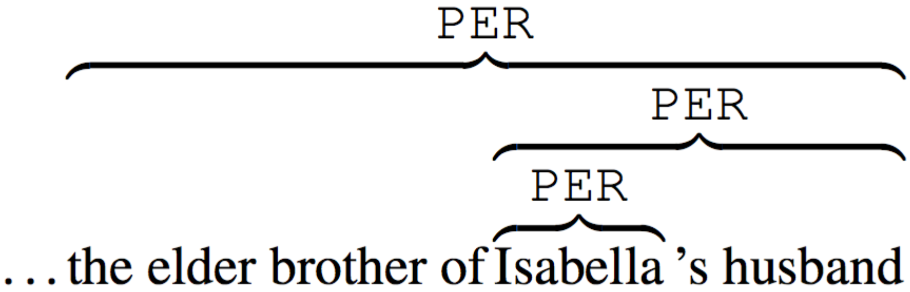
- Isabella’s husband

Mr. Knightley, a sensible man about seven or eight-and-thirty, was not only a very old and intimate friend of the family, but particularly connected with it, as the elder brother of Isabella's husband.

Austen, *Emma*

Nested entity recognition

- Recognize spans of text that correspond to categories of entities (whether named or not).



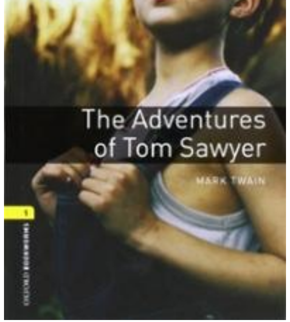
…the elder brother of Isabella ’s husband

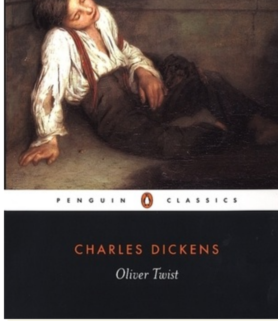
Dataset

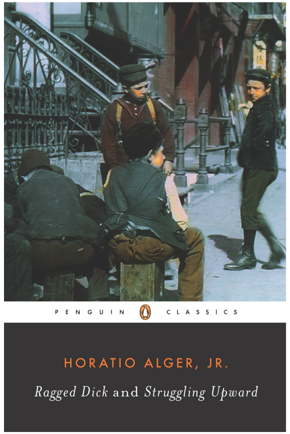
- 100 books from Project Gutenberg

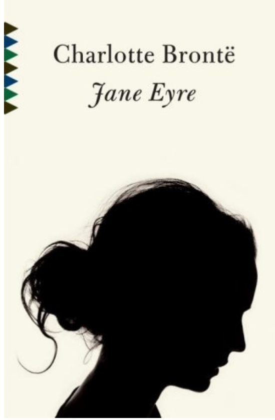
- Mix of high literary style (e.g., Edith Wharton's *Age of Innocence,* James Joyce's *Ulysses)* and popular pulp (Haggard's *King Solomon's Mines*, Alger's *Ragged Dick).*

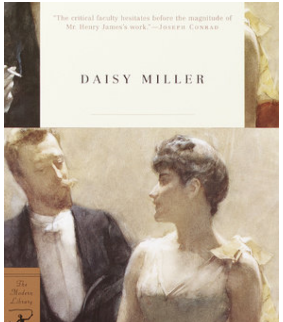
- Select first 2000 words from each text

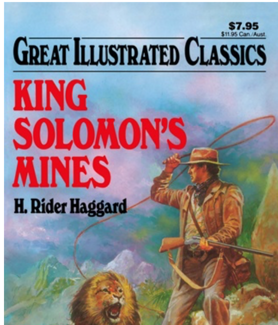






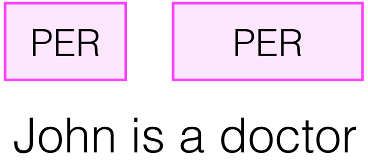



"The critical faculty hesitates before the magnitude of Mr. Henry James's work."—Joseph Conrad
DAISY MILLER



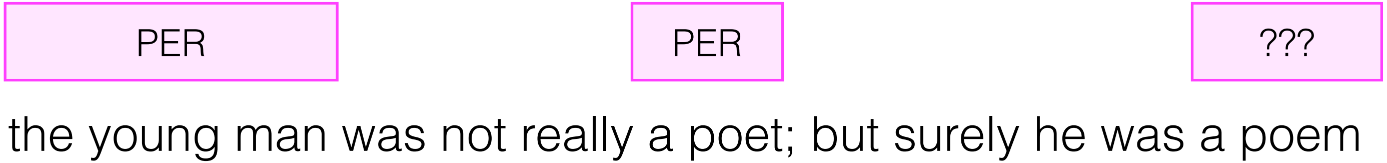
Metaphor

- Only annotate copular phrases whose types denotes an entity class.



PER PER

John is a doctor



PER PER ???

the young man was not really a poet; but surely he was a poem

Chesterton, *The Man Who Was Thursday*

Personification

- Person includes characters who engage in dialogue or have reported internal monologue, regardless of human status (includes aliens and robots as well).

> As soon as I was old enough to eat grass my mother used to go out to work in the daytime, and come back in the evening.

Sewell, *Black Beauty*

Data

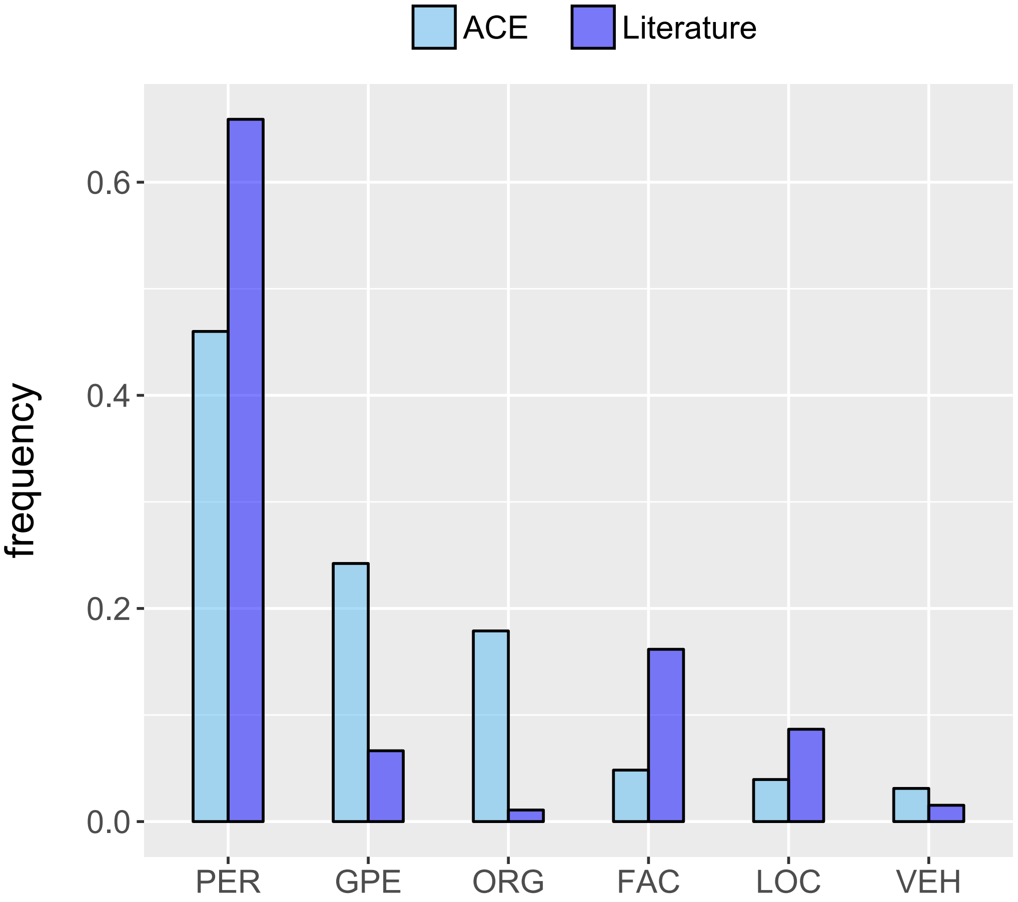
Cat	Count	Examples
PER	9,383	my mother, Jarndyce, the doctor, a fool, his companion
FAC	2,154	the house, the room, the garden, the drawing-room, the library
LOC	1,170	the sea, the river, the country, the woods, the forest
GPE	878	London, England, the town, New York, the village
VEH	197	the ship, the car, the train, the boat, the carriage
ORG	130	the army, the Order of Elks, the Church, Blodgett College

Prediction

How well can find these entity mentions in text as a function of the training domain?

