Predicting Text Authorship Using Alignment Free Network Comparison

Sam Alptekin, Sam Berning, and Joe Kimlinger

Motivation and Background

No two voices sound the same. Whether speaking or writing, people are distinguishable by the way they convey their ideas. We hypothesize that we will be able to predict the author of a text based on that authors writing style.

We wanted to determine whether the structure of speech (i.e. the patterns in which words are used together) was an important characteristic of an author's writing style and whether the structure could be used on its own to accurately determine the author of a text. In order to represent the underlying structure of a text, we encoded each of the texts as a network, where nodes are words in the text and edges exist between adjacent words in the text.

Only until recently have the advancements of network science been applied to the field of author attribution[1]. By coding text into a network, we can use established network analysis techniques in order to study the characteristics of the text. Can these characteristics help determine who the text was written by? Some have explored this using global network properties such as clustering coefficient, degree correlation, and a divergence metric used in dynamic network analysis[1, 2]. There have also been studies with successful attempts at determining the author of a text using network motifs[3] and a wide variety of other network features[4, 5, 6]. However, to our knowledge, no one has yet tried determining author using graphlets.

Graphlets are defined to be a subgraph of a graph such that every node in the subgraph is connected to some other node in the subgraph, every edge that exists between two nodes in the original graph is part of the subgraph. The analysis in this paper is based on graphlets of up to four nodes. Each graphlet has a set of automorphic orbits that are defined as follows: the automorphism of a graph is a bijection from the set of nodes of a graph to themselves that preserves the edges in the graph and an orbit is formed by a group of nodes that can be mapped to each other in an automorphism. There are nine graphlets of up to four nodes and fifteen automorphism orbits in these nine graphlets.

Methodology

To obtain our data, we downloaded five texts each from nine different authors from Project Gutenberg[7], which provides texts available in the public domain. The authors we chose were Henry David Thoreau, Ralph Waldo Emerson, Margaret Fuller, Leo Tolstoy, Fyodor Dostoyevsky, Gustave Flaubert, Mary Shelley, Samuel Taylor Coleridge, and Walter Scott.

Three from each of three different literary movements -- Transcendentalism, Romanticism, and Realism. The rationale behind choosing authors from the same literary movements was that authors from the same literary movement would write similarly enough to confuse the model, which would allow us to more rigorously test the model.

We then converted the texts to networks using Algorithm 1 below:

```
initialize previous word to none
initialize sets to store words and edges
for line in input do
      for word in line do
            clean word (convert to lowercase, split punctuation)
            if word is not in nodes then
                  add word to nodes
            generate edge between word and previous word
            if edge is not in edges then
                  add edge to edges
            let current word be the new previous word
```

Algorithm 1. *Converting a text to a network.*

Once the networks were generated we used the Orca software[4] to calculate the graphlet degree distribution (GDD) of the network -- that is, the number of times that each node participates in each orbit of every graphlet with four or fewer nodes. Next, we generated all possible pairs of networks and analyzed the similarity of the networks using their graphlet degree distribution agreement (GDD agreement). Each pair of texts was then ranked based on their GDD agreement with the most similar nodes ranked highly. We then created a similar ranking using the Average Clustering Coefficient Distance, where clustering coefficient is, out of all neighbors to a node, how many neighbors are connected to each other and Average Clustering Coefficient Distance is a comparison between two networks of the average clustering coefficient for every node in each network.

After we rank each pair of networks for each metric we use these rankings to generate a receiver operating characteristic (ROC) curve for the data. A ROC curve plots the true positive rate (the rate at which correct predictions are made) against the false positive rate (the rate at which incorrect predictions are made). We treat each pair of networks under a given rank threshold as a prediction that the corresponding texts were written by the same author. As we relax the rank threshold for a prediction, we calculate the true positive and false positive rates for each step to gather data points for the ROC curve. Additionally, we calculate the precision (percent of total guesses that are correct) and recall (percent of total actual pairs that we have guessed) of our predictions at each step as well, and graphed those values against each other.

