Complexity of Word Collocation Networks: A Preliminary Structural Analysis

Shibamouli Lahiri
Computer Science and Engineering
University of North Texas
Denton, TX 76207, USA
shibamoulilahiri@my.unt.edu

Abstract

In this paper, we explore complex network properties of word collocation networks (Ferret, 2002) from four different genres. Each document of a particular genre was converted into a network of words with word collocations as edges. We analyzed graphically and statistically how the *global properties* of these networks varied across different genres, and among different network types within the same genre. Our results indicate that the distributions of network properties are visually similar but statistically apart across different genres, and interesting variations emerge when we consider different network types within a single genre. We further investigate how the global properties change as we add more and more collocation edges to the graph of one particular genre, and observe that except for the number of vertices and the size of the largest connected component, network properties change in *phases*, via jumps and drops.

1 Introduction

Word collocation networks (Ferret, 2002; Ke, 2007), also known as collocation graphs (Heyer et al., 2001; Choudhury and Mukherjee, 2009), are networks of words found in a document or a document collection, where each node corresponds to a unique *word type*, and edges correspond to *word collocations* (Ke and Yao, 2008). In the simplest case, each edge corresponds to a unique bigram in the original document. For example, if the words w_A and w_B appeared together in a document as a bigram w_Aw_B, then the word collocation network of that particular document will contain an edge $w_A \rightarrow w_B$. Note that edges can be directed ($w_A \rightarrow w_B$) or undirected ($w_A - w_B$). Furthermore, they can be weighted (with the frequency of the bigram w_Aw_B) or unweighted.

It is interesting to note that word collocation networks display complex network structure, including power-law degree distribution and small-world behavior (Matsuo et al., 2001a; Matsuo et al., 2001b; Masucci and Rodgers, 2006; Liang et al., 2012). This is not surprising, given that natural language generally shows complex network properties at different levels (Ferrer i Cancho and Solé, 2001; Motter et al., 2003; Biemann et al., 2009; Liang et al., 2009). Moreover, researchers have used such complex networks in applications ranging from text genre identification (Stevanak et al., 2010) and Web query analysis (Saha Roy et al., 2011) to semantic analysis (Biemann et al., 2012) and opinion mining (Amancio et al., 2011). In Section 2, we will discuss some of these applications in more detail.

The goal of this paper is to explore some key structural properties of these complex networks (cf. Table 1), and study how they vary across different genres of text, and also across different network types within the same genre. We chose *global network properties* like diameter, global clustering coefficient, shrinkage exponent (Leskovec et al., 2007), and small-worldliness (Walsh, 1999; Matsuo et al., 2001a), and experimented with four different text collections – blogs, news articles, academic papers, and digitized books (Section 4.1). Six different types of word collocation networks were constructed on each document, as well as on the entire collections – two with directed edges, and four with undirected edges (Section 3). We did not take into account edge weights in our study, and kept it as a part of our future work (Section 5).

Tracking the variation of complex network properties on word collocation networks yielded several important observations and insights. We

noted in particular that different genres had considerable visual overlap in the distributions of global network properties like diameter and clustering coefficient (cf. Figure 2), although statistical significance tests indicated the distributions were sufficiently apart from each other (Section 4.2). This calls for a deeper analysis of complex network properties and their general applicability to tasks like genre identification (Stevanak et al., 2010).

We further analyzed distributions of global word network properties across six different network types *within the same genre* (Section 4.2). This time, however, we noted a significant amount of separation – both visually as well as statistically – among the distributions of different global properties (cf. Figure 3 and Table 5).

In our final set of experiments, we analyzed how global network properties change as we start with an empty network, and gradually add edges to that network. For this experiment, we chose the news genre, and tracked the variation of 17 different global network properties on four types of networks. We observed that all global network properties (except the number of vertices and edges, number of connected components and the size of the largest connected component) show unpredictability and *spikes* when the percentage of added edges is small. We also noted that most global properties showed at least one *phase transition* as the word collocation networks grew larger. Statistical significance tests indicated that the patterns of most global property variations were non-random and positively correlated (Section 4.3).

2 Related Work

That language shows complex network structure at the word level, was shown more than a decade ago by at least two independent groups of researchers (Ferrer i Cancho and Solé, 2001; Matsuo et al., 2001a). Matsuo et al. (2001b) went further ahead, and designed an unsupervised keyword extraction algorithm using the small-world property of word collocation networks. Motter et al. (2003) extended the collocation network idea to *concepts* rather than words, and observed a small-world structure in the resulting network. Edges between concepts were defined as entries in an English thesaurus. Liang et al. (2009) compared word collocation networks of Chinese and English text, and pointed out their similarities and differences. They further constructed *character collocation networks* in Chinese, showed their small-world structure, and used these networks in a follow-up study to accurately segregate Chinese essays from different literary periods (Liang et al., 2012).