Data

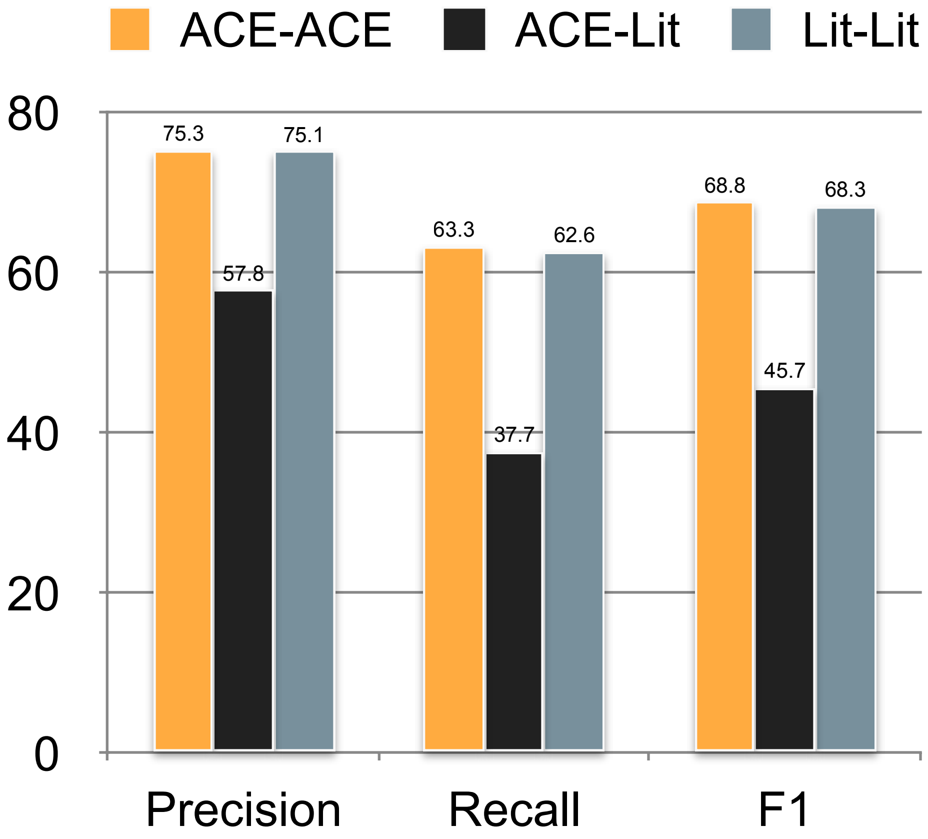
- ACE (2005) data from newswire, broadcast news, broadcast conversation, weblogs



Prediction

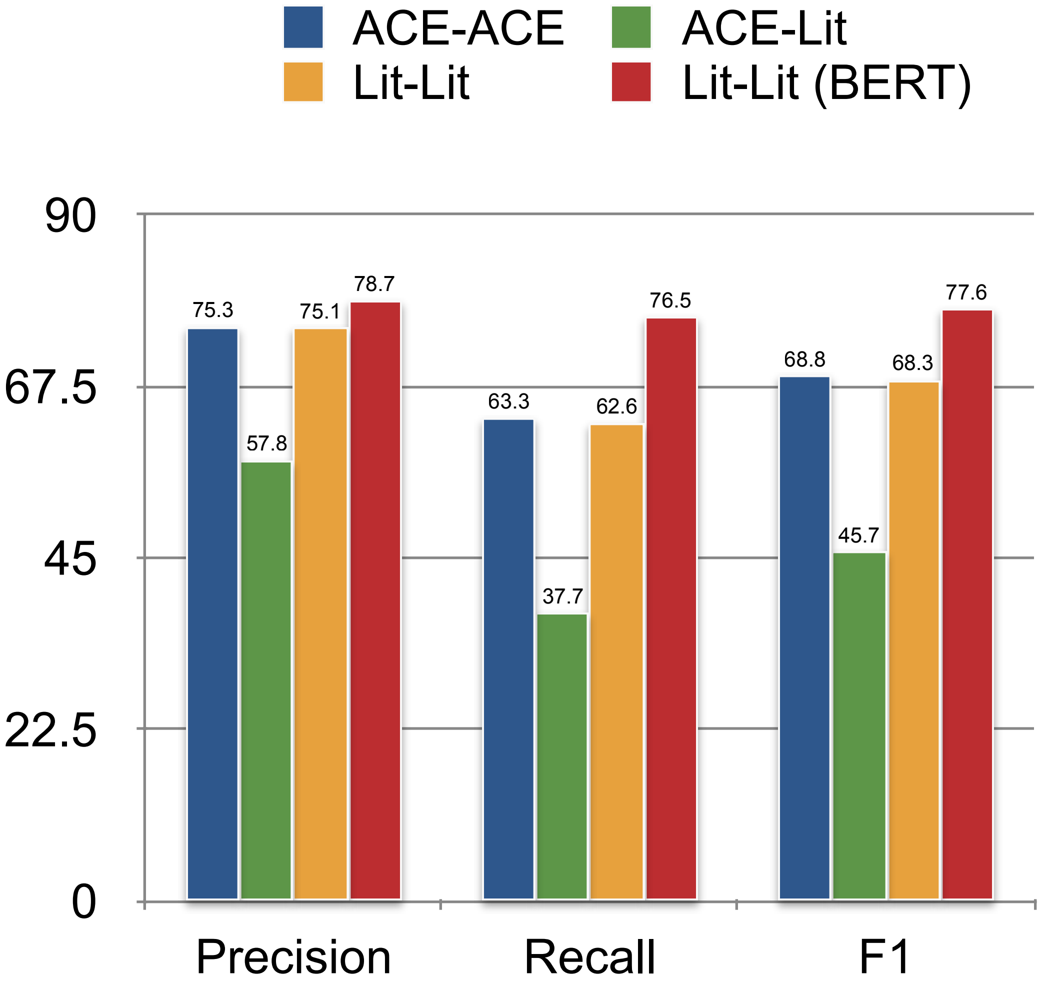
- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.

- Evaluate performance difference when altering the training/test domain.



Prediction

- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.
- Evaluate performance difference when altering the training/test domain.
- Adding BERT contextual embeddings (Devlin et al. 2019) yields +9.3 F1 score



Analysis

- Tag entities in 1000 new Gutenberg texts (78M tokens) using the two models (ACE vs. LIT) and analyze the difference in frequencies with which a given string is tagged as `PER` under both models.

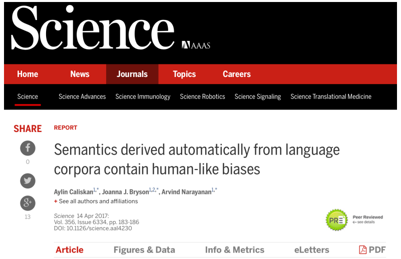
Mrs.
Miss
Lady
Aunt

Gender bias

WIKIPEDIA
The Free Encyclopedia

- 15% of Wikipedia biographies are of women

- Women's biographies are 2.58x more likely to mention divorce, 1.57x more likely to mention marriage [Bamman and Smith 2015]



REPORT

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]
+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

- Word embeddings encode cultural bias implicit in natural language usage [Caliskan et al. 2017; Bolukbasi et al. 2016]

MOSCOW, April 17 (AFP)

Silence is golden -- especially when your hand is weak -- top Moscow policy analysts said in an assessment of the fallout from Russia's vocal opposition to what turned out to be a swift US-led campaign in Iraq.

Several top diplomacy experts told a Kremlin-run forum that countries like China and India that said little about the conflict before its March 20 launch were already reaping the benefits.

Some suggested that Russian President Vladimir Putin will now be scrambling to contain the damage to his once-budding friendship with US President George W. Bush because he was poorly advised by his intelligence and defense aides.

AFP_ENG_20030417.0307

Chapter I: The Bertolini

“The Signora had no business to do it,” said Miss Bartlett, “no business at all. She promised us south rooms with a view close together, instead of which here are north rooms, looking into a courtyard, and a long way apart. Oh, Lucy!”

“And a Cockney, besides!” said Lucy, who had been further saddened by the Signora’s unexpected accent. “It might be London.”