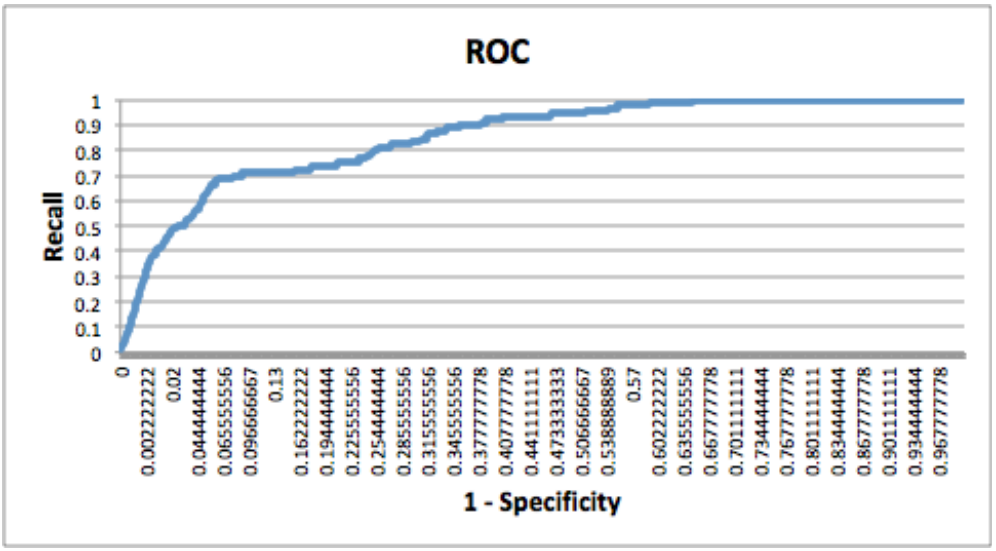

Figure 1. *ROC curve for GDD agreement*

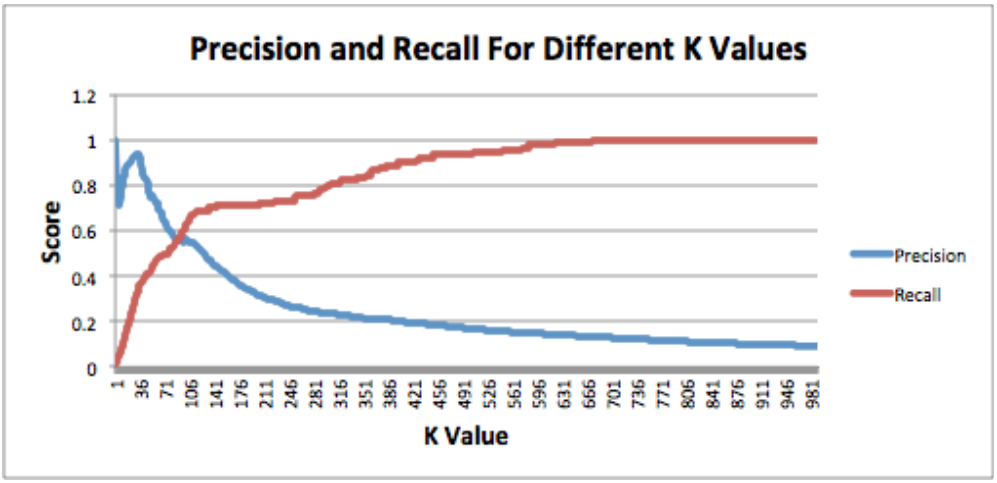

Figure 2. *Precision-recall curve for GDD agreement*

We used three different models and evaluated them. The first model uses GDD agreement, the second model uses the Average Clustering Coefficient Distance, and the final model combines the two metrics.

Results and Discussion

The first comparison metric we explored was GDD agreement. The ROC curve (Figure 1) shows that this comparison metric was effective at determining whether or not the same author wrote two given works. The Area Under the Curve (AUC) for this ROC curve was 0.8867.

