A Graph-based Approach to Text Genre Analysis

Ahmed Ragab Nabhan[1,2], Khaled Shaalan[3,4]

[1] Fayoum University, Faculty of Computers and Information,
Egypt

[2] Sears Holdings, Hoffman Estates, IL,
USA

[3] The British University in Dubai, Dubai,
United Arab Emirates

[4] School of Informatics, University of Edinburgh,
UK

ahmed.nabhan@gmail.com, khaled.shaalan@buid.ac.ae

Abstract. Genre characterization can be achieved by a variety of methods that employ lexical, syntactic, and presentation features of text to highlight key domain differences and stylistic preferences. However, these traditional methods cannot uncover some important macro-structural features that are embedded in text. Representation of text as a word graph can enable effective frameworks for analysis and identification of key topological features that characterize genres of text. In this study, we investigated graph features such as clustering coefficients, centralization, diameter, and average path lengths for eight text genres. The findings indicated key patterns that vary from a genre to another according to the stylistic differences in text. Furthermore, evidence of subgenres was found through some graph features such as number of connected components and node heterogeneity.

Keywords. Word graphs, genres analysis, topological features.

1 Introduction

Genres analysis is motivated by the idea that there are particular characteristics of texts used for certain communicative purposes such as in academia and business, and that these characteristics can be distinguished from other types of texts [7]. In addition to its importance in linguistic theory, genres analysis was proved practical for a variety of applications in business, information technology, and education fields. For instance, text genre analysis led to identification of some violations of government authoritative guidance related to writing of financial reports that communicate essential information to shareholders [16]. Characterization of text genres is of particular interest and has practical applications for information retrieval and extraction (e.g. automatic genre identification of webpages) [10, 17]. In educational research, genre-based text analysis was proposed as a technique that can aid the development of teaching materials [19, 6].

Some techniques to identify genre characteristics rely mainly on lexical and syntactic constructs of the language such as function words, part-of-speech tags, verb tense, and word class to name a few [11]. For instance, genres-related annotations in text corpora were shown to be useful for identification of genres characteristics [20]. However, these constructs can be only used to highlight frequent patterns of text units such as words, phrases, or sentences, and cannot uncover global patterns embedded in text as an integral entity. There can be important structural features of text (as an entity) that reflect genre

Computación y Sistemas, Vol. 20, No. 3, 2016, pp. 527–539
ISSN 1405-5546
doi: 10.13053/CyS-20-3-2471

characteristics. To this end, modeling of texts as graphs can enable a macro-level analysis of text genres characteristics. There are some patterns and regularities that stem from stylistics preferences and genre rules and can shape or affect the way a word graph is connected ("wired"). There are a set of graph features that can be used in text genres analysis, including average lengths of paths that links nodes representing words, number of connected components (isolated sub-graphs), clustering coefficients, and node centrality. These graph-based features go beyond word and sentence boundaries and therefore provide a global overview of the text being studied. These techniques of modeling text as graphs has drawn attention from physicists and linguistics to study language characteristics and evolution. Networks that represent text corpora were found to demonstrate some interesting complex systems properties [5].

Complex networks are large graphs demonstrating non-trivial characteristics and patterns that emerge from simple interactions between nodes. Complex characteristics include node degree distributions that can be characterized by power-law, hierarchical structures, and assortativity. Language is particularly an interesting instance of complex systems that can be modeled using network representation. Language evolution has been studied based on network-based models to study features such as efficiency and optimization during language growth (in terms of vocabulary). Previous studies suggested that when corpora are converted to graphs they were shown to demonstrate complex network characteristics such as power-law node degree distribution [1, 2].

Complex word networks provide powerful multi-dimensional representation of text utterances that can be useful in studying the linguistic phenomena. Beside theoretical research interests, graph representation of text provides a framework for some language technologies and has been utilized in developing practical methods for natural language applications such as text summarization and text mining [2, 12].

In this paper, graph-based analysis of a diverse set of text genres was shown to provide some practical features that can highlight and contrast differences between text genres. The significance of this graph-based approach stems from the fact that standard natural language processing methods of lexical and syntactic analysis can fall short to uncover non-local text characteristics that go beyond word, phrase, and sentence levels. Some of the graph parameters like global and local clustering coefficients can span multiple sentences and even documents within the same corpus. We also show that this approach can be useful in detecting subgenres in corpora. In addition to graph-based methods, traditional lexical analysis methods were used to explain some differences in the graph characteristics of different genres. In the next section, basic terminology of complex networks are presented.

2 Research Methods

In this section, the network parameters used for text genres analysis are briefly defined. Then, data selection and pre-processing steps as well as network construction methods are presented.

2.1 Complex networks

Network science [5] is a relatively new field with roots in graph theory, computer science, and statistical mechanics. Networks can be used to model many complex entities such as the Internet, social networks, and molecular interaction networks. Networks have popular classes including random networks, small world, and scale free networks [1]. A standard method for studying complex networks is the identification of topological information. There are topological characteristics of complex networks that we discuss in this section [4].