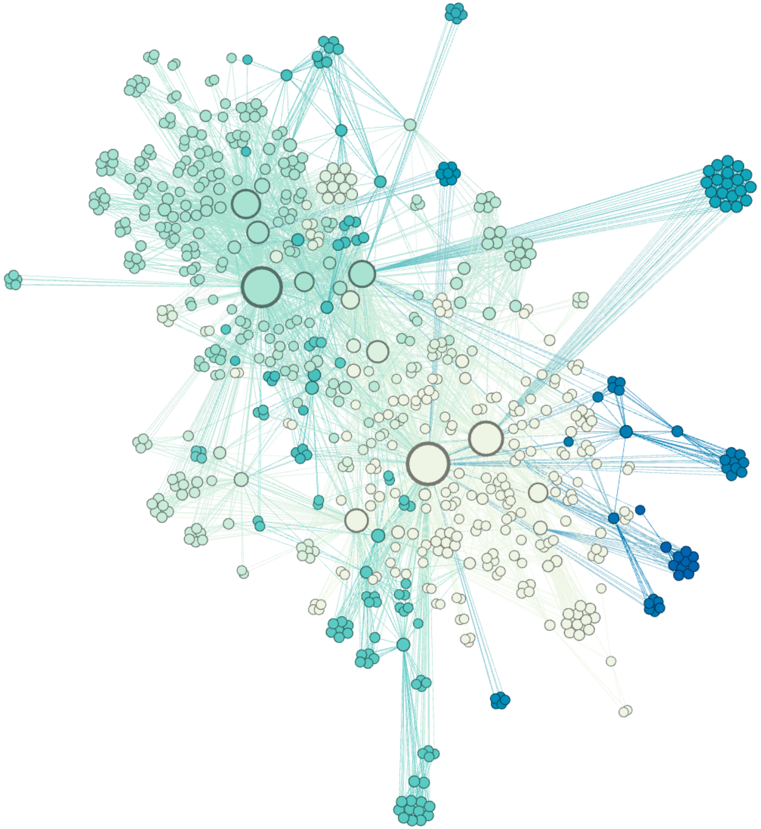
Forster, *A Room with a View*

Analysis

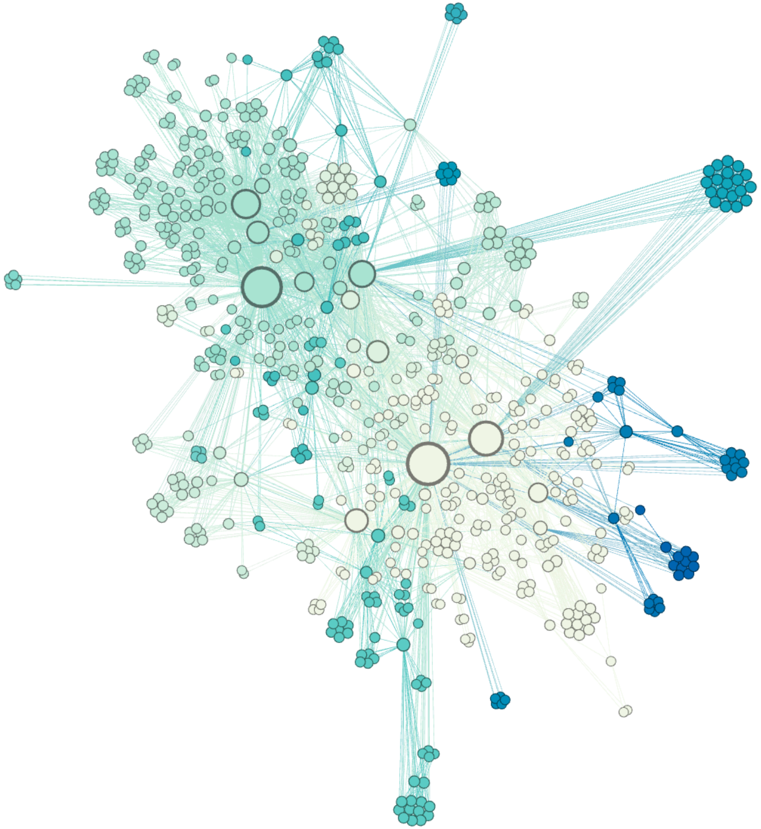
- How well does each model identify entities who are men and women?
- We annotate the gender for all `PER` entities in the literary test data and measure the recall of each model with respect to those entities.

Training	Women	Men	Diff
ACE	38.0	49.6	-11.6
Literary	69.3	68.2	1.1

How do we design NLP to drive insight into literary texts?

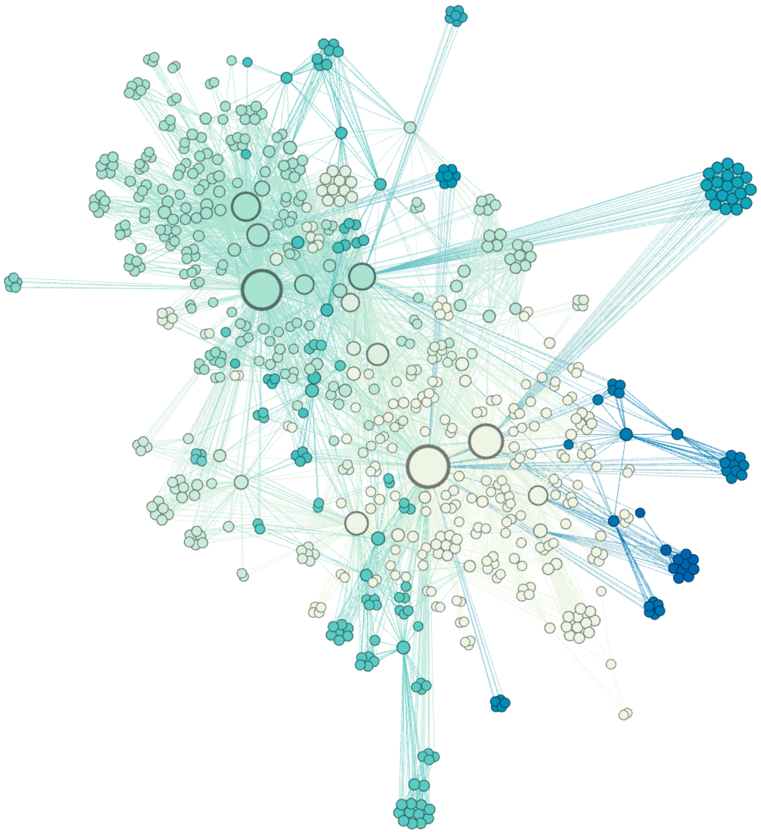


How does information propagate through implicit social networks?



- Information diffusion in blogs (Gruhl et al., 2004; Leskovec et al., 2007)
- Spread of rumor and misinformation (Kwon et al., 2013; Friggeri et al., 2014; Del Vicario et al., 2016; Vosoughi et al., 2018)
- Textual reuse across different legislative bills (Wilkerson et al., 2015)

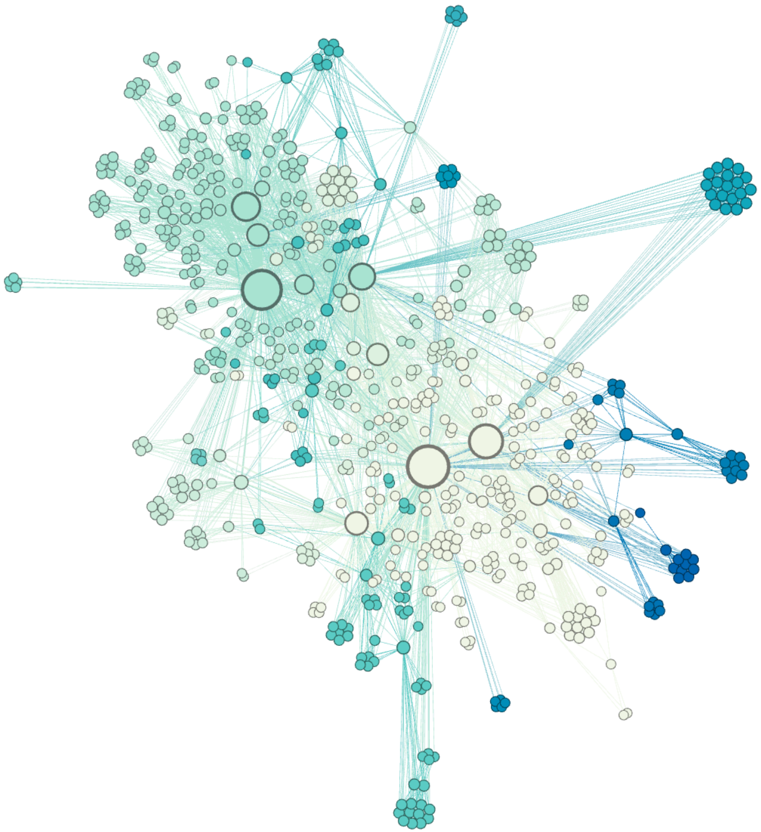
Great Expectations



● = Character

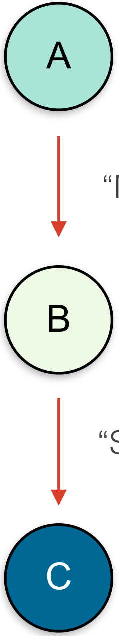
╱ = Conversational Interaction

Great Expectations



Information = Quoted Speech e.g., “Miss Havisham is dead”

Propagation = Repetition of information across a character triad



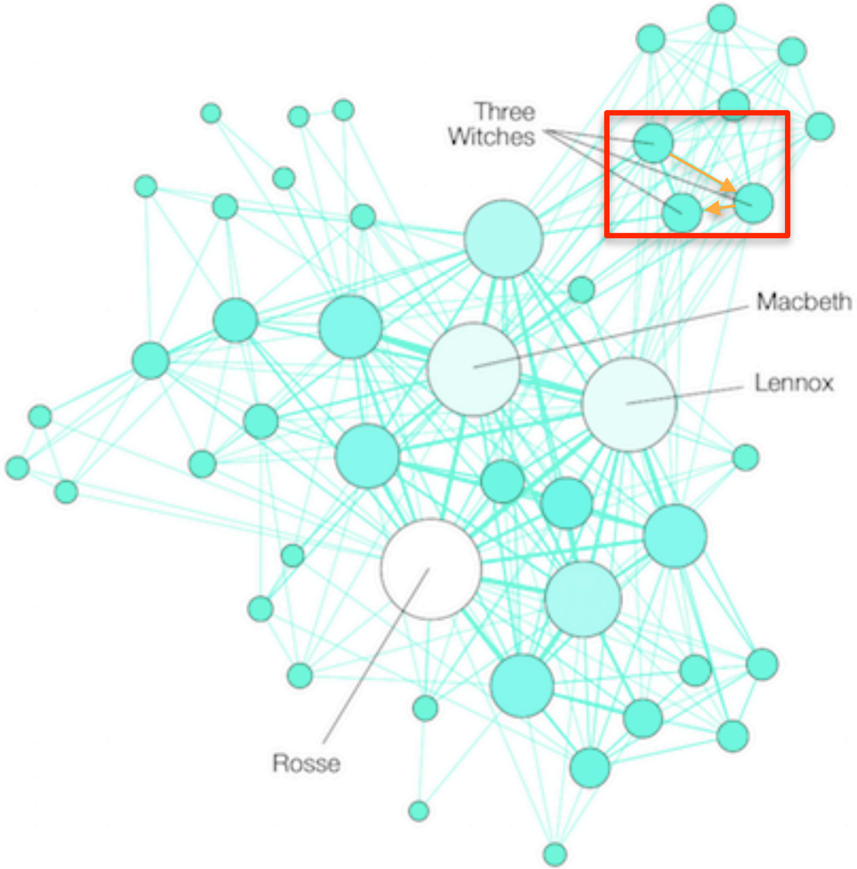
Research question

What are the structural properties of information-propagating nodes in fiction?

