Structural Analysis of Word Collocation Networks

Shibamouli Lahiri
University of North Texas

April 29, 2014

Three Views of Language

Structuralist View	Speech View
Grammars	n-gram language models
POS tags	PCFG language models
Dependencies	Speech recognition
Co-reference resolution	Machine translation

Three Views of Language

Structuralist View	Speech View	Complex Networks View
Grammars	n-gram language models	Networks of words, sentences, documents, ...
POS tags	PCFG language models	Small world, scale-free networks
Dependencies	Speech recognition	Power-law degree distribution
Co-reference resolution	Machine translation	High clustering coefficient, low diameter

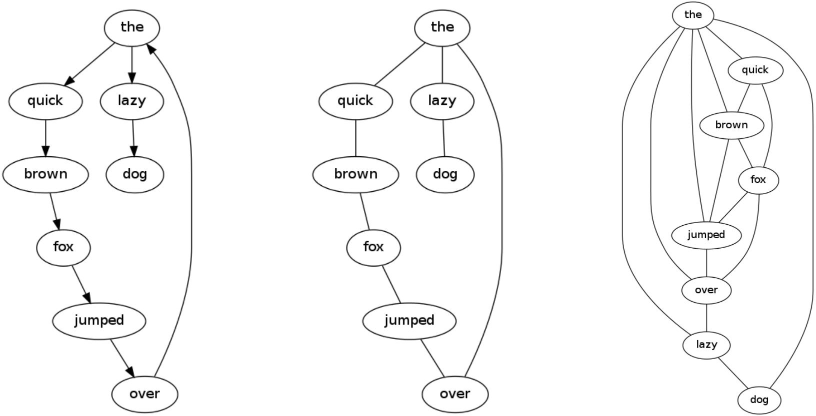
Where do we find them?

- Sentence networks
  - Summarization (LexRank, TextRank)
- Word networks
  - Keyword Extraction (TextRank, ExpandRank)
  - Authorship Attribution (Antiqueira et al., 2006)
  - Genre Identification (Stevanak et al., 2010)
  - Opinion Classification (Amancio et al., 2011)
  - Semantic Analysis (Biemann et al., 2012)

Outline

- Examples of word collocation networks
- Structural properties
- Three exploratory analyses

The quick brown fox jumped over the lazy dog.



Structural Properties

Network Property	Notation
Number of vertices	$\lvert V\rvert$
Number of edges	$\lvert E\rvert$
Shrinkage exponent (Leskovec et al., 2007)	$\log_{\lvert V\rvert} \lvert E\rvert$
Global clustering coefficient	C
Small-worldliness (Walsh, 1999; Matsuo et al., 2001)	$\mu = (\bar{C}/L)/(\bar{C}_{rand}/L_{rand})$
Diameter (directed)	$d_{directed}$
Diameter (undirected)	$d_{undirected}$
Power-law exponent of degree distribution	α
Power-law exponent of in-degree distribution	α_{in}
Power-law exponent of out-degree distribution	α_{out}
p-value for α	p_{α}
p-value for α_{in}	$p_{\alpha_{in}}$
p-value for α_{out}	$p_{\alpha_{out}}$
Number of connected components	#CC
Number of strongly connected components	#SCC
Size of the largest connected component	-
Size of the largest strongly connected component	-

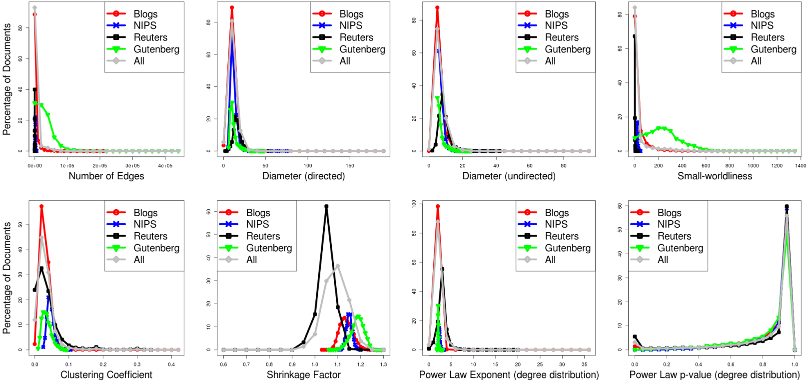
Exploratory Analysis

1. How do network properties vary across different **genres of text**?
2. How do network properties vary across different **network types**?
3. How do the properties **evolve** as a word network grows in size?