2.1.1 Node degree

Node degree k is the number of edges that link a node to its neighbors. A standard step in analyzing a complex network is to study the degree distribution of nodes and how node degrees vary with other network parameters. Some class of

complex networks, known as scale-free networks was found to follow a power-law degree [1]:

$$P(k) \approx k^{-\gamma}, \tag{1}$$

where γ is an exponent parameter estimated from data. For complex networks representing text corpora, this parameter was found to be near a value of 3.

2.1.2 Clustering Coefficient

This is a network parameter that represents the degree to which nodes in the network tend to cohere or form groups, communities or clusters. There are two clustering coefficients parameters, one is the global coefficient which is based on triplets of nodes. The global clustering coefficient can be measured by the ratio between count of triangles (closed paths between three nodes) and open triplets (paths that link three nodes without forming a triangle). The other parameter is the local clustering coefficient which is measured as the proportion of neighbors (of a specific node) that form a clique or complete graph. The local clustering coefficient C_i of a node i can be computed as

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \tag{2}$$

where k_i is the number of edges connecting node i to its neighboring nodes, and E_i is the number of edges between the k_i neighbors of node i [1].

2.1.3 Average Shortest Path Length

Average path length measures the expected distance or number of edges between two connected nodes. This is also known as characteristic path length. Average shortest path indicates the expected value of the shortest path in a network. This parameter is used in defining other parameters, such as closeness and radiality.

2.1.4 Network Diameter

The largest distance between two nodes in a network is called network diameter.

2.1.5 Number of Connected Components

This parameter is the number of node groups that are pairwise connected. The number of connected components indicates the connectivity of a network; fewer connected components can imply a stronger connectivity within the network.

2.1.6 Average Neighborhood Connectivity

This parameter measures expected number of neighbors to neighbors of a node. This is basically looking into a circle centered at the node of interest and whose diameter is four edges in length. This can be an important parameter to consider when studying the hierarchical characteristics of complex networks.

2.1.7 Network Centralization

This concept refers to tendency of nodes to cohere and staying connected to a center node in a star-like topology.

2.1.8 Network Heterogeneity

The tendency of a network to have hub nodes is termed as heterogeneity. Higher heterogeneity is found when nodes of varying degree tend to have links more often than when nodes of similar degree link together.

2.2 Selection of Text Genres

This study covers a selected set of eight text genres from the Brown corpus. Preprocessing steps were applied to raw text to remove punctuation characters. The categories were selected to represent different communication purposes like education, briefings, and entertainment. A summary of the pre-processed data is listed in Table 1.

Table 1. Statistics of Text Genres

Letter ID	Genre	No. Unique Words
A	News/Reportage	12649
B	Editorial	8873
D	Religion	5871
E	Hobbies	10436
J	Learned	14883
K	Fiction	8471
N	Adventure	8083
R	Humor	4611

Table 2. Complex Networks Summary for Text Genres

Letter ID	Genre	No. Nodes	No. Edges
A	News/Reportage	12649	50706
B	Editorial	8873	32307
D	Religion	5871	20098
E	Hobbies	10436	41727
J	Learned	14883	77689
K	Fiction	8471	31289
N	Adventure	8083	30974
R	Humor	4611	11944

2.3 Analysis of Word Networks

Word bigrams were generated from each text genre. Each bigram represents an edge between two word nodes. These word bigrams were generated such that sentence boundaries were preserved, meaning that no bigrams are formed between an end of a sentence and a start of another.

The frequency of the bigrams were ignored and hence the networks had unweighted edges. Cytoscape software platform [8] was used to manage the data and run the algorithms to compute network characteristics. Basic network summary data is shown in Table 2.

