Author Attribution using Network Motifs

Younis Al Rozz and Ronaldo Menezes

BioComplex Laboratory, Computer Science
Florida Institute of Technology, Melbourne, USA
yyounis2013@my.fit.edu
rmenezes@cs.fit.edu

Abstract. The problem of recognizing the author of unknown text has concerned linguistics and scientists for a long period of time. The authorship of the famous Federalist Papers remained unknown until Mosteller and Wallace solved the mystery in 1964 using the frequency of functional words. After that, many statistical and computational studies were published in the fields of authorship attribution and stylistic analysis. Complex networks, gaining much popularity in recent years, may have a role to play in this field. Furthermore, several studies show that network motifs, defined as statistically significant subgraphs within a network, have the ability to distinguish networks from distinctive disciplines. In this paper, we succeed in the utilization of network motifs to distinguish the writing style of 10 famous authors. Using statistical learning algorithms, we achieved an accuracy of 77% in classifying 100 books written by 10 different authors, which outperformed the results from other works. We believe that our method proved the importance of network motifs in author attribution.

Keywords: *word co-occurrence networks, author attribution, network motif, classification.*

1 Introduction

An author's writing style can be considered as an example of a behavioral biometric. The words used by people and the way they structure their sentences is unique, and can frequently be used to identify the author of a certain work. The task of author attribution gained attention among researchers in the fields of statistical physics, natural language processing, and data and information science. A thorough survey of the techniques used in authorship attribution can be found in [18]. Applications of authorship attribution are not only limited to literature stylometry [4] but also expands to other fields such as social media forensic [16] and email fraud detection [8]. As researchers find complex networks a promising field in linguistic studies [2], more and more authorship attribution works based on text networks saw the light of day. Measurement from word co-occurrence network topology combined with traditional statistical methods like frequency of functional words and intermittency were used to attribute authors [3, 1].

Network motif defined by Milo *et al.* [12] as a statically significant subgraphs pattern occurred in real-world networks compared to random ones, has gained a lot of attention because of its ability in discriminating networks from different discipline [19].

In this work, we utilized network motifs as a fingerprint to attribute authors by their writing style. More precisely, we extract network motifs from directed co-occurrence networks of 100 books by 10 well-known authors and then we use 5 machine learning algorithms to classify the authors by their network motif signature. We show that 4-nodes directed network motifs alone can be utilized to attribute authors of different books.

The paper is organized as follows. Section 2 is an overview of the efforts spent by other researchers on the subject of author attribution. In Section 3 we describe our dataset and steps taken place in order to extract the network motif from the text networks. The classification methods and results explained in section 4. Finally, we conclude our work in section 5 with a roadmap for future work.

2 Motivation and Related Work

Several studies exist that deal with the importance of network motifs in natural language networks. The first attempt to classify different networks including word co-occurrence using network motifs was made by Milo *et al.* [11]. Li *et al.* [9] extracted and studied three and four nodes directed motif structure of 72,923 two-character Chinese words network. They found that feed-forward loop (FFL) motif structure is significant in their network. Rizvić *et al.* [15] examined three nodes (triads) network motifs extracted from directed co-occurrence networks of five Croatian texts and compared their results with other languages. They realized that there is a similarity between the Croatian language networks triad significance profiles and other previously studied languages. Cabatbat *et al.* [7] compared five-nodes network motifs among other network measures of the Bible and the Universal Declaration of Human Rights (UDHR) translations in eight languages. Pearson correlation coefficient and mutual information were used to compare the metrics of real texts with random texts from other sources. Their finding is that the distribution of network motif frequency is beneficial in recognizing similar texts. Biemann *et al.* [6] realized that motif signatures serve to discriminate co-occurrence networks of natural language from artificially generated ones. To assist their finding, they present additional results on peer-to-peer streaming, co-authorship, and mailing networks. The directed motif of size 3 and undirected motif of size 4 was used in their work.

All the previous works showed the ability of various size network motifs of discriminating text from different languages and genre. They did not utilize machine learning algorithms to support their findings. On the other hand, Marinho *et al.* [10] achieved 57.5% accuracy in their best scenario of attributing eight authors of 40 novels with three nodes directed network motifs. An important aspect of author attribution task is the feature frequency [18]. To capture an author style more preciously, the feature should be more frequent. This motivates us to use the frequency of the 199 four nodes directed network motif in an attempt to attribute the authors under study.