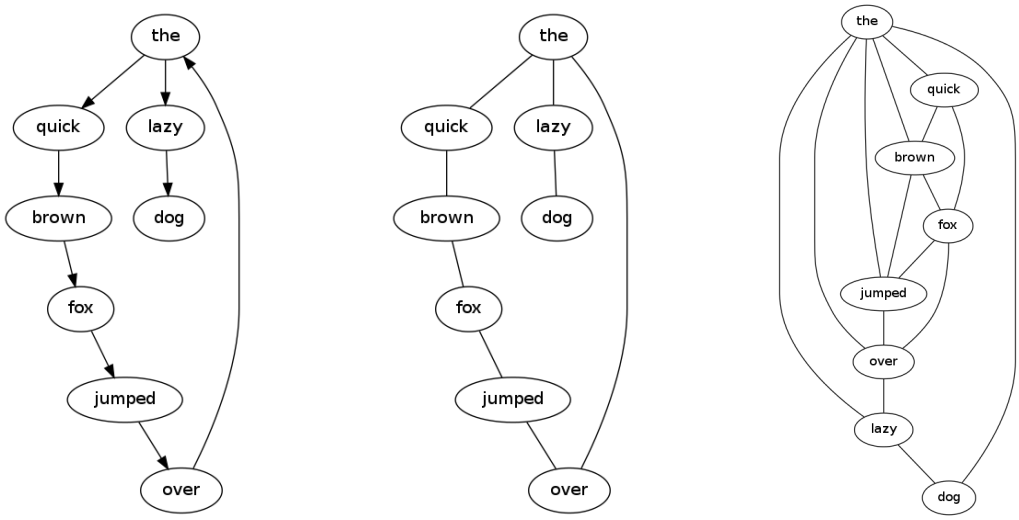
Word collocation networks have also been successfully applied to the authorship attribution task.[1] Antiqueira et al. (2006) were among the first to apply complex network features like clustering coefficient, *component dynamics deviation* and *degree correlation* to the authorship attribution problem.

Biemann et al. (2009) constructed syntactic and semantic distributional similarity networks (DSNs), and analyzed their structural differences using spectral plots. Biemann et al. (2012) further used *graph motifs* on collocation networks to distinguish real natural language text from generated natural language text, and to point out the shortcomings of n-gram language models.

Word collocation networks have been used by Amancio et al. (2011) for opinion mining, and by Mihalcea and Tarau (2004) for keyword extraction. While the former study used complex network properties as features for machine learning algorithms, the latter ran PageRank (Page et al., 1998) on word collocation networks to sieve out most important words.

While all the above studies are very important, we found none that performed a thorough and systematic exploration of different global network properties on different network types across genres, along with statistical significance tests to assess the validity of their observations. Stevanak et al. (2010), for example, used word collocation networks to distinguish between novels and news articles, but they did not perform a distributional analysis of the different global network properties they used, thereby leaving open how good those properties truly were as features for genre classification, and whether there exist a better and simpler set of global network properties for the same task. On the other hand, Masucci and Rodgers (2006), Ke (2007), and Ke and Yao (2008) explored several global network properties on word collocation networks, but they did not address the problem of analyzing within-genre and cross-genre variations of those properties.

[1] For details on authorship attribution, please see the surveys by Juola (2006), Koppel et al. (2009), and Stamatatos (2009).



(a) Directed (b) Undirected, Variant 1 (c) Undirected, Variant 2

Figure 1: Word collocation networks of the sentence *"the quick brown fox jumped over the lazy dog"*. Note that for all three network types, the word "the" appeared as the most central word. It is in general the case that stop words like "the" are the most central words in collocation networks, especially since they act as *connectors* between other words.

Network Property	**Mathematical Expression**
Number of vertices	$\|V\|$
Number of edges	$\|E\|$
Shrinkage exponent (Leskovec et al., 2007)	$\log_{\|V\|} \|E\|$
Global clustering coefficient	C
Small-worldliness (Walsh, 1999; Matsuo et al., 2001a)	$\mu = (\bar{C}/L)/(\bar{C}_{rand}/L_{rand})$
Diameter (directed)	d
Diameter (undirected)	d
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 the power-law exponent of degree distribution	N/A
p-value for the power-law exponent of in-degree distribution	N/A
p-value for the power-law exponent of out-degree distribution	N/A
Number of connected components*	N/A
Size of the largest connected component*	N/A
Number of strongly connected components*	N/A
Size of the largest strongly connected component*	N/A

Table 1: Different global network properties used in our study. The ones marked with an asterisk ("*") are only used in Section 4.3 in the context of incrementally constructing networks by gradually adding edges. For document networks, these four properties do not make sense, because the number of connected components is always one, and the size of the largest connected component always equals the number of vertices in the document network. Note also that in-degree distribution, out-degree distribution, and the directed version of diameter do not make sense for undirected networks, and same goes with the number of strongly connected components and the size of the largest strongly connected component. Here we report them separately for conceptual clarity.