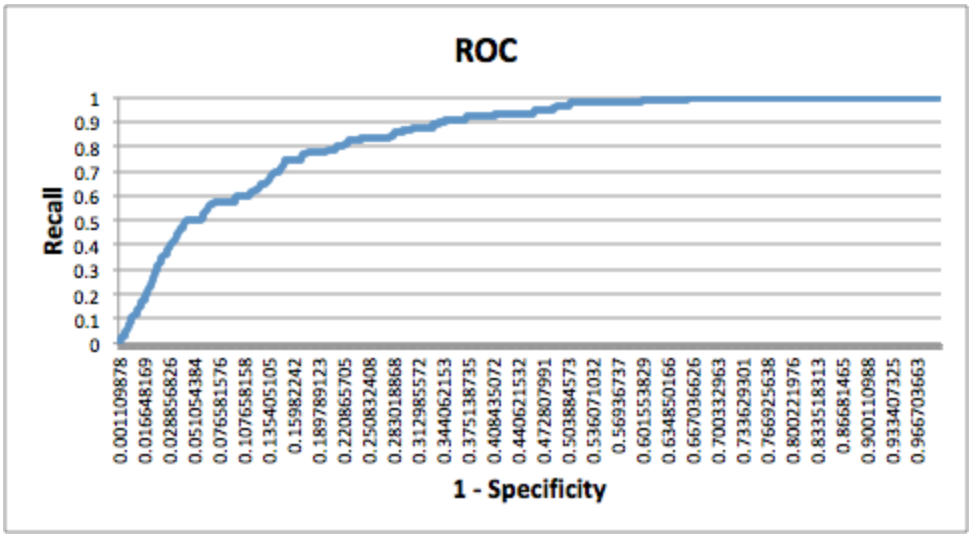

Figure 3. *ROC curve for average clustering coefficient distance*

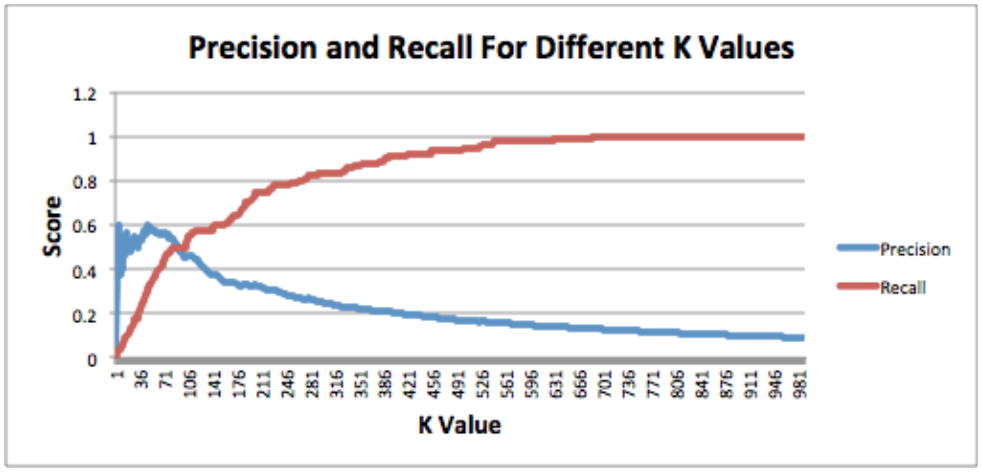

Figure 4. *Precision-recall for average clustering coefficient distance*

As the precision-recall curve for GDD agreement (Figure 2) shows, the model had a high precision for the first several predictions before it leveled off. This reflects the fact that 55 of 90 possible same author pairings (61.11%) occured in the top 10% of the 990 author combinations. Additionally, the text that was most similar to a given work was by the same author 38 out of 45 times (84.44%).

The next similarity metric we examined was clustering coefficient. We did not expect that clustering coefficient alone would result in a good model for determining similarity of works by the same author, but it vastly outperformed our expectations. The ROC curve (Figure 3) for the clustering coefficient is visibly worse than the ROC curve for the GDD agreement model. We are most interested, for the sake of this problem, with the very beginning of the curve, and the ROC curve for GDD is much steeper in its first incline than the one for the clustering coefficient. This is reflected by the fact that only 51.11% of same