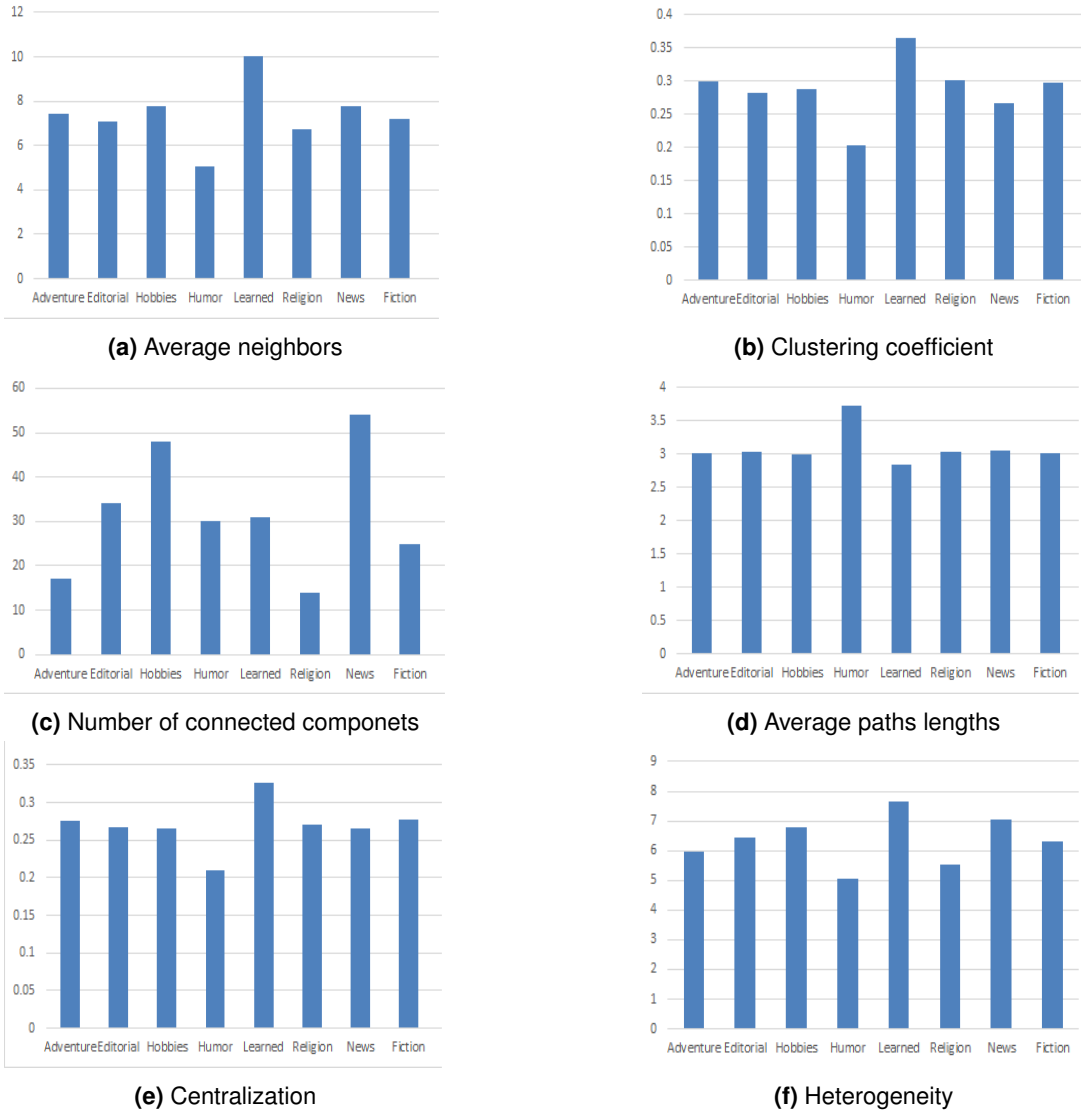
In the next section of the paper, findings of networks analysis are presented. Further lexical analysis steps were applied to data to learn more about the genres differences to highlight some of the network characteristics of genres.

3 Results

In this section, we present graph features measured for the above text genres and discuss implications of these measurements. The below graph features are related and explained according to common stylistic preferences for the genres covered in this study. For some graph parameters, additional lexical features are computed for illustration.

3.1 Network Characteristics

Six simple and complex properties of the network have been calculated for each of the eight text genre. Figure 1 presents the distribution of the six properties across eight genres.



(a) Average neighbors

(b) Clustering coefficient

(c) Number of connected componets

(d) Average paths lengths

(e) Centralization

(f) Heterogeneity

Fig. 1. Basic network statistics

3.1.1 Average Number of Neighbors

The average number of neighbors, also known as average node degree, gives an indication of the density of the network: that is how dense it is wired. For word graphs, a node degree is equivalent to the unigram frequency in traditional language models.

As presented in Figure 1(a), scientific/academic writing genre Learned has the highest average node degree, and the Humor genre has the lowest of all. Higher average node degree may indicate a preference to use existing or established vocabulary than inventing or introducing new ones.

For academic writings, this can be an expected case, where concepts and ideas are introduced and then being clarified across many chapters and sections.

Table 3. Frequency of pronouns in News, Editorial, Fiction, and Adventure text genres

	You	He	She	They
News	66	642	77	267
Editorial	90	364	61	184
Fiction	282	1308	425	313
Adventure	446	1283	377	305

Table 4. Frequency of modal verbs in selected genres

Source/Modal verbs	can	could	may	might	must	will	shall	would	should
Political	22	23	17	9	21	98	1	132	29
Spot	6	11	6	2	4	44	3	27	10
Society Reportage/Cultural	22	7	19	2	6	144	0	9	4

3.1.2 Clustering Coefficients

This network parameter reflects the modularity or structural organization of a given network. A high value would mean there is a tendency of a set of nodes to cohere or stay connected in a subnetwork. Figure 1(b) shows the distribution of clustering coefficient across the eight genres. When there is a particular theme or idea that becomes the focus of the text, this parameter can be expected to have higher value, and to observe substructure/subnetworks emerge.