In addition to addressing these problems, in this paper we introduce a new analysis - how the global network properties change as we gradually add more collocation edges to a network (Section 4.3).[2]

3 Collocation Networks of Words

Before constructing collocation networks, we lowercased the input text and removed all punctuation, but refrained from performing stemming in order to retain subtle distinctions between words like "vector" and "vectorization". Six different types of word collocation networks were constructed on each document (used in Section 4.2) as well as on document collections (used in Section 4.3), where nodes are unique words, and an edge appears between two nodes if their corresponding words appeared together as a bigram or in a trigram in the original text. All the network types have the *same* number of vertices (i.e., words) for a particular document or a document collection, and they are only distinguished from each other by the type (and potentially, number) of edges, as follows:

Directed – Directed edge $w_A \rightarrow w_B$ if $w_A w_B$ is a bigram in the given text.

Undirected, Variant 1 – Undirected edge $w_A - w_B$ if $w_A w_B$ is a bigram in the given text.

Undirected, Variant 2 – Undirected edges $w_A - w_B$, $w_B - w_C$ and $w_A - w_C$, if $w_A w_B w_C$ is a trigram in the given text.

Directed Simplified – Same as the directed version, with *self-loops* removed.[3]

Undirected Variant 1, Simplified – Same as the undirected variant 1, with self-loops removed.

Undirected Variant 2, Simplified – Same as the undirected variant 2, with self-loops removed.

We did not take into account edge weights in our study, and all our networks are therefore unweighted networks. Furthermore, since we removed all punctuation information *before* constructing collocation networks, sentence boundaries were implicitly ignored. In other words, the last word of a sentence *does* link to the first word of the next sentence in our collocation networks. An example of the first three types of networks (directed, undirected variant 1, and undirected variant 2) is shown in Figure 1. Here we considered a sentence *"the quick brown fox jumped over the lazy dog"* as our document. Note that all the collocation networks in Figure 1 contain at least one cycle, and the directed version contains a directed cycle. In a realistic document network, there can be many such cycles.

We constructed word collocation networks on *document collections* as well. In this case, the six network types remain as before, and the only difference comes from the fact that now the whole collection is considered a single *super-document*. Words in this super-document are connected according to bigram and trigram relationships. We respected document boundaries in this case, so the last word of a particular document *does not* link to the first word of the next document. The *collection networks* have only been used in Section 4.3 of this paper, to show how global network properties change as we add edges to the network.

With the networks now constructed, we went ahead and explored several of their global properties (cf. Table 1). Properties were measured on each type of network on each document, thereby giving us property distributions across different genres of documents for a particular network type (cf. Figure 2), as well as property distributions across different network types for a particular genre (cf. Figure 3). We used the *igraph* software package (Csardi and Nepusz, 2006) for computing global network properties.

Among the properties in Table 1, number of vertices ($|V|$) and number of edges ($|E|$) are self-explanatory. The *shrinkage exponent* ($\log_{|V|} |E|$) is motivated by the observations that the number of edges ($|E|$) follows a power-law relationship with the number of vertices ($|V|$), and that as a network evolves, both $|V|$ and $|E|$ continue to grow, but the diameter of the network either *shrinks* or plateaus out, thereby resulting in a *densified* network (Leskovec et al., 2007). We explored two versions of graph diameter (d) in our study - a directed version (considering directed edges), and an undirected version (ignoring edge directions).[4]

The *global clustering coefficient* (C) is a mea-

[2] All code, data, and supplementary material are available at `https://drive.google.com/file/d/0B2Mzhc7popBgODFKZVVnQTFMQkE/edit?usp=sharing`. The data includes – among other things – the corpora we used (cf. Section 4.1), and code to construct the networks and analyze their properties.

[3] Note that self-loops may appear in word collocation networks due to punctuation removal in the pre-processing step. An example of such a self-loop is: *"The airplane took off. Off we go to Alaska."* Here the word "off" will contain a self-loop.

[4] For undirected collocation networks, these two versions yield the same results, as expected.

sure of how interconnected a graph's nodes are among themselves. It is defined as the ratio between the number of closed triplets of vertices (i.e., the number of ordered triangles or *transitive triads*), and the number of connected vertex-triples (Wasserman and Faust, 1994). The *small-worldliness* or *proximity ratio* (μ) of a network measures to what extent the network exhibits small-world behavior. It is quantified as the amount of deviation of the network from an equally large random network, in terms of average local clustering coefficient ($\bar{C}$) and average shortest path length (L)[5]. The exact ratio is $\mu = (\bar{C}/L)/(\bar{C}_{rand}/L_{rand})$, where $\bar{C}$ and L are the average local clustering coefficient and the average shortest path length of the given network, and $\bar{C}_{rand}$ and L_{rand} are the average local clustering coefficient and the average shortest path length of an equally large random network (Walsh, 1999; Matsuo et al., 2001a).