- "Gossip" among close friends, family, etc.: nodes that circulate information among densely-connected strong ties.

- Information bridges: nodes that pass information between otherwise disconnected communities.

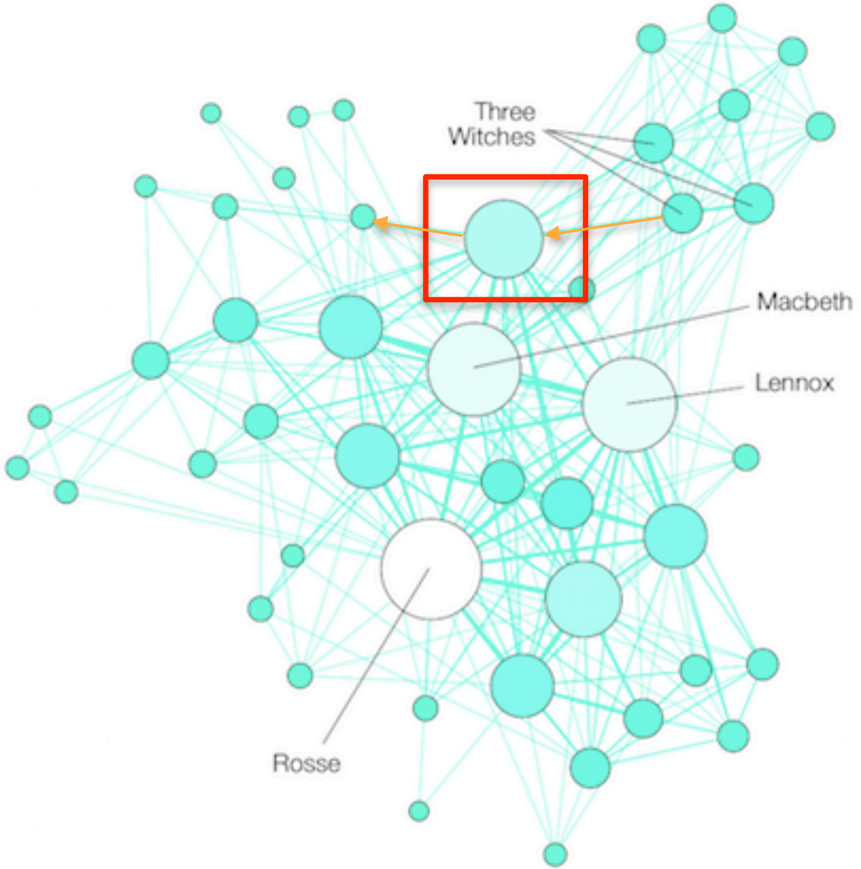
- “Gossip” among close friends, family, etc.: nodes that circulate information among densely-connected strong ties.



Macbeth

Source: http://www.martingrandjean.ch/network-visualization-shakespeare/

- Information bridges: nodes that pass information between otherwise disconnected communities.

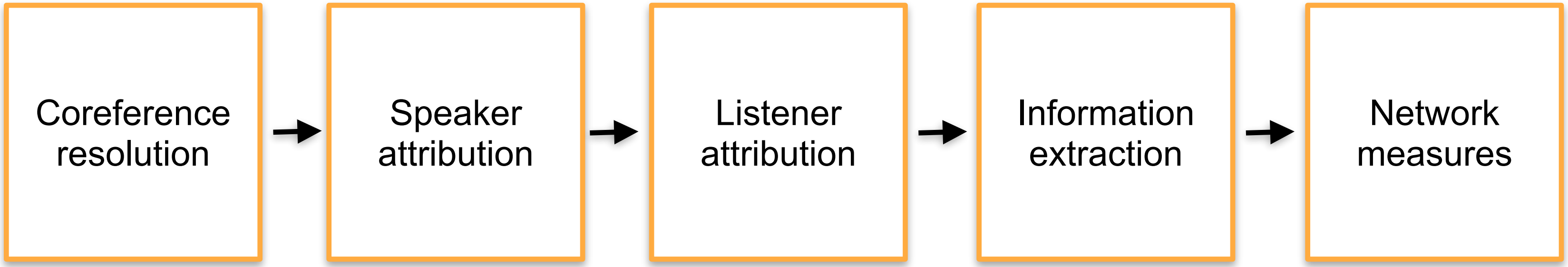


Source: http://www.martingrandjean.ch/network-visualization-shakespeare/

Macbeth

NLP Pipeline

What are the structural properties of information-propagating nodes in fiction?



Coreference resolution

Speaker attribution

Listener attribution

Information extraction

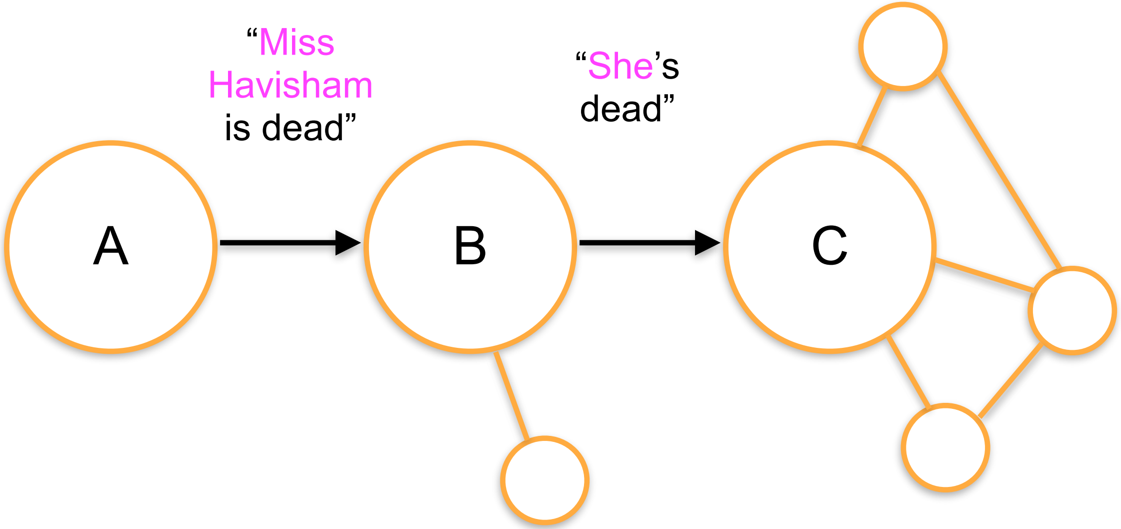
Network measures

Coreference resolution

- Identify unique characters from mentions of names, pronouns to construct network

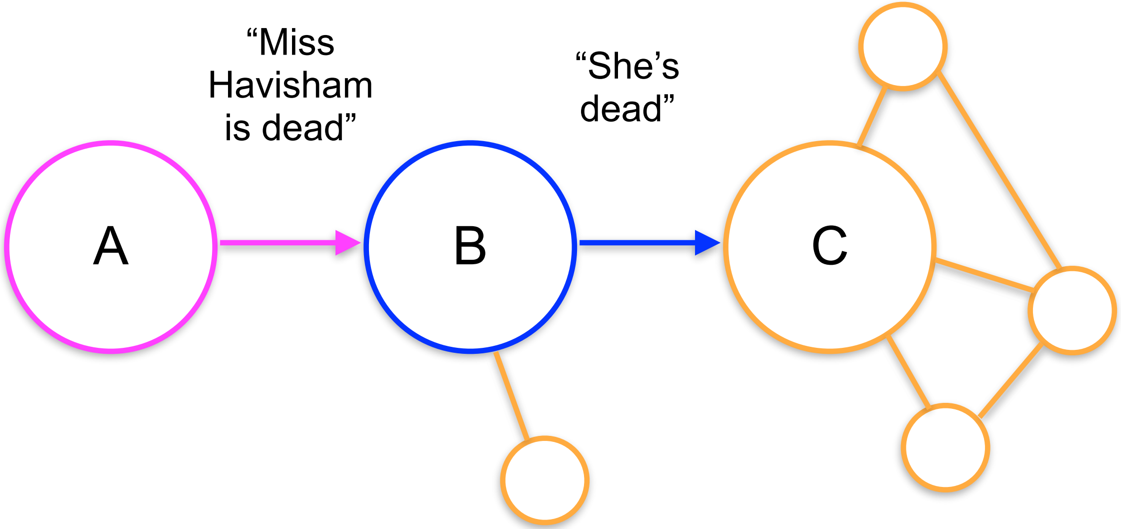
- Identify when two mentions in quotations refer to the same individual