3 Datasets and Methodology

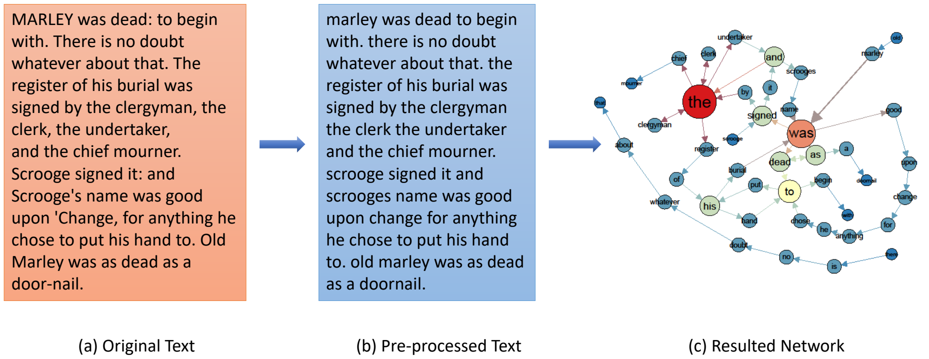
3.1 Data Collection and Network Creation

The dataset used in this work comprised of 100 literature books authored by 10 different authors; 10 books for each individual author. The books are listed in Table 1, and were collected from the Project Gutenberg website[1]. Each book was limited to 20 thousand words which is the length on the shortest book in the set. Text pre-processing steps were applied to remove punctuation, numbers and non-Latin alphabets and all letters were converted to lowercase. We preserved functional words (stop words) in the text as their frequency has been proven to reflect stylistic aspects of the text and improve authorship attribution task [13, 5, 17]. A sample text from Charles Dickens's "A Christmas Carol" novel and the resulted pre-processed text are shown in (Fig. 1(a) and (b) respectively) to illustrate this process.

Table 1: Authors used in our experiments and their book titles.

Authors	Book Titles
Bernard Shaw 1856-1950	Man and Superman, Candida, Arms and the Man,The Philanderer, Caesar and Cleopatra, Pygmalion, Major Barbara, Heartbreak House, The Devil's Disciple, Cashel Byron's Profession.
Charles Dickens 1812-1870	A Christmas Carol, A Tale of Two Cities, The Pickwick Papers, Oliver Twist, Great Expectations, David Copperfield, Little Dorrit, Our Mutual Friend, The Life and Adventures of Nicholas Nickleby, Dombey And Son.
George Eliot 1819-1880	The Essays of George Eliot, Impressions of Theophrastus Such, Silas Marner, Scenes of Clerical Life, The Mill on the Floss, Adam Bede, Romola, Daniel Deronda, Felix Holt The Radical, Middlemarch.
Herbert George Wells 1866-1946	Tales of Space and Time, The Food of the Gods and How It Came to Earth, The Country of the Blind, And Other Stories, The Invisible Man, The First Men in The Moon, The Island of Doctor Moreau, The War of the Worlds, The Time Machine, In the Days of the Comet, Ann Veronica.
Jack London 1876-1916	The Call of the Wild, White Fang, The Iron Heel, Before Adam, Martin Eden, The People of the Abyss, The Night-Born, The Sea Wolf, South Sea Tales, The Valley of the Moon.
Mark Twain 1835-1910	The Adventures of Tom Sawyer, Adventures of Huckleberry Finn, Life on The Mississippi, The Mysterious Stranger and Other Stories, A Tramp Abroad, Following the Equator, The Innocents Abroad, Roughing It, The Prince and The Pauper, A Connecticut Yankee in King Arthur's Court.
Oscar Wilde 1854-1900	A House of Pomegranates, The Duchess of Padua, Vera, Lady Windermere's Fan, A Woman of No Importance, Intentions, An Ideal Husband, Lord Arthur Savile's Crime and Other Stories, The Importance of Being Earnest, The Picture of Dorian Gray.
Sir Arthur Conan Doyle 1859-1930	Rodney Stone, The Adventures of Sherlock Holmes, A Duet, The Tragedy of The Korosko, The Refugees, Uncle Bernac, The Valley of Fear, The Hound of the Baskervilles, Sir Nigel, The Lost World.
William Henry Giles Kingston 1814-1880	Hendricks the Hunter, The Three Lieutenants, The Three Midshipmen, The Three Commanders, Peter the Whaler, Ben Burton, The Three Admirals, Adventures in Africa, In the Wilds of Florida, Peter Trawl.
William Shakespeare 1564-1616	Hamlet, Prince of Denmark, The Life of Henry the Fifth, The Merchant of Venice, The Tragedy of Antony And Cleopatra, The Tragedy of Coriolanus,The Tragedy of Julius Caesar, The Tragedy of King Lear, The Tragedy of Othello, Moor of Venice, The Tragedy of Romeo And Juliet, The Winter's Tale.