Since collocation networks have been found to display scale-free (power-law) degree distribution in several previous studies (see, e.g., (Ferrer i Cancho and Solé, 2001; Masucci and Rodgers, 2006; Liang et al., 2009)), we computed power-law exponents of in-degree, out-degree, and degree distributions on each of our collocation networks.[6] We also computed the corresponding p-values, following a procedure outlined in (Clauset et al., 2009). These p-values help assess whether the distributions are power-law or not. If a p-value is < 0.05, then there is statistical evidence to believe that the corresponding distribution is *not* a power-law distribution.

Finally, we computed the number of connected components, size of the largest ("giant") connected component, number of strongly connected components, and size of the largest strongly connected component, to be used in Section 4.3.

4 Analysis of Network Properties

4.1 Datasets

We used four document collections from four different genres – blogs, news articles, academic papers, and digitized books. For blogs, we used the **Blog Authorship Corpus** created by (Schler et al., 2006). It consists of 19,320 blogs from authors of different age groups and professions. The unprocessed corpus has about 136.8 million word tokens.

Our news articles come from the **Reuters-21578, Distribution 1.0** collection.[7] This collection contains 19,043 news stories, and about 2.6 million word tokens (unprocessed).

For the academic paper dataset, we used **NIPS Conference Papers Vols 0-12**.[8] This corpus comprises 1,740 papers and about 4.8 million unprocessed word tokens.

Finally, we created our own corpus of 3,036 digitized books written by 142 authors from the **Project Gutenberg** digital library.[9] After removing metadata, license information, and transcribers' notes, this dataset contains about 210.9 million word tokens.

That the word collocation networks of individual documents are indeed scale-free and small-world, is evident from Tables 2, 3, and 4, and Figure 2h. Irrespective of network type, a majority of the median α (power-law exponent of degree distribution) values hovers in the range $[2, 3]$, with low dispersion. This corroborates with earlier studies (Ferrer i Cancho and Solé, 2001; Liang et al., 2009; Liang et al., 2012). Similarly, the median μ (small-worldliness) is high for all genres except *news* (irrespective of network type), thereby indicating the document networks are indeed small-world. This finding is in line with previous studies (Matsuo et al., 2001a; Matsuo et al., 2001b). Moreover, Figure 2h shows that a majority of documents in different genres have a very high p-value, indicating that the networks are significantly power-law. The *news* genre poses an interesting case. Since many news stories in the Reuters-21578 collection are small, their collocation networks are not very well-connected, thereby resulting in very low small-worldliness values, as well as higher estimates of the power-law exponent α (cf. Tables 2, 3, and 4).

4.2 Distribution of Global Network Properties

We plotted the histograms of eight important global network properties on directed collocation networks in Figure 2. All histograms were plot-

[5] Also called *"characteristic path length"* (Watts and Strogatz, 1998).

[6] For undirected graphs, the exponents on all three distributions are the same.

[7] Available from `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

[8] Available from `http://www.cs.nyu.edu/~roweis/data.html`.

[9] `http://www.gutenberg.org/`.

Dataset	Median α on Digraph	Median α on Undigraph 1	Median α on Undigraph 2	Median μ on Digraph	Median μ on Undigraph 1	Median μ on Undigraph 2
	(quartile deviations are in parentheses)			(quartile deviations are in parentheses)		
Blog	2.34 (0.17)	2.34 (0.17)	2.41 (0.19)	16.63 (17.16)	22.50 (22.01)	14.93 (9.49)
News	3.38 (0.42)	3.38 (0.42)	4.35 (0.98)	0.63 (0.50)	0.95 (0.76)	1.75 (0.71)
Papers	2.35 (0.09)	2.35 (0.09)	2.45 (0.11)	20.69 (2.96)	27.87 (3.93)	14.95 (1.80)
Digitized Books	2.12 (0.04)	2.12 (0.04)	2.16 (0.05)	244.31 (98.62)	296.73 (116.98)	88.46 (31.78)
All together	2.58 (0.53)	2.58 (0.53)	2.70 (0.90)	5.03 (11.93)	7.27 (15.85)	7.31 (8.47)