- $B^3 = 68.1$



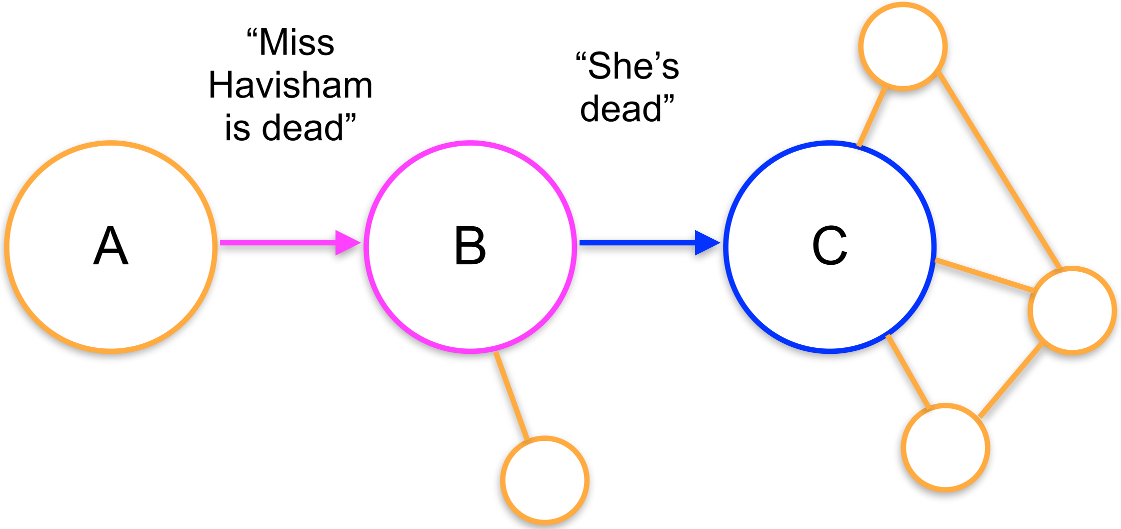
Speaker attribution

- Link quotations to their speakers.
- $B^3 = 71.3$



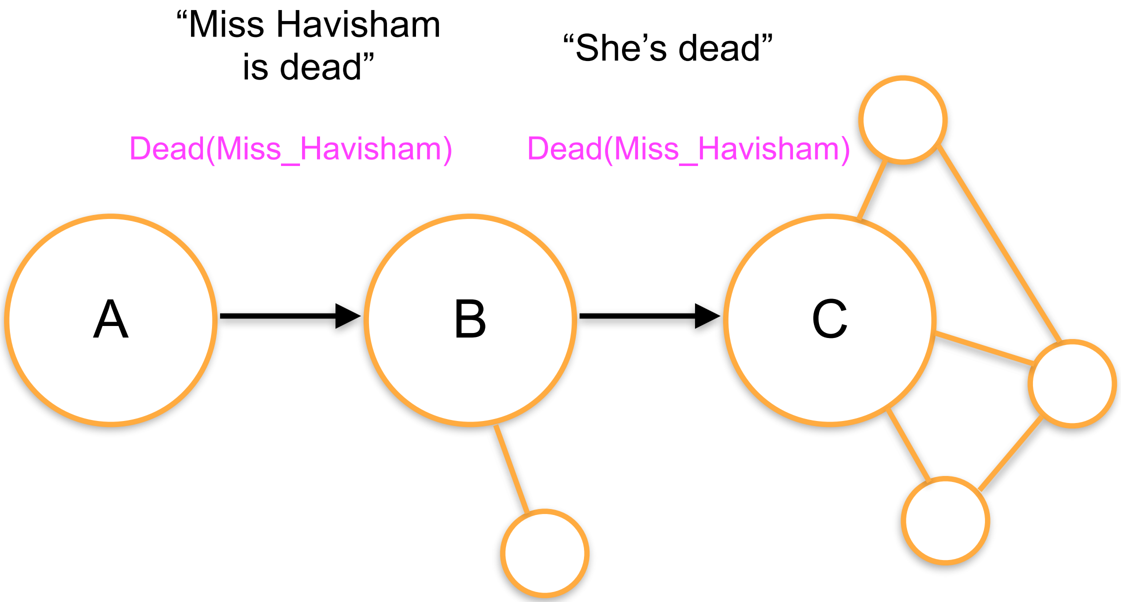
Listener attribution

- Identify which characters were present when a given quotation was spoken.
- Identify blocks of dialogue; listeners = all characters mentioned in narrative



Information extraction

- Extract an atomic unit of information in order to track its propagation.
- Resolve coreference and select propositional tuples of the form: [subject, verb, object]



Research question

What are the structural properties of information-propagating nodes in fiction?

- “Gossip” among close friends, family, etc.: nodes that circulate information among densely-connected strong ties.

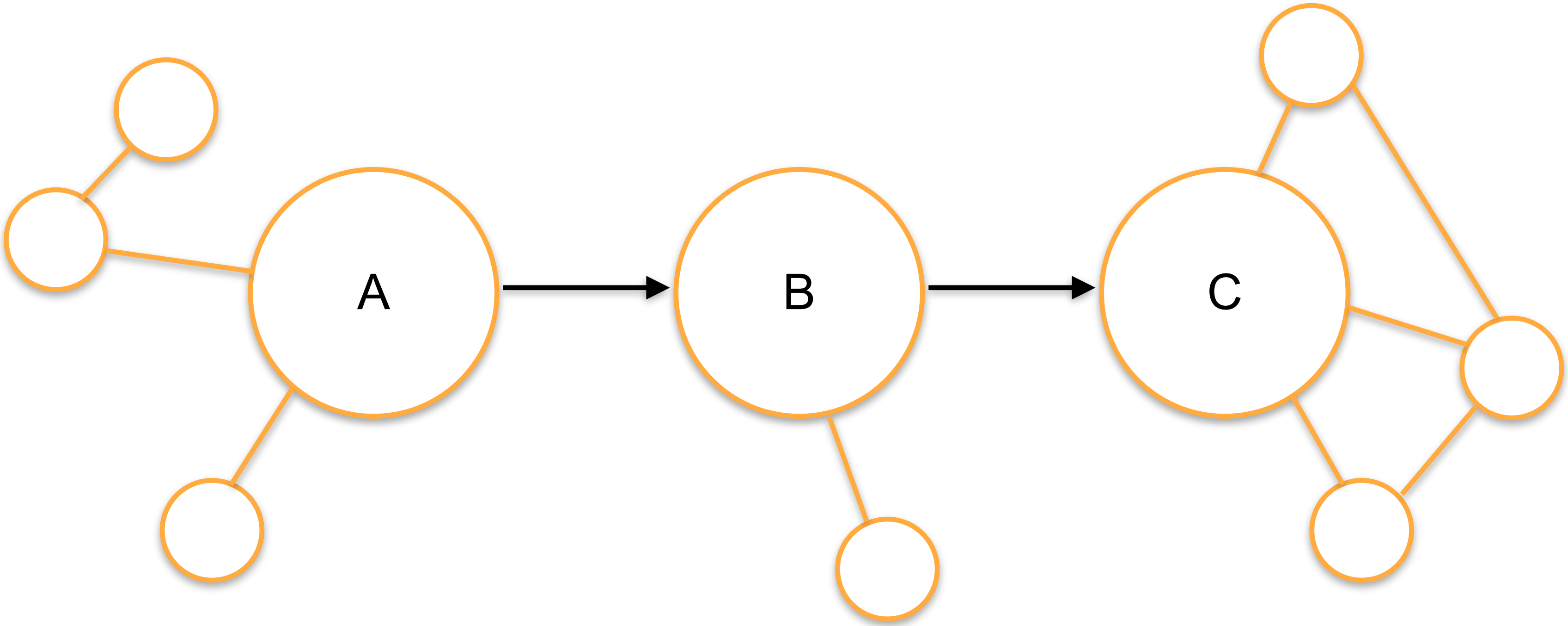
- Information bridges: nodes that pass information between otherwise disconnected communities.