[1] http://www.gutenberg.org



MARLEY was dead: to begin with. There is no doubt whatever about that. The register of his burial was signed by the clergyman, the clerk, the undertaker, and the chief mourner. Scrooge signed it: and Scrooge's name was good upon 'Change, for anything he chose to put his hand to. Old Marley was as dead as a door-nail.

marley was dead to begin with. there is no doubt whatever about that. the register of his burial was signed by the clergyman the clerk the undertaker and the chief mourner. scrooge signed it and scrooges name was good upon change for anything he chose to put his hand to. old marley was as dead as a doornail.

(a) Original Text

(b) Pre-processed Text

(c) Resulted Network

Fig. 1: Sample text from Charles Dickens's "A Christmas Carol" novel showing the stages of text preprocessing and the co-occurrence network created from the text.

Next, we created the directed co-occurrence networks from the result of the pre-processed text of the 100 books. Co-occurrence networks can be constructed based on the sentence, paragraph, or the whole text boundary. We chose the sentence boundary as it produces less dense network hence, reduces the amount of time required to extract network motifs. Sentence boundary is defined by period, exclamation point, and question mark [14]. The network constructed from the pre-processed text is depicted in (Fig. 1(c)).

3.2 Feature Extraction

A plethora of network motif extraction tools exist, each one has its pros and cons related to the number of motif's nodes count and the algorithm speed. We chose the iGraph [2] implementation for its flexibility and fast execution time. Tran *et al.* [19] suggested that small undirected network motifs cannot reveal differences among networks from different disciplines, while large ones do. Based on this argument and the importance of feature frequency explained at the end of section 2, we chose the directed 4-node network motifs shown in Fig. 2. For each book in the dataset, we extracted the 199 motifs from their directed network and then a data frame contains the motifs frequencies was created. Fig. 3 illustrate a sample 4-node directed motif extracted from the example network of (Fig. 1(c)). The frequency distribution of the extracted 4-node motifs from the books of Bernard Shaw, H. G. Wells, Jack London and William Shakespeare shown in Fig. 4.

4 Motif-based Classification

For this part of the work, we utilized 5 supervised machine learning classification algorithms namely K nearest neighbors (KNN), decision trees, random forests, support

[2] http://igraph.org

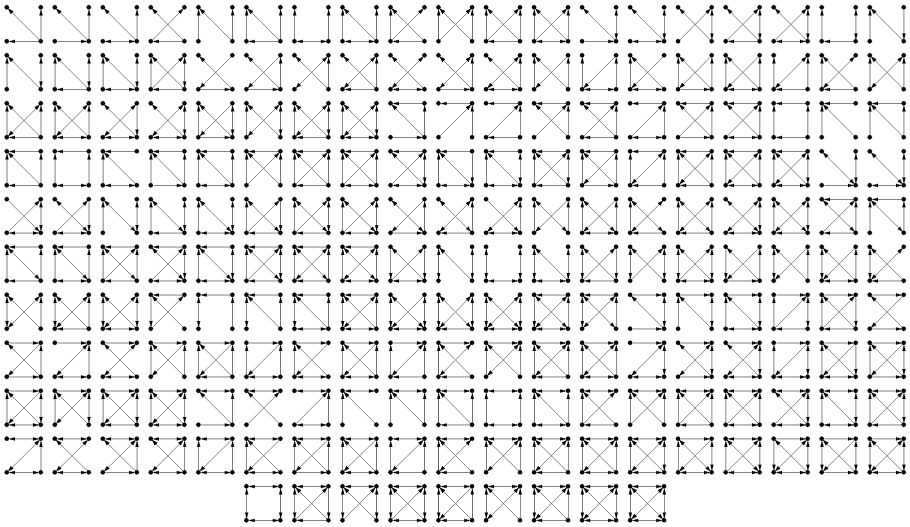


Fig. 2: 199 different orientation of the directed 4-node network motif.