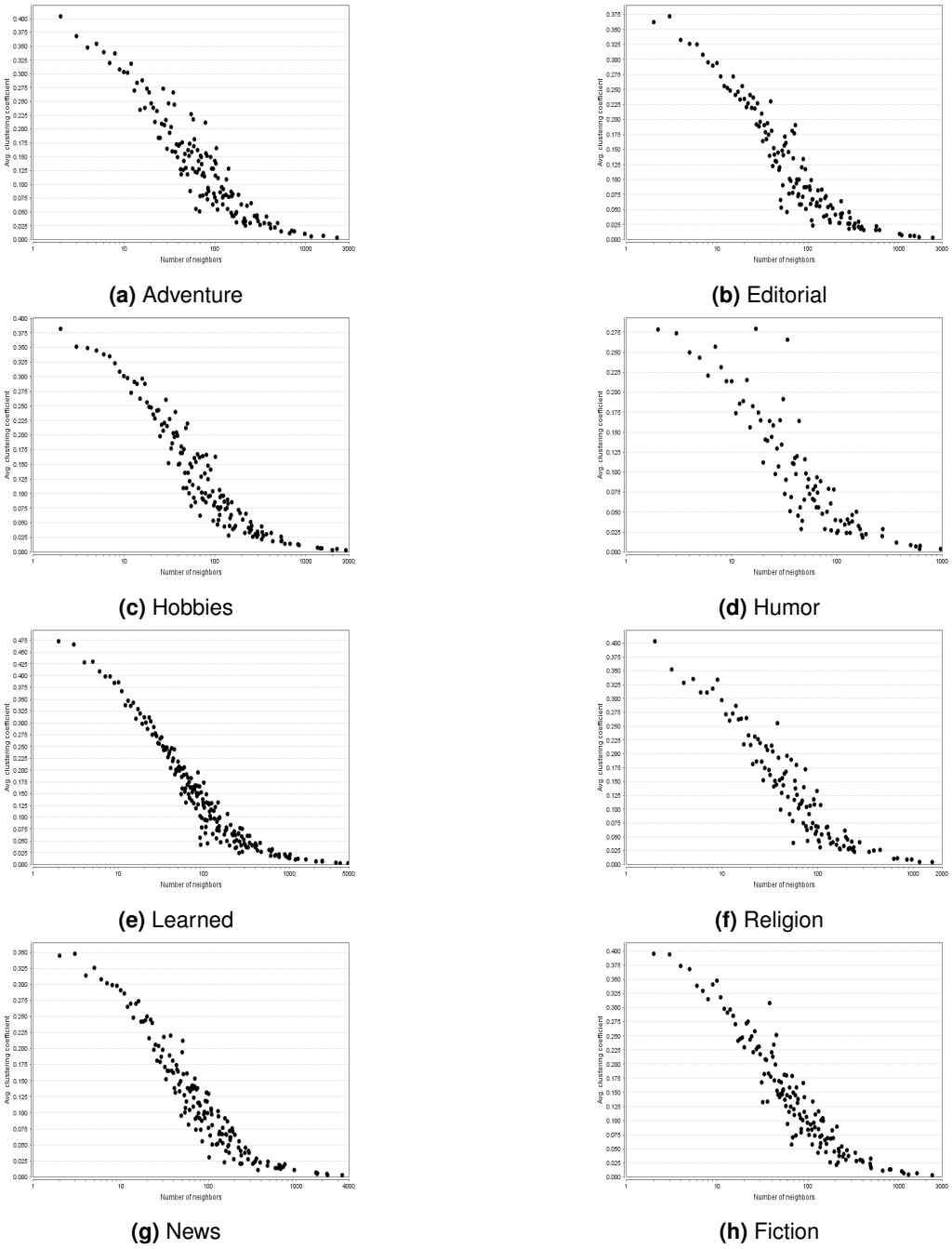
For the text genres covered in this study, genres of Learned, Religion, Fiction, and Adventure had higher clustering coefficient values. Other text genres that are more diverse (in terms of vocabulary) such as News, Editorial, and Hobbies, were found to have a slightly lower value of this parameter. It can be interesting to see a correlation between this parameter and average node degree parameter (Figure 1(c) and Figure 1(a)). Higher clustering coefficients were associated with higher average node degree in respective networks.

3.1.3 Number of Connected Components

This parameter is an indication of connectivity of the network as a whole. High number can be an indication of topic diveristy within a given genre. Figure 1(c) shows that News and Hobbies genres had the biggest number of connected components in the study. On the other hand, Fiction, Adventure, and Religion genres had the smallest number of connected components. Explanation of these differences might be based on the nature of content in each genre.

For instance, Fiction, Adventure, Religion articles tend to have stories revolving around certain concepts and ideas. That means words correlate with a particular theme or topic and can mostly link to words within a community or cluster that is associated with the theme or topic.