Data

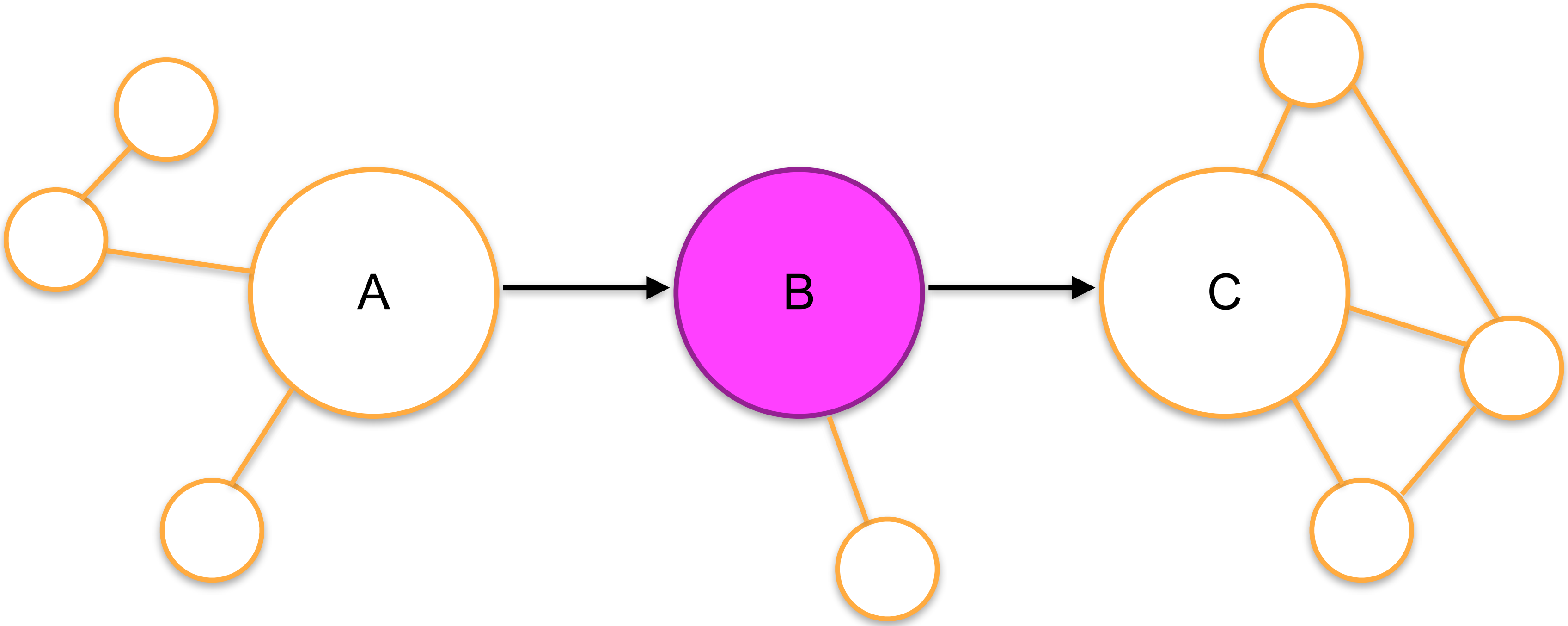
- 15K books in Project Gutenberg (in the public domain in the US).
- Examine four high-precision topics:
- Each topic has > 100 repeated tuple instances
- 4k of 15k books contain one of these repeated tuples

amorous	love, marry
hostile	hurt, hit, shoot, kill
juridical	arrest, escape, innocent, guilty
vital	alive, sick, dead

Propagation

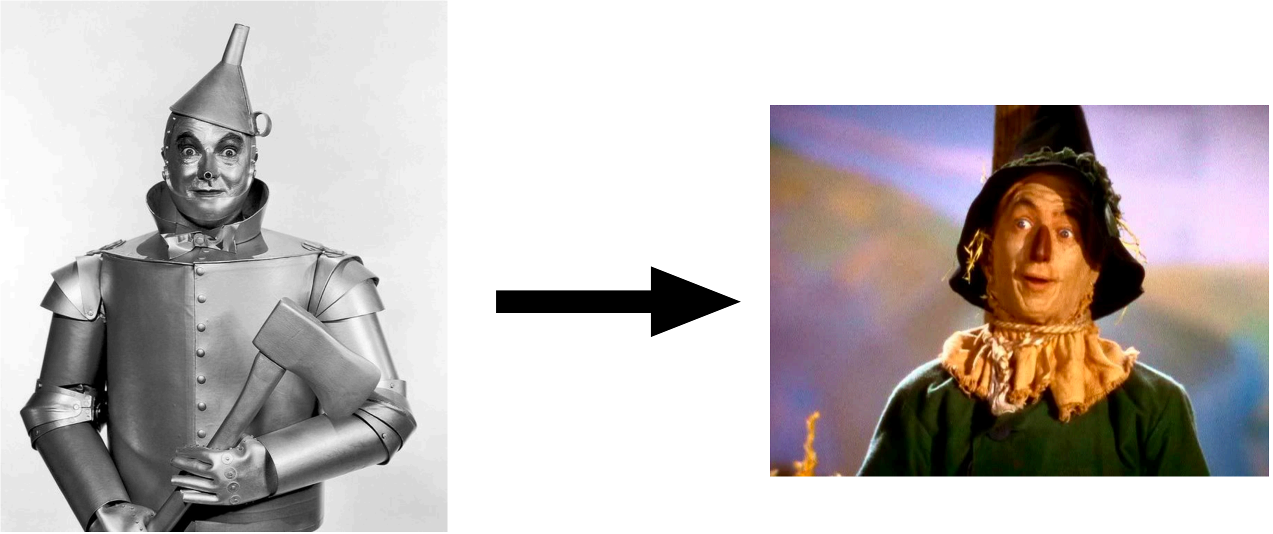


Propagation



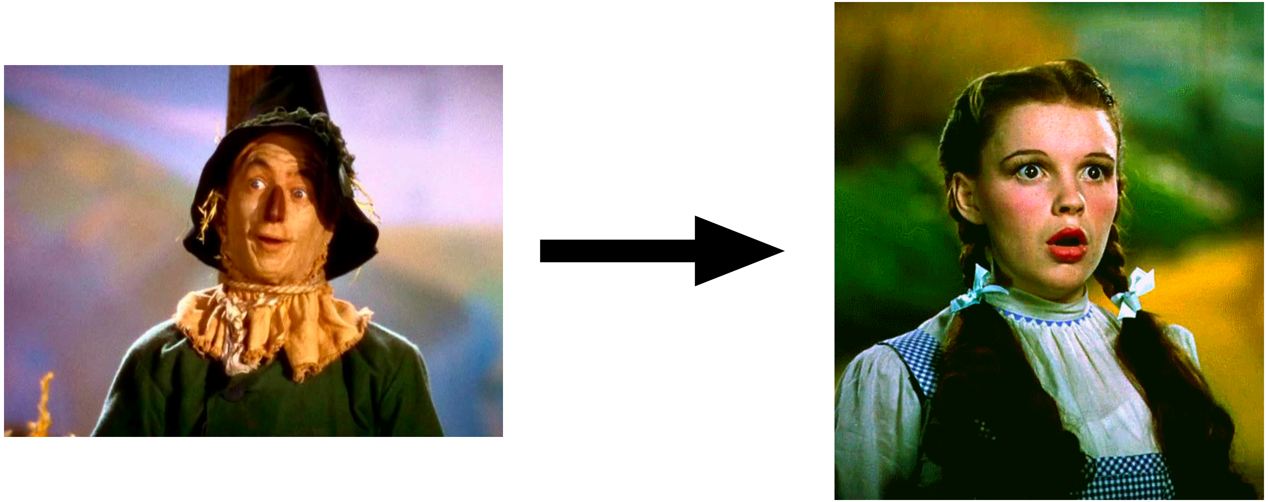
What distinguishes successful propagation from unsuccessful propagation?

Propagation



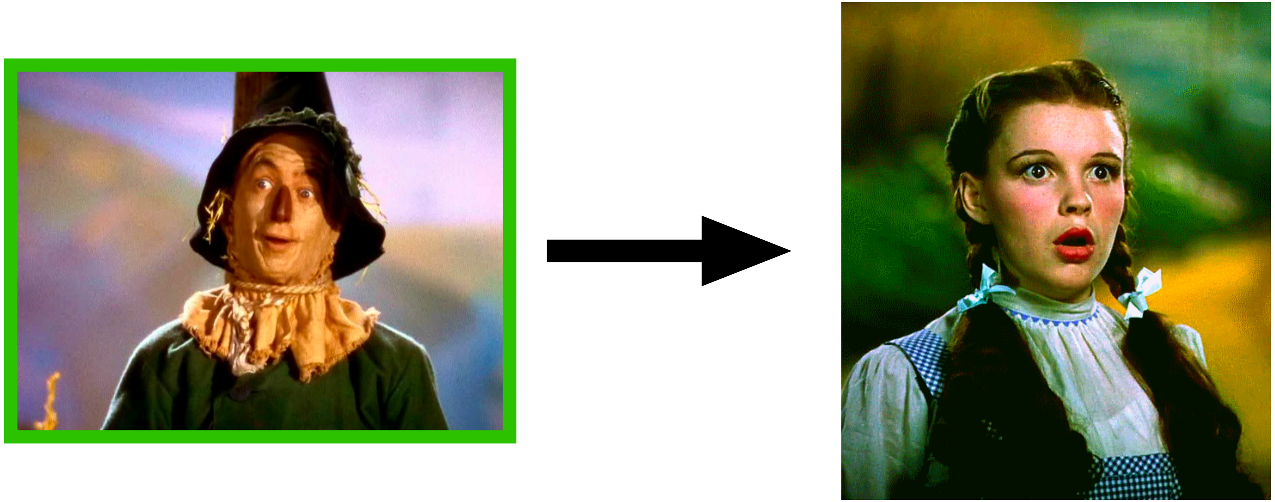
“The wicked witch is dead”