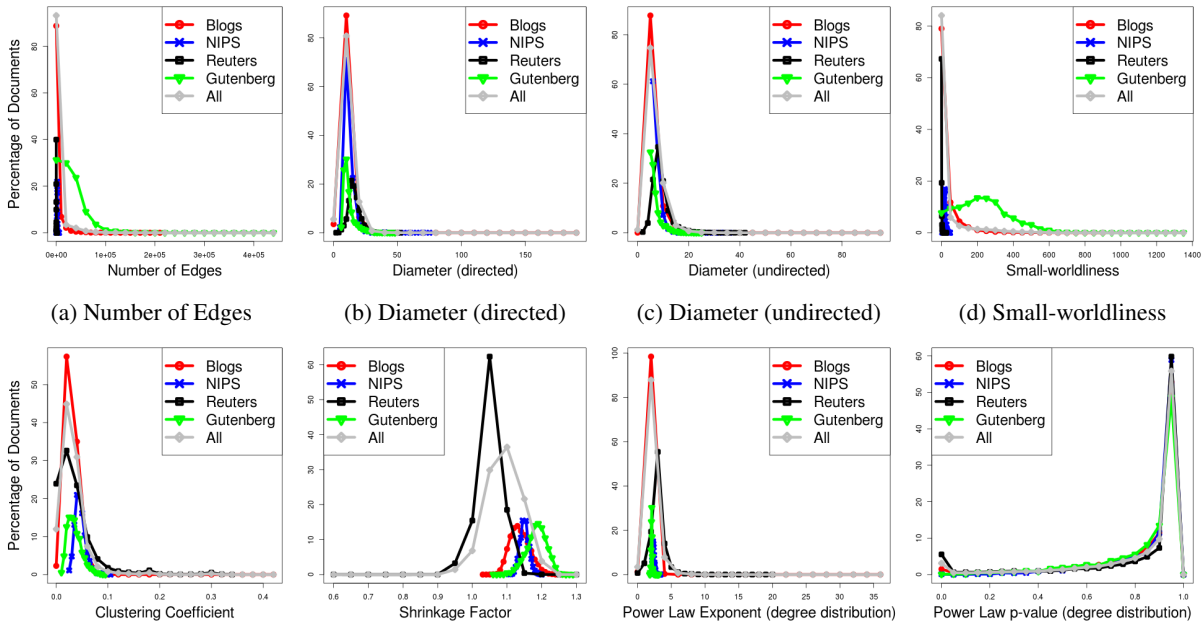
Table 2: Power-law exponent of degree distribution (α) and small-worldliness (μ) of word collocation networks. Here we report the median across documents in a particular dataset (genre), and also the median across all documents in all datasets (last row).

Dataset	Median α on Simplified Digraph	Median α on Simplified Undigraph 1	Median α on Simplified Undigraph 2	Median μ on Simplified Digraph	Median μ on Simplified Undigraph 1	Median μ on Simplified Undigraph 2
	(quartile deviations are in parentheses)			(quartile deviations are in parentheses)		
Blog	2.34 (0.17)	2.34 (0.16)	2.36 (0.18)	16.67 (17.18)	23.28 (22.98)	39.13 (24.03)
News	3.39 (0.42)	3.40 (0.42)	3.88 (0.77)	0.63 (0.50)	0.96 (0.77)	4.96 (1.93)
Papers	2.36 (0.09)	2.37 (0.09)	2.40 (0.11)	20.78 (2.98)	29.18 (4.09)	38.81 (4.75)
Digitized Books	2.12 (0.04)	2.13 (0.04)	2.14 (0.05)	244.53 (98.81)	317.49 (127.14)	218.77 (78.02)
All together	2.58 (0.53)	2.58 (0.54)	2.65 (0.72)	5.04 (11.97)	7.45 (16.52)	19.64 (21.82)

Table 3: Power-law exponent of degree distribution (α) and small-worldliness (μ) of word collocation networks. Here we report the median across documents in a particular dataset (genre), and also the median across all documents in all datasets (last row).

Network Type	Median α on Blogs	Median α on Papers	Median α on News	Median α on Books	Median α on All	Median μ on Blogs	Median μ on Papers	Median μ on News	Median μ on Books	Median μ on All
	(quartile deviations are in parentheses)					(quartile deviations are in parentheses)				
Digraph	2.34 (0.17)	2.35 (0.09)	3.38 (0.42)	2.12 (0.04)	2.58 (0.53)	16.63 (17.16)	20.69 (2.96)	0.63 (0.50)	244.31 (98.62)	5.03 (11.93)
Undigraph 1	2.34 (0.17)	2.35 (0.09)	3.38 (0.42)	2.12 (0.04)	2.58 (0.53)	22.50 (22.01)	27.87 (3.93)	0.95 (0.76)	296.73 (116.98)	7.27 (15.85)
Undigraph 2	2.41 (0.19)	2.45 (0.11)	4.35 (0.98)	2.16 (0.05)	2.70 (0.90)	14.93 (9.49)	14.95 (1.80)	1.75 (0.71)	88.46 (31.78)	7.31 (8.47)
Simplified Digraph	2.34 (0.17)	2.36 (0.09)	3.39 (0.42)	2.12 (0.04)	2.58 (0.53)	16.67 (17.18)	20.78 (2.98)	0.63 (0.50)	244.53 (98.81)	5.04 (11.97)
Simplified Undigraph 1	2.34 (0.16)	2.37 (0.09)	3.40 (0.42)	2.13 (0.04)	2.58 (0.54)	23.28 (22.98)	29.18 (4.09)	0.96 (0.77)	317.49 (127.14)	7.45 (16.52)
Simplified Undigraph 2	2.36 (0.18)	2.40 (0.11)	3.88 (0.77)	2.14 (0.05)	2.65 (0.72)	39.13 (24.03)	38.81 (4.75)	4.96 (1.93)	218.77 (78.02)	19.64 (21.82)