For instance, these genres can be expected to employ dialogues between authors and readers, and thus it is expected to observe more of the pronoun “you“ in these genres as compared to News or Learned genres for instance. These pronouns would serve as hubs that connect words within the fiction or adventure genres. Also, in story-telling styles, it is not uncommon to refer to third person pronouns (he, she, and they). These pronouns also can act as hubs in the networks of these genres and hence larger and fewer connected components can be observed. Table 3 shows that there is a tendency to use dialogue in Fiction and Adventure genres as implied by use of second person pronoun “you“. On the other hand, News and Hobbies cover diverse ideas and concepts that might lead to isolated graph components. One possible explanation of this difference in number of connected components is that there might be sub-genres as a result of domain and stylistic differences (in addition



(a) Adventure **(b)** Editorial

(c) Hobbies **(d)** Humor

(e) Learned **(f)** Religion

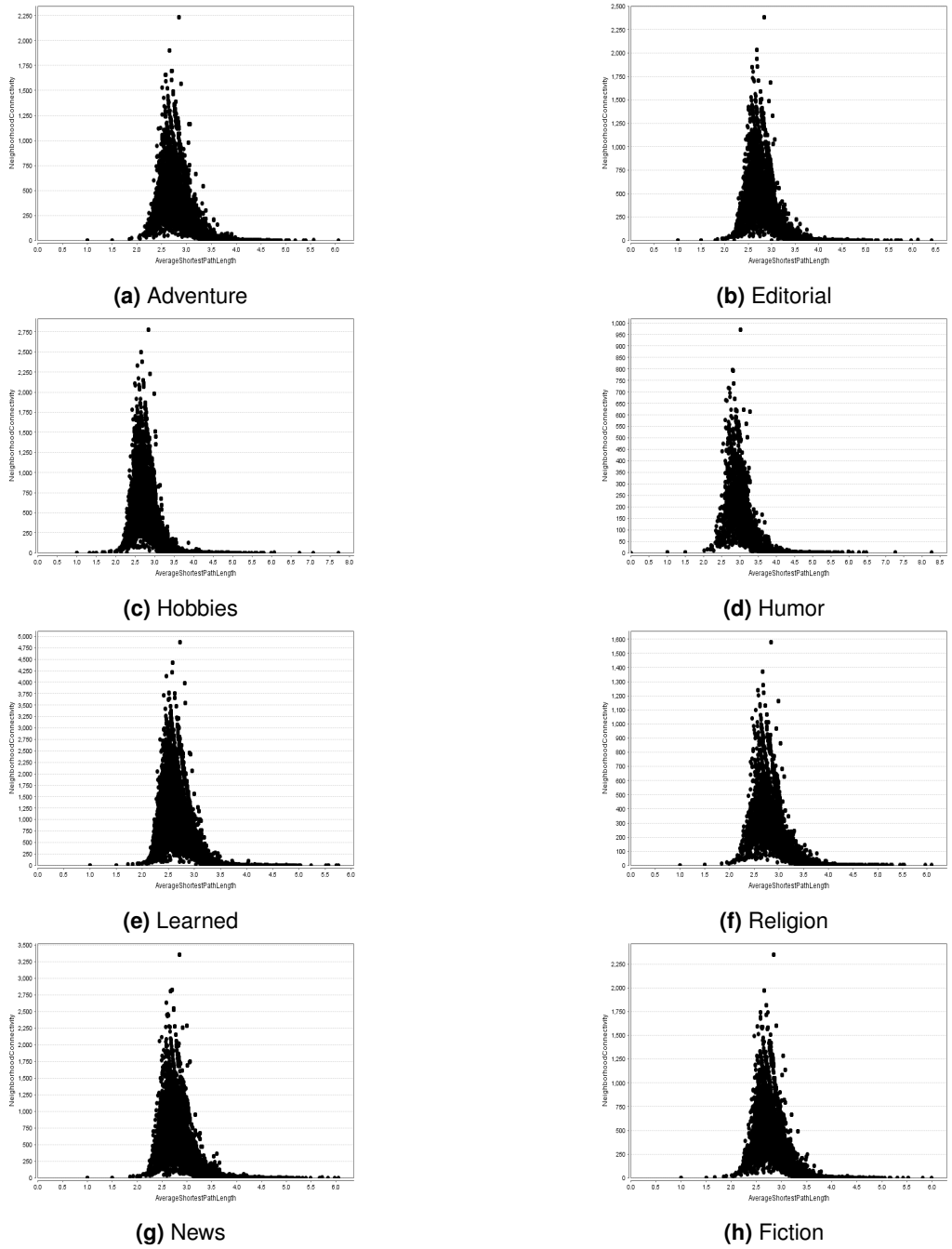
(g) News **(h)** Fiction

Fig. 2. Distribution of average local clustering coefficient as a function of node degree

to content difference) that might lead to isolated subnetworks. This point will be elaborated on in a subsequent section.