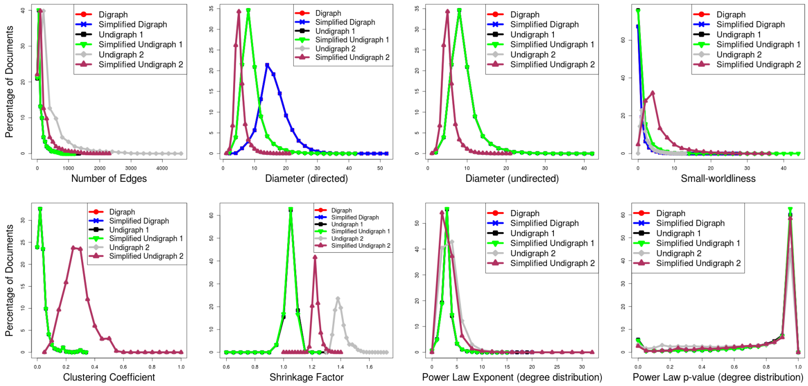
Datasets

- **Blog Authorship Corpus** (Schler et al., 2006)
  - 19,320 blogs, 136.8 million words
- **Reuters-21578, Distribution 1.0**
  - 19,043 news articles, 2.6 million words
- **NIPS Conference Papers Vols 0-12**
  - 1,740 papers, 4.8 million words
- **E-books from Project Gutenberg**
  - 3,036 books, 210.9 million words

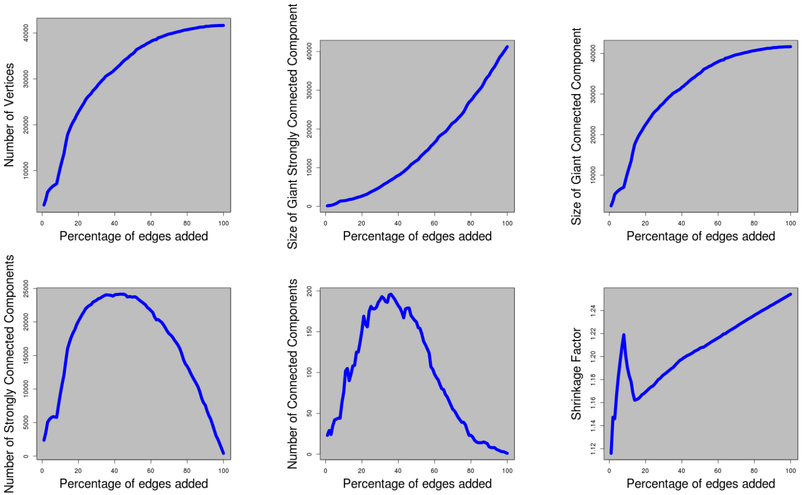
Exploratory Analysis 1: Across Genres



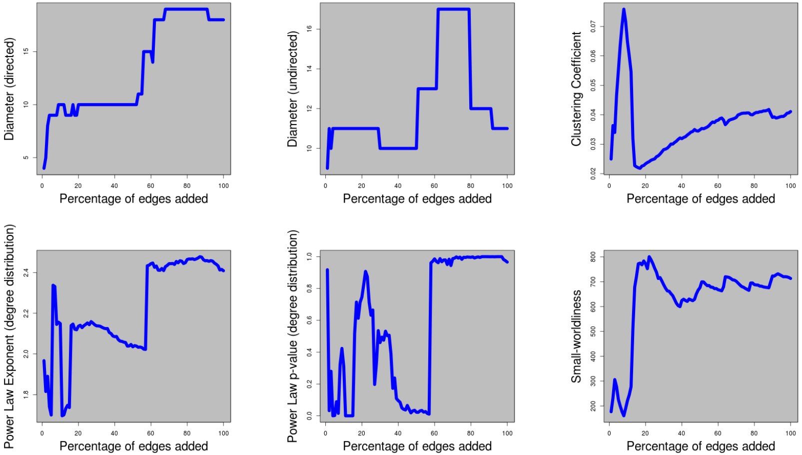
Exploratory Analysis 2: Across Network Types



Exploratory Analysis 3: Evolution of Properties



Exploratory Analysis 3: Evolution of Properties



Observations

- There are statistically significant variations between **genres** for distributions of network properties.
- There are also statistically significant variations between **network types** for distributions of network properties.
- As word networks grow in size, key structural properties **evolve** in *phases*, via spikes and drops.
- Networks **densify** as they grow, evidenced by the shrinkage in diameter.
- Networks from news articles have low small-worldliness (μ) and high α.

Code and Data

- https://drive.google.com/file/d/0B2Mzhc7popBgODFKZVVnQTFMQkE/edit?usp=sharing

Looking Ahead . . .

- **Visualization of collocation networks**
  - IBM Many Eyes project (http://www.manyeyes.com/)
- **Application of collocation networks**
  - Authorship Attribution (Lahiri and Mihalcea, 2013)
  - Keyword Extraction (Lahiri et al., 2014)
  - Gender and Genre Classification
  - Stylometry

Thank You!

Questions?