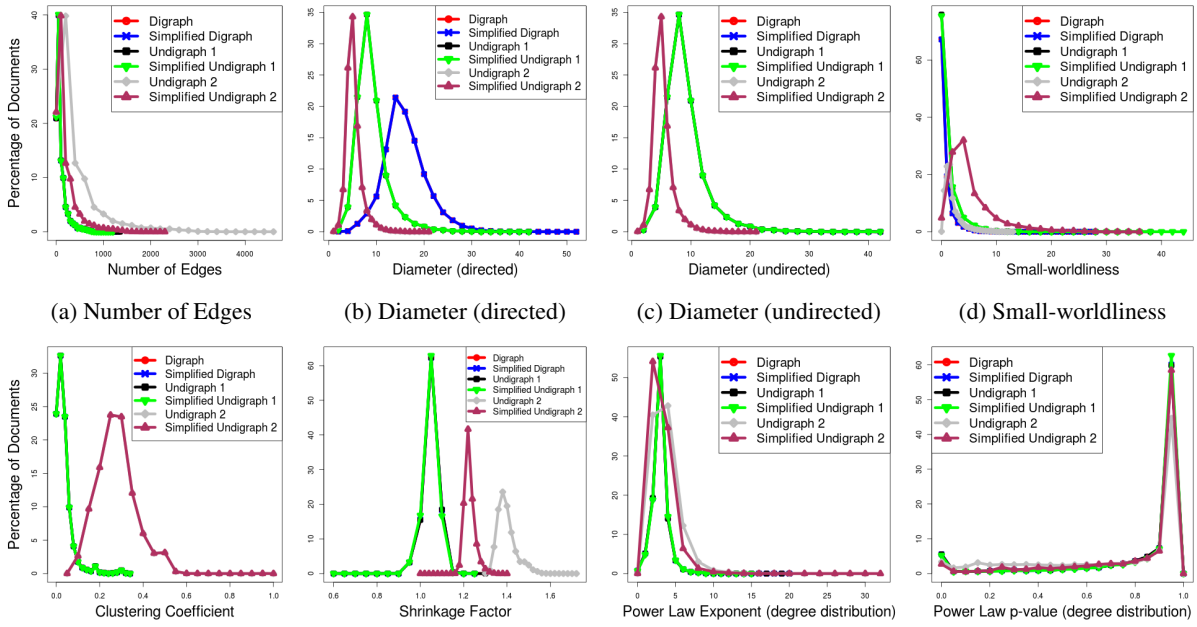
Table 4: Power-law exponent of degree distribution (α) and small-worldliness (μ) of word collocation networks. Here we report the median across documents for a particular network type.



(a) Number of Edges (b) Diameter (directed) (c) Diameter (undirected) (d) Small-worldliness

(e) Clustering Coefficient (f) Shrinkage Exponent (g) α (h) p-value for α

Figure 2: Distributions of eight global network properties across different genres for directed collocation networks. Y-axes represent the percentage of documents for different genres.



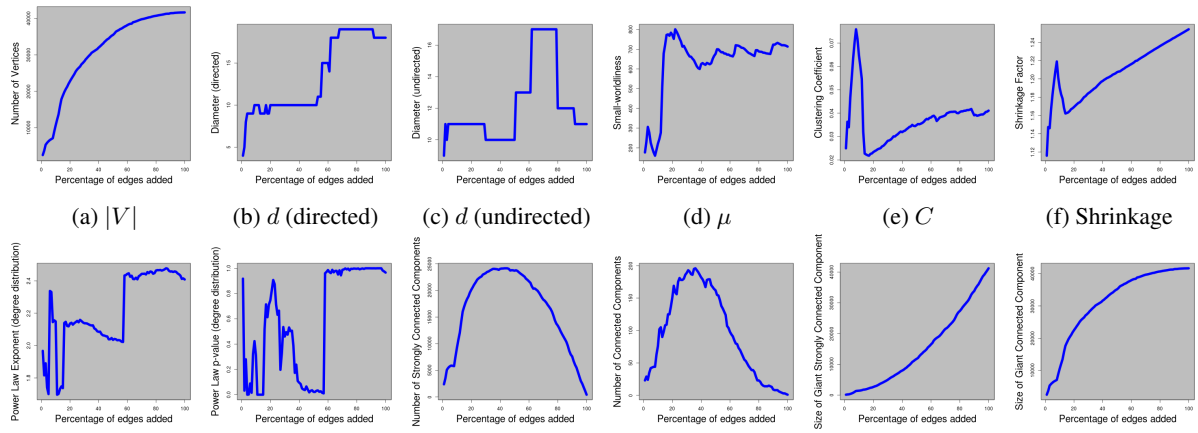
(a) Number of Edges (b) Diameter (directed) (c) Diameter (undirected) (d) Small-worldliness

(e) Clustering Coefficient (f) Shrinkage Exponent (g) α (h) p-value for α

Figure 3: Distributions of eight global network properties across different network types on the *news* genre. Y-axes represent the percentage of documents for different network types.