Propagation



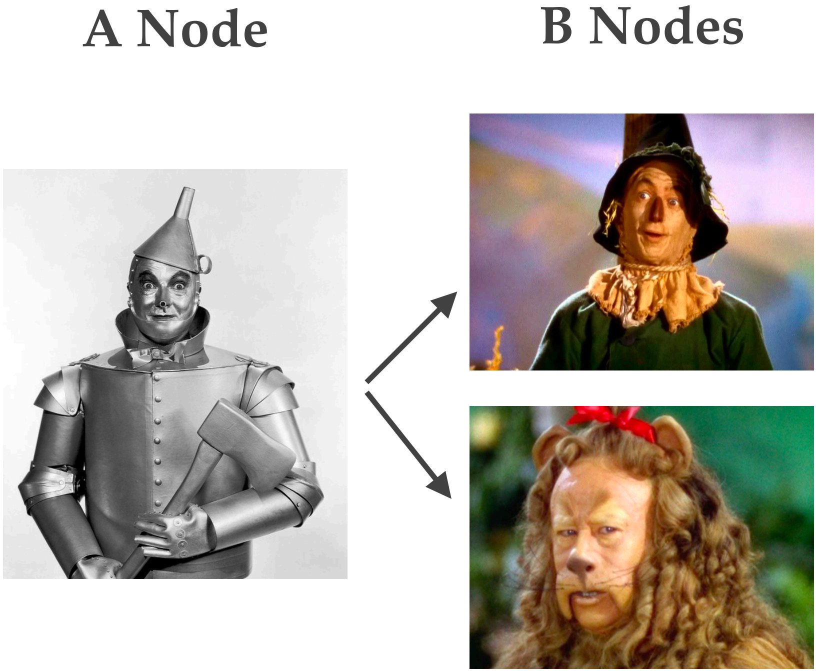
“You’ll never guess what happened to the wicked witch… she’s dead”

Propagation



“You’ll never guess what happened to the wicked witch… she’s dead”

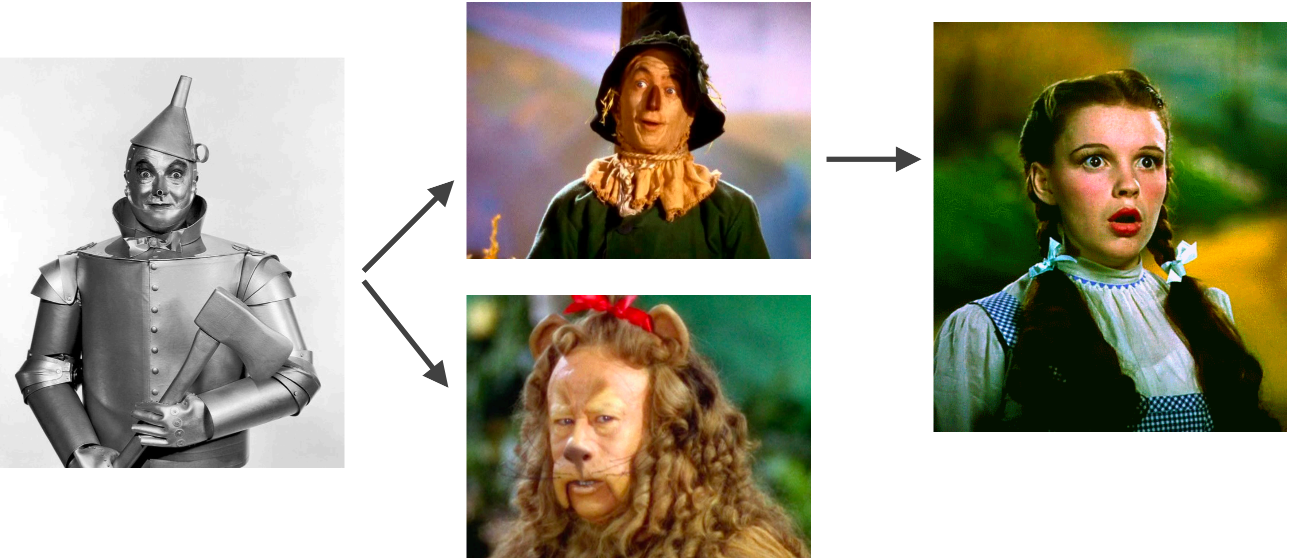
Propagation



A Node

B Nodes

Propagation

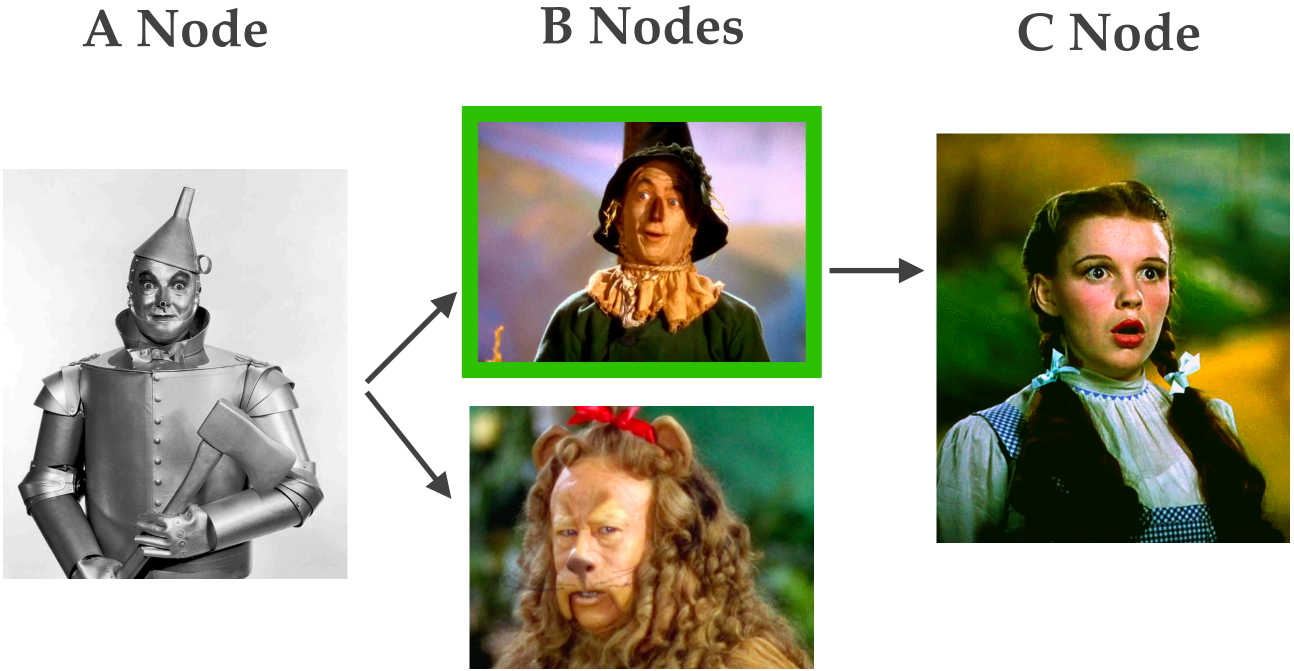


A Node

B Nodes

C Node

Propagation

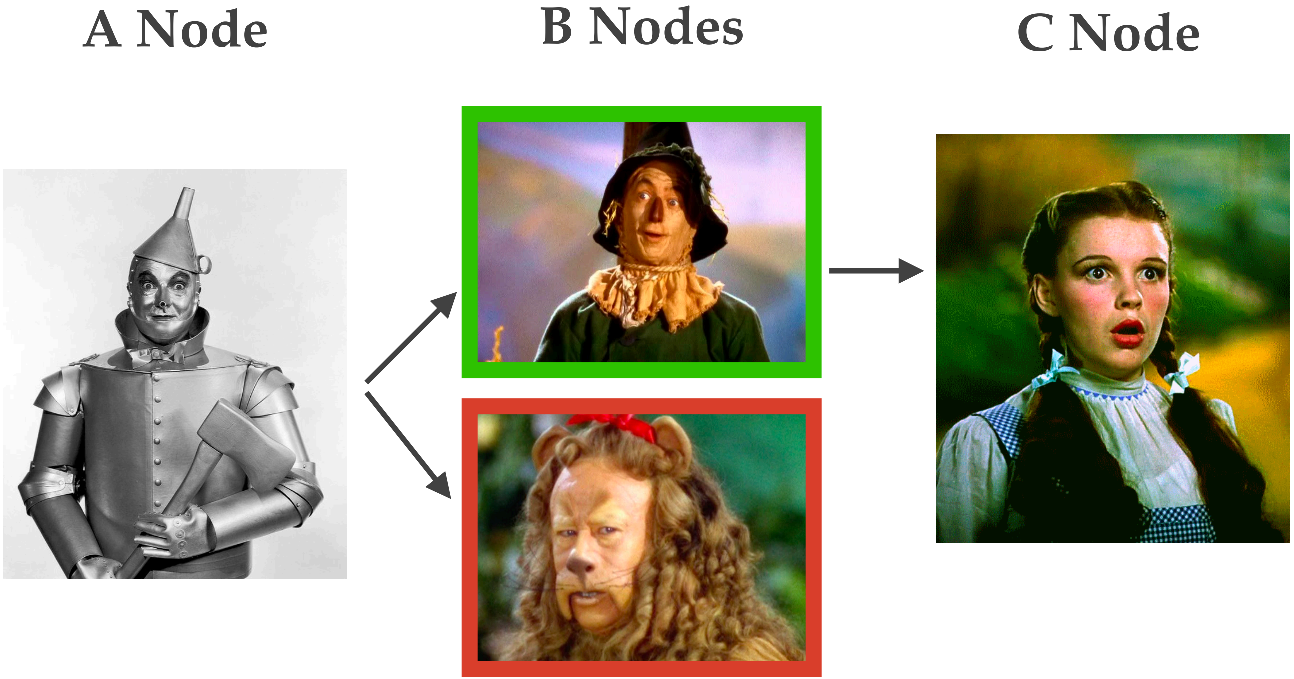


A Node

B Nodes

C Node

Propagation



A Node

B Nodes

C Node

Network Measures

B Nodes

- Betweenness Centrality
- Effective Size
- Efficiency

Information bridges

- Closeness Centrality
- Average Neighbor Degree
- Number of Triangles

Modeling

B Nodes

x 1730

x 1730

- Use logistic regression to identify which node measures are meaningful
- Contextualize results to determine the network dynamics that are most integral to information propagation in literary fiction