3.1.4 Average Shortest Path Length

Figure 1(d) shows global average of shortest path lengths of networks of the eight genres. As shown,



(a) Adventure **(b)** Editorial

(c) Hobbies **(d)** Humor

(e) Learned **(f)** Religion

(g) News **(h)** Fiction

Fig. 3. Neighborhood connectivity per genre as a function of average shortest path length

Learned genre has the smallest average value, indicating greater centrality in the network and that there are more hubs (high degree nodes) in the network. This network parameter is related to another parameter which is the expected number of neighbors or node degree shown in Figure 1(a) .

Table 5. Statistical significance of parameter values of experimental vs. control data. Significant values are in bold text

Network Parameter	Mean(E)	Stddev(E)	Mean(C)	Stddev(C)	t-test
Average Number of Neighbors	7.01	1.38	7.37	0.04	-0.52
Clustering Coefficient	0.28	0.01	0.31	0.01	**-5.76**
Number of Connected Components	28.43	10.3	36	5.54	-1.8
Average Path Length	3.0	0.02	2.94	0.1	**2.13**
Centralization	0.27	0.03	0.28	0.01	-0.86
Heterogeneity	6.16	0.83	7.63	0.05	**-3.54**

3.1.5 Network Centralization

Centralization values are shown in Figure 1(f) Learned, Fiction, and Adventure genres have the highest network centralization values. This is expected given high average node degree and low average shortest path across these networks. High centralization can be an indication of focused topics and ideas in texts.

3.1.6 Network Heterogeneity

Learned, News, and Hobbies have the highest network heterogeneity among the eight genres as shown in Figure 1(f). This reflects tendency of the networks to have links between high degree nodes (hubs) and low degree nodes. This can indicate tendency to use functional keywords in text.

3.2 Genre-specific network characteristics

In the preceding section, a number of graph structural features were shown to vary from one genre to another. Text genre rules and stylistic preferences can influence the way a word graph is wired and thus can give rise to some characteristic patterns. One observed pattern is the relatively high global clustering coefficient of Learned (academic writing) and Religious text genres. On one hand, this might reflect the organization theme and the degree of formality of these texts. On the other hand, it can be an indication of existence of core set of concepts or terminologies and ideas that are used in relation with different subjects within the same genre. This can be due to the nature of academic or religious writings that contains a core set of terms that are used to explain or illustrate topics within these genres.

While global clustering coefficient takes into account high-degree nodes, local clustering coefficient focuses on low degree nodes. Average local clustering coefficient gives an indication of embeddedness of the local node ("The embeddedness measure assesses how much the direct neighbors of a node belong to its own community" [13]). In Figure 2, it can be seen that the average local clustering strongly correlate with node degree and the measured points in that graph looks coherent to reflect that trend. Thus, the formal or academic writing style imposes local clusters that corresponds to specific concepts or ideas (for example, as in Mathematics).

A third feature of the hierarchical organization of academic writing as in Learned genre is the lower average shortest path length for this network as in Figure 1(d). Shorter paths indicate there are more nodes with high connectivity that act as hubs in the network. A fourth key feature of this text genre is high centralization value (Figure 1(e)), which consolidate the idea of a hierarchical structure of this network.

On the other hand, by looking at characteristics of News genre, it is observed that there is lower global clustering coefficient value than that of Religious and Learned genres. This can be an indication of existence of diverse set of terminology and concepts within these genres. In addition, the average path length is higher in cases of News and Fiction genres versus Religion text genre.

3.3 Evidence of Subgenres

Figure 1(c) shows higher number of connected components of News and Hobbies networks. That might suggest that these text genres have more diverse groups of nodes than that can be identified within the network. On the other hand, Religion and Learned genres do have lower number of connected components. A second network parameter that provides an indication of subgenres is neighborhood connectivity of nodes as a function of average lengths of shortest paths passing through these nodes. Figure 3 shows the distribution of neighborhood connectivity as per average shortest path length. The data points are centered around the shortest paths values that are close to the global average shortest paths lengths previously shown in Figure 1(d). The trend in Figure 3 suggests that high neighborhood connectivity is expected where there are shorter paths connecting the nodes, which also supports the evidence of centralization of the network.

For the News genre (Figure 3(h)), it can be seen that there is a high node neighborhood connectivity associated with an average shortest paths length value of 2.5 (this is compared to neighborhood connectivity of other genres at the same average shortest path length). Given that this genre also has a higher number of connected components (isolated groups of nodes) as per Figure 1(c) , higher neighborhood connectivity can further support the hypothesis that these connected components may form sub genres. The Hobbies genre, on the other hand, has a lower number of connected components compared to News genre. Figure 1(d) shows that both networks have very close average shortest paths length across the entire network. However, the neighborhood connectivity is lower per average shortest path in Hobbies genre (Figure 3(c)). Therefore, Figure 3(h) suggests there are centralized connected components in the News/Reportage network which might suggest evidence of existence of subgenres.