Test	$\lvert E\rvert$	**Directed** d	**Undirected** d	μ	C	**Shrinkage**	α	**p-value for** α
ANOVA	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Kruskal-Wallis	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
ANOVA	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Kruskal-Wallis	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 5: p-values from ANOVA and Kruskal-Wallis tests. The top two rows are p-values for Figure 2, and the bottom two rows are p-values for Figure 3. Each column corresponds to one subfigure of Figure 2 and Figure 3. p-values in general were extremely low - close to zero in most cases.



(a) $|V|$ (b) d (directed) (c) d (undirected) (d) μ (e) C (f) Shrinkage

(g) α (h) p-value for α (i) Number of SCCs (j) Number of CCs (k) Giant SCC Size (l) Giant CC Size

Figure 4: Change of global network properties with incremental addition of edges to the directed network of *news* genre. SCC = Strongly Connected Component, CC = Connected Component. By "giant" CC and "giant" SCC, we mean the largest CC and the largest SCC. See Table 1 for other properties.

ted with 20 bins. Figure 2e, for example, shows the global clustering coefficient (C) on the X-axis, divided into 20 bins, and the percentage of document networks (directed) with C values falling into a particular bin, on the Y-axis. Histograms from different genres are overlaid. Note from Figure 2e that most distributions are highly overlapping across different genres, thereby putting into question if they are indeed suitable for genre identification. But when we performed ANOVA and Kruskal-Wallis tests to figure out if the distributions were similar or not across different genres, we observed that the corresponding p-values were all < 0.001 (cf. Table 5, top two rows), thereby showing that at least a pair of mean values were significantly apart. Follow-up experiments using unpaired t-tests, U-tests, and Kolmogorov-Smirnov tests (all with Bonferroni Correction for multiple comparisons) showed that indeed almost all distributions across different genres were significantly apart from each other. Detailed results are in the supplementary material. This, we think, is an important and interesting finding, and needs to be delved deeper in future work.

Figure 3 shows histograms of the eight properties from Figure 2, but this time on a *single genre* (news articles), across different network types. This time we observed that many histograms are significantly apart from each other (see, e.g., Figures 3b, 3c, 3e, and 3f). ANOVA and Kruskal-Wallis tests corroborated this finding (cf. Table 5, bottom two rows). Detailed results, including t-tests, U-tests, and Kolmogorov-Smirnov tests are in the supplementary material.

4.3 Change of Global Network Properties with Gradual Addition of Edges

To see how global network properties change as we gradually add edges to a network, we took the whole news collection, and constructed a directed word collocation network on the whole collection, essentially considering the collection as a *super-document* (cf. Section 3). We studied how properties change as we consider top $k\%$ of edges in this super-network, with k ranging from 1 to 100 in steps of 1. The result is shown in Figure 4. Note that the number of connected components and the number of strongly connected components increase first, and then decrease. The number of vertices, size of the largest strongly connected component, and size of the largest connected component increase monotonically as we consider more and more collocation edges. For other properties, we see a lot of unpredictability and spikes (see, e.g., Figures 4d, 4e, 4g, and 4h), especially when the percentage of added edges is small. We performed Runs Test, Bartels Test, and Mann-Kendall Test to figure out if these trends are random, and the resulting p-values indicate that they are not random, and in fact positively correlated (i.e., *increasing*). Details of these tests are in the supplementary material. Note also that all figures except Figures 4a, 4k, and 4l show at least one *phase transition* (i.e., a "jump" or a "bend").

5 Conclusion

We performed an exploratory analysis of global properties of word collocation networks across four different genres of text, and across different network types within the same genre. Our analyses reveal that cross-genre and within-genre variations are statistically significant, and incremental construction of collocation networks by gradually adding edges leads to non-random and positively correlated fluctuations in many global properties, some of them displaying single or multiple *phase transitions*. Future work consists of the inclusion of edge weights; exploration of other datasets, network properties, and network types; and applying those properties to the genre classification task.

Acknowledgments

We would like to acknowledge Dr Rada Mihalcea for her support. This work emerged from a class project in a graduate course on Network Science, given by Prof Réka Albert at Penn State. Sagnik Ray Choudhury provided us with the primary inspiration to do the hard work and write the paper. Finally, we thank the anonymous reviewers whose comments greatly improved this draft. All results, discussions, and comments contained herein are the sole responsibility of the author, and in no way associated with any of the above-mentioned people. The errors and omissions, if any, should be addressed to the author, and will be gratefully acknowledged.

References

[Amancio et al.2011] Diego R. Amancio, Renato Fabbri, Osvaldo N. Oliveira Jr., Maria G. V. Nunes, and Luciano da Fontoura Costa. 2011. Opinion Discrimination Using Complex Network Fea-

tures. In Luciano F. Costa, Alexandre Evsukoff, Giuseppe Mangioni, and Ronaldo Menezes, editors, *Complex Networks*, volume 116 of *Communications in Computer and Information Science*, pages 154–162. Springer Berlin Heidelberg.