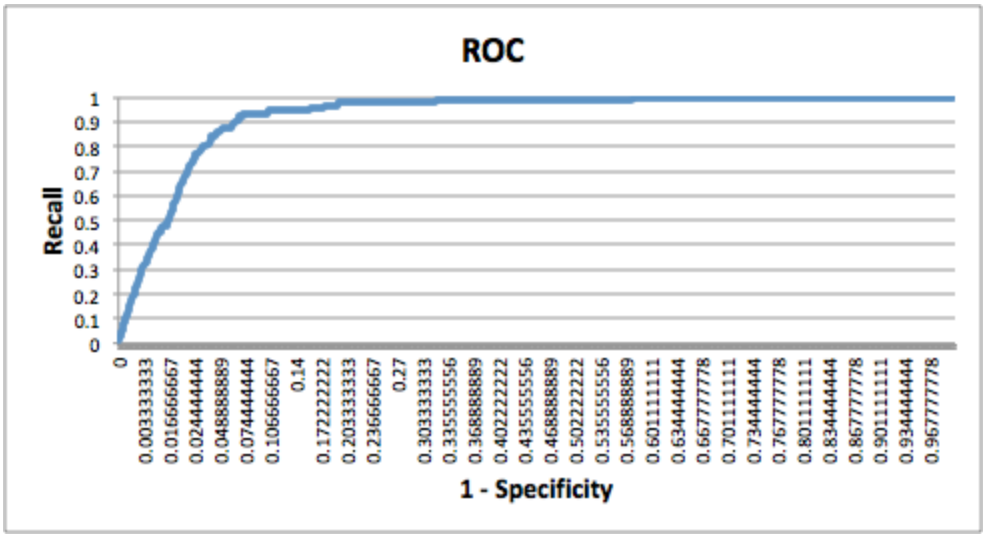

Figure 5. *ROC curve for rank metric*

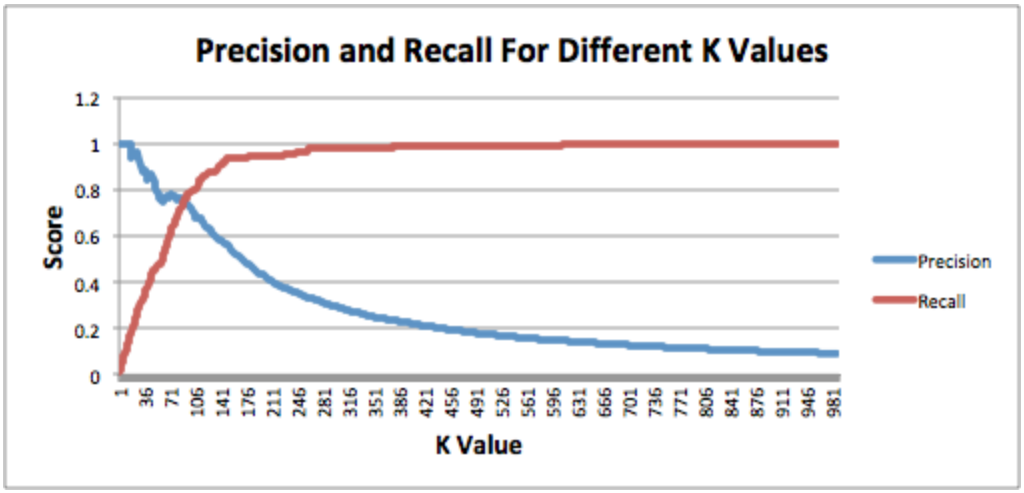

Figure 6. *Precision-recall for rank metric*

author pairings (46/90) occured in the top 10% of possible pairs, 10% less than the GDD agreement model. This is also apparent in the Precision-Recall curve (Figure 4), which shows that the first few pairs were incorrectly predicted, and the precision never really gets above 0.6 in the first 90 pairs.

Additionally, only 32 of 45 (71.11%) works were most closely related to a work by the same author, which is worse than the GDD model (84.44%). However, the AUC for the clustering coefficient was 0.8759, which is very close to the AUC for the GDD metric. This shows that the clustering coefficient model, although it performs worse than the GDD agreement model, is still a pretty reasonable model.

When taking into account how poorly the correlation coefficient similarity metric does for the top matches as shown by both the ROC and Precision-Recall curves, it is somewhat surprising that the AUC is as high as it is. We believe that this has to do with the clustering

coefficient being a stronger metric for ignoring the noise as networks become less similar. As a result, we decided for formulate our own metric by combining the similarity scores from GDD and clustering coefficient.