There are some categories with communicative purpose and style that are different from those of the mainstream News genre. Examples of different kind of News include the Society, Political, Sports, Financial and Cultural reportage. The society and cultural reportage is expected to refer to named entities likes persons names, places, months more often than expected in financial reportage or political reportage. Society reportage has the communicative purpose of informing audience about social events such as marriages. In addition to content differences within the News/Reportage genre, there are some stylistics differences within the presumed sub-genres. As an example of stylistic differences within the Reportage genre, Table 4 shows differences in usage of modal verbs of cultural, political, and spot Reportage genre.

3.4 Statistical Significance Test

The above network parameter values were compared to a control model consisting of random corpora sampled from five text genres in Brown corpus: government, mystery, romance, belles lettres, and lore genres. Each control corpus was compiled by randomly selecting 20% of sentences from each of the five control genres (i.e. each random corpus represented content from five mixed genres). A sample of 12 random corpora was used as a null model. Word networks were constructed from the random corpora. Each of these networks had approximately 12,000 nodes and 47,000 edges. Network parameter values were computed for items in the control sample, and then statistically evaluated against parameter values of the experimnetal sample of eight genres. A two-sample t-test of the mean of the two populations (experimental vs. control) was conducted for each of the six network parameters. The null hypothesis was set as there was no difference between the two population means. The two samples were assumed to be random, independent, and to follow a normal distribution. The maximum value of each of the six parameters was excluded from the samples. Samples mean, standard deviation, and a degree of freedom of 7 was used to compute the test statistic. A two-tailed test with 0.10 level of significance was performed. Test results are presented in Table 5. Values between braces in table header indicate experimental (E) vs. control (C) measurements.

As shown in Table 5, parameter values of clustering coefficients, average path lengths,

and network heterogeneity is suggested to be statistically significant. The parameter values of average number of neighbors and number of connected components are not suggested as significant and this might be due to the larger standard deviation of values derived from the experimental data. Centralization parameter is shown to have very close values in experimental and control data.

4 Discussion and Conclusions

This study addressed the problem of text genres analysis using graph-based methods to identify macro-structural characteristics embedded in text. While traditional methods of genre analysis rely mainly on lexical, syntactic and presentation features, the current study presents a new approach to genres analysis using complex networks analysis methods. The data analysis demonstrated that network parameters can be used to characterize genres. For instance, with respect to focused and educational genres, average shortest path lengths tend to be smaller compared to other genres such as news. The proposed method relied on global parameters of text, compared to other genre characterization methods that rely on limited set of lexical terms or specific syntactic structure. Moreover, the proposed method were shown to be feasible in detecting sub-genres, as shown in analysis of the News genre. The proposed network based approach in this paper advances the state-of-the-art of genres analysis by demonstrating the practical use of complex network parameters and graph algorithmic methods (e.g. shortest path finding) in identifying differences of genres. Furthermore, the approach was shown to be feasible for more interesting genres analysis tasks such as recognition of sub-genres.

There are network topological differences between text genres that can be attributed to stylistic and content domain differences in text. Key network parameters, such as average shortest paths lengths, clustering coefficient, neighborhood connectivity and number of connected components, were observed to vary systematically from one genre to another. In addition, evidence of existing sub-genres was found in network characteristics of the News text genre. Network models of text, such as the ones used in this study, can answer some interesting questions that are subtle to address via traditional approaches to corpus linguistics.

The computed parameters values of the experimental network data were compared to a control model of an ensemble of corpora compiled by randomly selecting sentences from five Brown text genres not used in the experimental data. Three parameters, namely: clustering coefficients, average path lengths, and network heterogeneity, showed statistically significant values compared to parameter values of the control sample. The values of the sample network parameter of the average number of neighbors were shown not to be statistically significant. This can be expected because complex word networks are known to belong to the scale-free class of networks, which demonstrate a universal power-law degree distribution. The number of connected components parameter did not demonstrate significant values from the sample, but the control sample showed a slightly higher average number of connected components. Also, the standard deviation of the experimental data was higher, even after removing a maximum value from the sample, which might had affected the calculation of t-test statistics. However, the mean values of this parameter suggest that there was a greater variability in the eight genres of the experimental data which might indicate genre-specific characteristics.

This study presented an approach to the problem of genre characterization, in contrast with the genre classification problem. The scope of this study was genre characterization based on global and local features of complex word networks constructed from a corpus. Using network parameters (e.g. average path length and clustering coefficients) for classification might not be robust enough to produce good classification accuracy. Previous studies had focused on genre classification problem from a feature selection perspective. For instance, Lee and Miaeng developed a method for genre classification based on genre and topic specific features [9]. Petrenz and Webber proposed a method for genre

classification using lexical and syntactic features [14]. Graph-based methods were developed for various text analysis problems including text categorization [3] and sentiment analysis [18].