Feature Coefficients

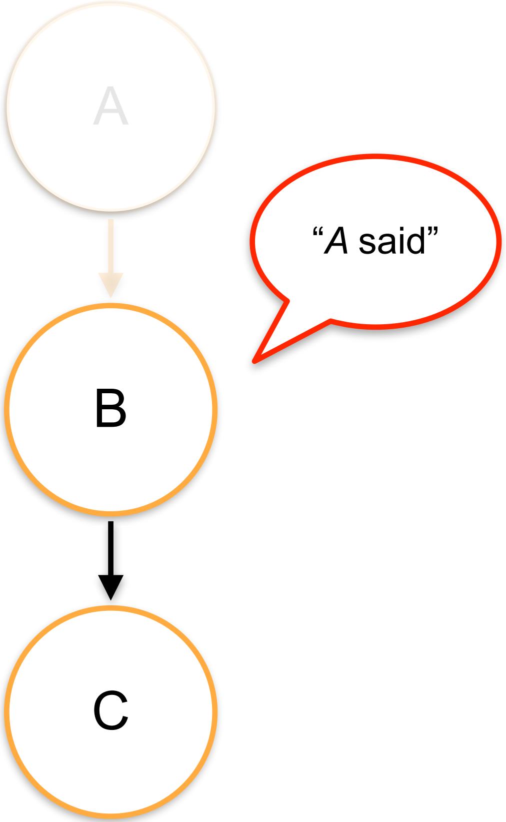
Graph Measure	Model Coefficient
Efficiency	3.0*
Effective size	2.7
Betweenness Centrality	0.5
Closeness centrality	0.1
Triangles	-0.4
Average Neighbor Degree	-4.9*

Feature Coefficients

Graph Measure	Model Coefficient
Efficiency	3.0*
Effective size	2.7
Betweenness Centrality	0.5
Closeness centrality	0.1
Triangles	-0.4
Average Neighbor Degree	-4.9*

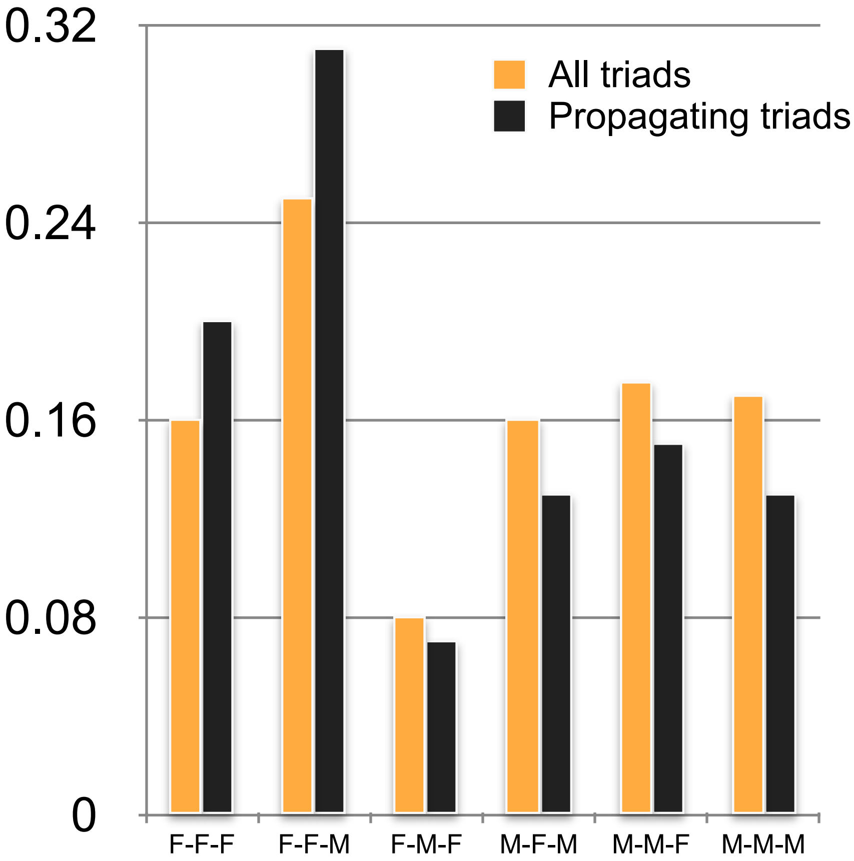
Explicit propagation

- Test explicit propagation: when another character states: "… Jane *said* …"
- 93,948 instances of propagation; 258,619 triads
- (Couldn't use this for testing successful propagation since we can't identity co-present B nodes during A's initial statement but we can analyze its properties for other questions.)

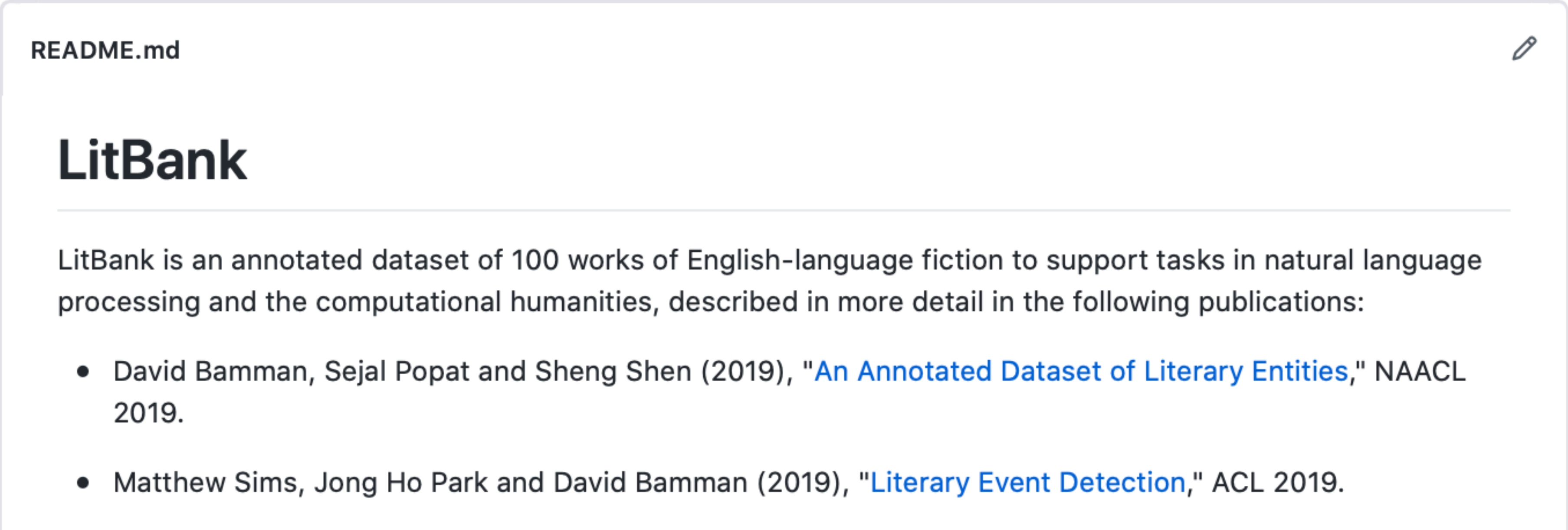


Explicit propagation

- What is the role of gender in in the *depiction* of propagation?
- Women are often stereotyped as more likely to engage in gossip within literature (Spacks 1985); often cast as intermediaries between men (Selisker 2015).
- Calculate the relative proportions of different gender configurations for propagating triads compared to all triads



https://github.com/dbamman/litbank



README.md

LitBank

LitBank is an annotated dataset of 100 works of English-language fiction to support tasks in natural language processing and the computational humanities, described in more detail in the following publications:

- David Bamman, Sejal Popat and Sheng Shen (2019), "An Annotated Dataset of Literary Entities," NAACL 2019.
- Matthew Sims, Jong Ho Park and David Bamman (2019), "Literary Event Detection," ACL 2019.

Thanks!

David Bamman
dbamman@berkeley.edu

- David Bamman, Olivia Lewke and Anya Mansoor, "An Annotated Dataset of Coreference in English Literature" (LREC 2020)
- David Bamman, Sejal Popat and Sheng Shen, "An Annotated Dataset of Literary Entities" (NAACL 2019)
- Matthew Sims and David Bamman, "Information Propagation in Literary Social Networks" (EMNLP 2020)

```
https://github.com/dbamman/litbank
```