The combined model takes into effect how similar two works are on both metrics. Due to the ambiguous relationship between the raw values of the distance between two networks based on the two metrics, we created a model based on rank. For every network, we examine the 44 possible pairs it can make with the other networks and rank each pair by GDD and correlation coefficient separately to generate ranks. For instance, network A could be most similar to network B based on GDD and therefore B would have a GDD-rank of 1, but based on correlation coefficient, network B could be the 5th most similar network to network A, giving it a correlation coefficient rank of 5. Due to the nature of our rankings, network B's most similar network based on GDD may not be A; network A may be the 3rd most similar network to B based on GDD for a GDD-rank of 3. Therefore, to calculate the similarity rank for a network pair (A,B), we sum together four different values: (1) The GDD-rank of B for A, (2) the correlation coefficient rank of B for A, (3) the GDD-rank of A for B, and (4) the correlation coefficient rank of A for B. Therefore, the lowest, and therefore best, possible rank is 4. We expected this model to perform better than either of the previous models since it combined information from both.

The ROC curve (Figure 5) demonstrates that this model improves upon the previous models. It has a large slope at the beginning, which is reflected in the very high AUC value of 0.9682, and is evident in the Precision-Recall curve (Figure 6). We can see that the precision stays around or above 0.8 throughout the first 90 pairs.

Additionally, we see 68 of the 90 of same author pairings (75.56%) occured in the top 10% of possible pairs which is an increase from both of the previous models. However, interestingly, only 37 of the 45 (82.22%) works that are most closely related to another are works by the same author, which means it matches 1 less work than the GDD agreement model. Unlike the other metrics, here we see a decline in accuracy when using the rank metric.

This data tells us that the combined ranking model is more robust at determining two works were written by the same author. It isn't better than the GDD model in determining the best match for a network, but it is far better at determining the best author down the line, especially given more data.

Conclusion

Based on the relatively high AUC values and F1 scores for all models, the structure of an author's writing is distinguishable enough that we can determine authorship reasonably well given few metrics. Future work could involve incorporating new features into our model to improve the accuracy of our predictions. Additionally, it would be interesting to see how writers who are said to have similar, or different, writing styles compare in our framework.

Student Contributions

Sam Alptekin 30%, Sam Berning, 33.3%, Joe Kimlinger 36.6%
Generating networks - Sam B and Joe did programming, Sam A did conceptual planning
Generating metrics - Joe calculated similarity metrics for all pairs of networks
Analyzing metrics - Sam A and Sam B did the planning, programming, and analysis
Writing paper - All students worked together on this

References

1. Antiqueira, L & Pardo, T.A.S. & Nunes, Maria & Oliveira, Osvaldo. (2007). Some issues on complex networks for author characterization. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601, Nº. 36, 2007 (Ejemplar dedicado a: From Natural Language Processing to Information and Human Language Technology), pags. 51-58. 11. 10.4114/ia.v11i36.891.
2. R. Arun, V. Suresh and C. E. V. Madhavan, "Stopword Graphs and Authorship Attribution in Text Corpora," *2009 IEEE International Conference on Semantic Computing*, Berkeley, CA, 2009, pp. 192-196. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5298613
3. H. El-Fiqi, E. Petraki and H. A. Abbass, "A computational linguistic approach for the identification of translator stylometry using Arabic-English text," *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, Taipei, 2011, pp. 2039-2045. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6007535&isnumber=6007305 .
4. Shibamouli Lahiri, Rada Mihalcea, “Authorship Attribution Using Word Network Features,” *CoRR*, Volume abs/1311.2978 2013, http://arxiv.org/abs/1311.2978
5. Zhenhao Ge, Yufang Sun, Mark J. T. Smith, “Authorship Attribution Using a Neural Network Language Model”, *CoRR*, Volume abs/1602.05292, 2016, http://arxiv.org/abs/1602.05292.
6. Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, Nataša Pržulj, “Revealing the Hidden Language of Complex Networks,” *Scientific Reports*, Volume 4, 1 April 2014, https://doi.org/10.1038/srep04547
7. Project Gutenberg: https://www.gutenberg.org/.
8. Tomaž Hočevar, Janez Demšar; A combinatorial approach to graphlet counting, *Bioinformatics*, Volume 30, Issue 4, 15 February 2014, https://doi.org/10.1093/bioinformatics/btt717.