vector machines (SVM), and multi-layer perceptrons (MLP). They are all part of the scikit-learn [3] machine learning package for Python. As we try to attribute 10 authors, we have a multi-class classification problem with the number of samples ($N = 100$) which represents the number of books and the dimension of the feature set ($D = 199$) was relatively high. We used two cross-validation methods, the first one is to split our dataset into 75% training set and 25% testing set and then shuffle the dataset and repeat the operation for 100 times. The second method was leave-one-out, where the dataset is split into 99 sample for training and one sample for testing then iterate through the remaining samples. The average classification accuracy was calculated with both methods for all the algorithms used in the work. All the dataset were standardized by scaling to unit variance and removing the mean. The classification was performed on all the feature sets, that is the whole 199 4-nodes directed motifs and then recursive feature elimination (RFE) feature selection method used to find the best 75%, 50%, 25%, and 10% features respectively. An alternative method mostly used in the literature is to choose significant motifs based on the highest Z-scores, but we preferred to collect the whole set of motifs and then use feature selection methods to choose the best set.

The results of classification using the first cross-validation method of shuffling and splitting the dataset are listed in Table 2, while Table 3 lists the results of the leave-one-out cross-validation method. As can be seen from both tables, the two basic classification algorithms KNN and the decision trees did not perform well compared to more sophisticated algorithms. Although KNN gives us an average accuracy of 60% when using 25% of the dataset and the leave-one-out cross-validation method, it is still lower than the accuracy obtained from the other classification methods. The best classifica-

[3] http://scikit-learn.org

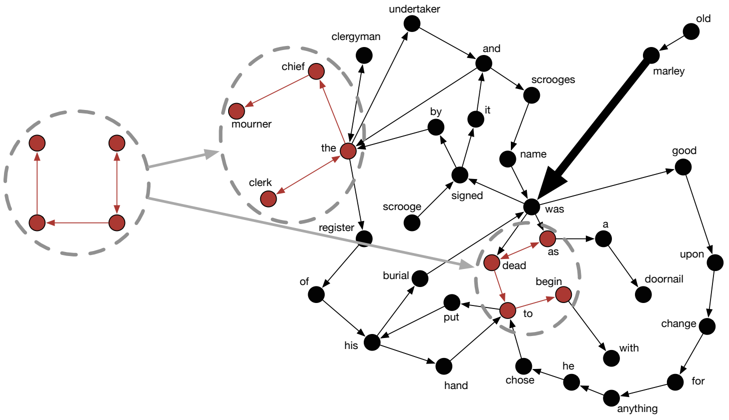


Fig. 3: 4-node directed network motif sample from the network of Fig. 1.

tion accuracy of 77% was obtained when the MLP classifier used with leave-one-out validation method.

Table 2: Average classification accuracy results for the four nodes directed motifs when splitting the dataset into 75% for training set and 25% for testing set with 100 times random shuffling.

	Complete set	75%	50%	25%	10%
KNN	42	42	48	53	52
Decision Tree	41	45	46	45	50
Random Forest	56	58	58	59	64
SVM	53	56	62	63	67
MLP	66	68	70	70	68

5 Conclusion and Future Work

Throughout this work, we attempted to attribute 10 authors of 100 books using 4-nodes directed network motifs. Functional words (stop words) were kept during text pre-processing as they proven by many previous works to reflect author style and increase the accuracy of attributing authors. The results we obtained herein outperformed other works when network motifs were the only feature used in attributing authors. Also, the number of 100 books used in this work are much higher than other works, which statistically means if we used the same smaller dataset, we will get better classification accuracy. This proves the importance of network motifs in recognizing the variety of writing styles among different authors. This opens the door for future work