[Antiqueira et al.2006] Lucas Antiqueira, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes, Osvaldo Novais Oliveira Jr., and Luciano da Fontoura Costa. 2006. Some issues on complex networks for author characterization. In Solange Oliveira Rezende and Antonio Carlos Roque da Silva Filho, editors, *Fourth Workshop in Information and Human Language Technology (TIL'06) in the Proceedings of International Joint Conference IBERAMIA-SBIA-SBRN*, Ribeiro Preto, Brazil, October 23-28. ICMC-USP.

[Biemann et al.2009] Chris Biemann, Monojit Choudhury, and Animesh Mukherjee. 2009. Syntax is from Mars while Semantics from Venus! Insights from Spectral Analysis of Distributional Similarity Networks. In *ACL/IJCNLP (Short Papers)*, pages 245–248.

[Biemann et al.2012] Chris Biemann, Stefanie Roos, and Karsten Weihe. 2012. Quantifying Semantics Using Complex Network Analysis. In *Proceedings of COLING*.

[Choudhury and Mukherjee2009] Monojit Choudhury and Animesh Mukherjee. 2009. The Structure and Dynamics of Linguistic Networks. In *Dynamics on and of Complex Networks*, pages 145–166. Springer.

[Clauset et al.2009] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, November.

[Csardi and Nepusz2006] Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

[Ferrer i Cancho and Solé2001] Ramon Ferrer i Cancho and Ricard V. Solé. 2001. The Small World of Human Language. *Proceedings: Biological Sciences*, 268(1482):pp. 2261–2265.

[Ferret2002] Olivier Ferret. 2002. Using collocations for topic segmentation and link detection. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Heyer et al.2001] Gerhard Heyer, Martin Läuter, Uwe Quasthoff, Thomas Wittig, and Christian Wolff. 2001. Learning Relations using Collocations. In *Proceedings of the IJCAI Workshop on Ontology Learning, Seattle, USA*.

[Juola2006] Patrick Juola. 2006. Authorship Attribution. *Found. Trends Inf. Retr.*, 1(3):233–334, December.

[Ke and Yao2008] Jinyun Ke and Yao Yao. 2008. Analysing Language Development from a Network Approach. *Journal of Quantitative Linguistics*, 15(1):70–99.

[Ke2007] Jinyun Ke. 2007. Complex networks and human language. *CoRR*, abs/cs/0701135.

[Koppel et al.2009] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26, January.

[Leskovec et al.2007] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph Evolution: Densification and Shrinking Diameters. *ACM Trans. Knowl. Discov. Data*, 1(1), March.

[Liang et al.2009] Wei Liang, Yuming Shi, Chi K. Tse, Jing Liu, Yanli Wang, and Xunqiang Cui. 2009. Comparison of co-occurrence networks of the Chinese and English languages. *Physica A: Statistical Mechanics and its Applications*, 388(23):4901 – 4909.

[Liang et al.2012] Wei Liang, YuMing Shi, Chi K. Tse, and YanLi Wang. 2012. Study on co-occurrence character networks from Chinese essays in different periods. *Science China Information Sciences*, 55(11):2417–2427.

[Masucci and Rodgers2006] Adolfo Paolo Masucci and Geoff J. Rodgers. 2006. Network properties of written human language. *Physical Review E*, 74(2):026102+, August.

[Matsuo et al.2001a] Yutaka Matsuo, Yukio Ohsawa, and Mitsuru Ishizuka. 2001a. A Document as a Small World. In *Proceedings of the Joint JSAI 2001 Workshop on New Frontiers in Artificial Intelligence*, pages 444–448, London, UK, UK. Springer-Verlag.

[Matsuo et al.2001b] Yutaka Matsuo, Yukio Ohsawa, and Mitsuru Ishizuka. 2001b. KeyWorld: Extracting Keywords from a Document as a Small World. In *Proceedings of the 4th International Conference on Discovery Science*, DS '01, pages 271–281, London, UK, UK. Springer-Verlag.

[Mihalcea and Tarau2004] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

[Motter et al.2003] Adilson E. Motter, Alessandro P. S. de Moura, Ying-Cheng Lai, and Partha Dasgupta. 2003. Topology of the conceptual network of language. *Physical Review E*, 65.

[Page et al.1998] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia.

[Saha Roy et al.2011] Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury, and Naveen Kumar Singh. 2011. Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries. In *Proceedings of SIGIR Workshop on Query Understanding and Representation*.

[Schler et al.2006] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, March.

[Stamatatos2009] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, March.

[Stevanak et al.2010] J. T. Stevanak, David M. Larue, and Lincoln D. Carr. 2010. Distinguishing Fact from Fiction: Pattern Recognition in Texts Using Complex Networks. *CoRR*, abs/1007.3254.

[Walsh1999] Toby Walsh. 1999. Search in a Small World. In Thomas Dean, editor, *IJCAI*, pages 1172–1177. Morgan Kaufmann.

[Wasserman and Faust1994] Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November.

[Watts and Strogatz1998] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10.