One of the limitations of this study is that it relies solely on the Brown corpus, which has been in use for decades. This corpus is limited in size compared to recent corpora, thus the findings may need to be treated with caution. The choice of Brown corpus was motivated by the fact that it has single labels for the documents, compared to other public corpora that can have multiple categories per document (e.g. Reuters corpus [15]). Despite of this data representativeness issue, the approach proposed in this paper can be generally applied to analysis of other corpora. Besides, the proposed graph-based approach does not have language-specific parameters, and hence can be applied to texts from languages other than English.

Acknowledgment

The authors would like to thank an anonymous reviewer for critical views and positive feedback that helped improve this study.

References

1. **Albert, R. & Barabási, A.-L. (2002).** Statistical mechanics of complex networks. *Reviews of modern physics*, Vol. 74, No. 1, pp. 47.
2. **Amancio, D. R., Antiqueira, L., Pardo, T. A., da F. COSTA, L., Oliveira Jr, O. N., & Nunes, M. G. (2008).** Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, Vol. 19, No. 04, pp. 583–598.
3. **Angelova, R. & Weikum, G. (2006).** Graph-based text classification: learn from your neighbors. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 485–492.
4. **Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., & Albrecht, M. (2008).** Computing topological parameters of biological networks. *Bioinformatics*, Vol. 24, No. 2, pp. 282–284.
5. **Barabási, A.-L. (2013).** Network science. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 371, No. 1987, pp. 20120375.
6. **Chang, C.-F. & Kuo, C.-H. (2011).** A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes*, Vol. 30, No. 3, pp. 222–234.
7. **Dudley-Evans, T. (2000).** Genre analysis: a key to a theory of esp? *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*, , No. 2, pp. 3–11.
8. **Kohl, M., Wiese, S., & Warscheid, B. (2011).** Cytoscape: software for visualization and analysis of biological networks. *Data Mining in Proteomics: From Standards to Applications*, pp. 291–303.
9. **Lee, Y.-B. & Myaeng, S. H. (2002).** Text genre classification with genre-revealing and subject-revealing features. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 145–150.
10. **Lim, C. S., Lee, K. J., & Kim, G. C. (2004).** Automatic genre detection of web documents. *International Conference on Natural Language Processing*, Springer, pp. 310–319.
11. **Luštrek, M. (2006).** Overview of automatic genre identification. *Ljubljana, Slovenia: Jožef Stefan Institute, Department of Intelligent Systems*.
12. **Nabhan, A. R. (2014).** *Graph Pattern Mining Techniques to Identify Potential Model Organisms*. Ph.D. thesis, The University of Vermont.
13. **Orman, G. K., Labatut, V., & Cherifi, H. (2012).** Comparative evaluation of community detection algorithms: a topological approach. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2012, No. 08, pp. P08001.
14. **Petrenz, P. & Webber, B. (2011).** Stable classification of text genres. *Computational Linguistics*, Vol. 37, No. 2, pp. 385–393.
15. **Rose, T., Stevenson, M., & Whitehead, M. (2002).** The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. *LREC*, volume 2, pp. 827–832.
16. **Rutherford, B. A. (2005).** Genre analysis of corporate annual report narratives a corpus linguistics–based approach. *Journal of Business Communication*, Vol. 42, No. 4, pp. 349–378.

17. **Santini, M. (2007).** Characterizing genres of web pages: Genre hybridism and individualization. *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, IEEE, pp. 71–71.
18. **Taboada, M., Brooke, J., & Stede, M. (2009).** Genre-based paragraph classification for sentiment analysis. *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Association for Computational Linguistics, pp. 62–70.
19. **Thompson, S. (1994).** Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes*, Vol. 13, No. 2, pp. 171–186.
20. **Webber, B. (2009).** Genre distinctions for discourse in the penn treebank. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, Association for Computational Linguistics, pp. 674–682.

Ahmed Ragab Nabhan has a PhD in Computer Science from University of Vermont, USA. He is a senior software engineer with Sears Holdings Corporation sepcializing in information retrieval. He is also a lecturer in Computer Science, Faculty of Computers and Information, Fayoum University, Egypt. Dr. Nabhan's research is focused on graph data mining, computational biology, complex networks, and statistical natural language processing.

Khaled Shaalan is a full professor of Computer Science at the British University in Dubai (BUiD), UAE, an Honorary Fellow at the School of Informatics, University of Edinburgh (UoE), UK, and a tenured full professor of Computer Science and Information (on Secondment) at the Faculty of Computers and Information (FCI), Cairo University (CU), Egypt. Recently, Prof Shaalan has been contributing to a wide range of research topics in Arabic Natural Language Processing, including machine translation, parsing, spelling and grammatical checking, named entity recognition, and diacritization. He has published over 100 referred publications and the impact of my research using GoogleScholar H index metric is 20. Prof Shaalan has actively and extensively supported the local and international academic community. He is the founder and CoChair of The International Conference on Arabic Computational Linguistic (ACLing).

Article received on 09/01/2016; accepted on 04/03/2016.
Corresponding author is Ahmed Ragab Nabhan.