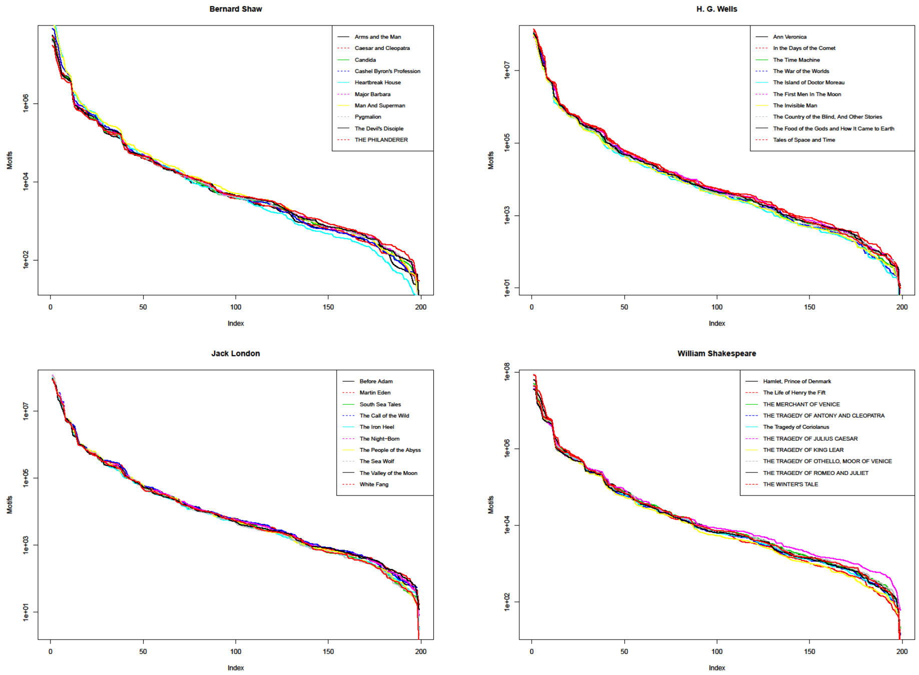


Fig. 4: 4-node network motif sorted frequency of the networks created from the books by Bernard Shaw, H. G. Wells, Jack London and William Shakespeare.

to generalize this method in attributing text from a different genre and translation assessment. Other possibilities are to study the effect of extracting higher motif order on the accuracy of classification.

References

1. Camilo Akimushkin, Diego Raphael Amancio, and Osvaldo Novais Oliveira Jr. Text authorship identified using the dynamics of word co-occurrence networks. *PloS one*, 12(1):e0170527, 2017.
2. Younis Al Rozz, Harith Hamoodat, and Ronaldo Menezes. Characterization of written languages using structural features from common corpora. In *Workshop on Complex Networks CompleNet*, pages 161–173. Springer, 2017.
3. Diego Raphael Amancio. A complex network approach to stylometry. *PloS one*, 10(8):e0136076, 2015.
4. Ahmed Shamsul Arefin, Renato Vimieiro, Carlos Riveros, Hugh Craig, and Pablo Moscato. An information theoretic clustering approach for unveiling authorship affinities in shakespearean era plays and poems. *PloS one*, 9(10):e111445, 2014.
5. Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1991.
6. Chris Biemann, Lachezar Krumov, Stefanie Roos, and Karsten Weihe. Network motifs are a powerful tool for semantic distinction. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 83–105. Springer, 2016.

Table 3: Average classification accuracy results for the four nodes directed motifs with leave-one-out cross-validation method.

	Complete set	75%	50%	25%	10%
KNN	41	42	51	60	54
Decision Tree	44	55	47	55	57
Random Forest	61	56	62	60	65
SVM	61	66	70	66	71
MLP	72	73	75	77	72

7. Josephine Jill T Cabatbat, Jica P Monsanto, and Giovanni A Tapang. Preserved network metrics across translated texts. *International Journal of Modern Physics C*, 25(02):1350092, 2014.
8. Xiaoling Chen, Peng Hao, Rajarathnam Chandramouli, and KP Subbalakshmi. Authorship similarity detection from email messages. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 375–386. Springer, 2011.
9. Jianyu Li, Feng Xiao, Jie Zhou, and Zhanxin Yang. Motifs and motif generalization in chinese word networks. *Procedia Computer Science*, 9:550–556, 2012.
10. Vanessa Queiroz Marinho, Graeme Hirst, and Diego Raphael Amancio. Authorship attribution via network motifs identification. In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pages 355–360. IEEE, 2016.
11. Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.
12. Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
13. Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.
14. G. Nunberg. *The Linguistics of Punctuation*. CSLI lecture notes. Cambridge University Press, 1990.
15. Hana Rizvić, Sanda Martinčić-Ipšić, and Ana Meštrović. Network motifs analysis of croatian literature. *arXiv preprint arXiv:1411.4960*, 2014.
16. Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33, 2017.
17. Santiago Segarra, Mark Eisen, and Alejandro Ribeiro. Authorship attribution through function word adjacency networks. *IEEE transactions on signal processing*, 63(20):5464–5478, 2015.
18. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, 2009.
19. Ngoc Tam L Tran, Luke DeLuccia, Aidan F McDonald, and Chun-Hsi Huang. Cross-disciplinary detection and analysis of network motifs. *Bioinformatics and Biology insights*, 9:49, 2015.