RELATE Final Report
Lucas Jarmin & Benjamin Malley
Networks: Theory & Application

Introduction and Background

In this paper, we examine first a network of 88 texts retrieved from Project Gutenberg linked by their superficial textual characteristics and then a network of the same texts linked by customer purchasing data retrieved from Amazon.com. This project born both out of some measure of frustration with the prevailing method of determining user preferences employed by Amazon.com among other sites as well as a desire to determine if valuable connections between works can be made using superficial characteristics. The goal of the project is to determine interesting relationships between works using network analysis of text characteristics and to compare those relationships to the Amazon.com customer preference data.

Project Gutenberg is a digital library of works mostly in the public domain. As such, most of the works available on Project Gutenberg are from before the mid-1920s when new copyright law was enacted and some are significantly older than that. For example, we used the Odyssey by Homer, which is from the 8th century BC. Our most recent text, *Down and Out in the Magic Kingdom* by Cory Doctorow was published in 2003. A complete list of texts used is available as an appendix to this report.

Because we are seeking relationships between texts based on their superficial textual characteristics, we have created a network where the nodes represent texts and edges connecting them are representative of shared textual characteristics as explained in the next section. We compare this network to a network made from Amazon.com customer purchasing data where the nodes are also individual texts and edges were made when Amazon customers who bought one text also bought another.

Information on shallow textual analysis was gained from several sources. In *Linguistic correlates of style: authorship classification with deep linguistic analysis features*, Gamon discusses determining authorship using a combination of the shallow textual analysis performed here as well as "deep" analysis using context-free grammars. Gamon's approach correctly identifies the author of a text a maximum of 81.65% of the time using only shallow characteristics and a maximum of 85% percent of the time using both shallow and deep characteristic. Our approach, which does not leverage deep analysis, uses network properties to investigate connections between texts using shallow analysis.
We gained further insight into the field from *Learning to classify documents according to genre* by Finn and Kushmerick in which the authors indicated ‘keywords,’ as a potential method of identifying the ‘genre’ of documents. Out of 152 characteristics of documents obtained in that research, the majority of the characteristics were function words as a percentage of overall words (Finn and Kushmerick).

**Methodology**

While our initial project plan called for a large number of texts to be used, in the interest of time, this was scaled down to just 88 texts. These texts were the most popular works on Project Gutenberg available in the English language. We removed extraneous information from the documents, including Project Gutenberg boilerplate and licensing information, tables of contents and introductions by third parties. We feel that more complete analysis could be made using a significantly larger corpus of texts, which would require a method of scaling the preparation work we performed on our texts.

Once our corpus was gathered, we performed "shallow" textual analysis of the texts using the method described by Finn and Kushmerick. Specifically, we gathered the following characteristics of each text:

- the occurrence of specific common words (i.e., the, and & to) as a percentage of total words
- the average number of words per sentence
- the average number of characters per word
- the Gunning Fog index of the text
- the total number of words (used for finding percentages of common words)

For the occurrence of common words as a percentage of total words of the document, we chose "the," "and," and "to" for several reasons. First, these are all function words, which create grammatical relationships between other words, and therefore are used by most authors regardless of word choice. Thus, these words had the consistently highest percentage of words in our sample of texts. Second, there is historical precedent for textual analysis using function words identified in Mosteller & Wallace's *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* (Gamon, 2004). Generally, our function words had a frequency of between 1% and 10% of the total words. For example, the range of values for use of the word "and" as a percentage of total words ranged from 1.729% to 9.047%, depending on the text. From these, we
created a composite "score" based on the occurrence of these words as follows:

\[
(\text{occurrence of and} / \text{total words}) + (\text{occurrence of the} / \text{total words}) + \\
(\text{occurrence of to} / \text{total words})
\]

The Gunning Fog index is a measure of the readability of the document, developed by Robert Gunning in 1952. The index is an estimate of the minimum grade of formal education needed in order to understand the document, which it predicts roughly 80% of the time. The Gunning Fog index of a document can be calculated using the following algorithm:

1. Divide the number of words by the number of sentences to find the average sentence length (asl).
2. Count the number of words with three or more syllables, not including proper nouns, compound words, or common suffixes such as -es, -ed, or -ing as a syllable (these are "complex words").
3. Find the percentage of complex words to total words in the document.
4. Add the average sentence length and the percentage of complex words (ex., +13.37%, not simply + 0.1337).
5. Multiply the result by 0.4 for index.

The range of Gunning Fox scores for our texts was 3.00 for Hedda Gabler to 15.28 for the Metaphysical Element of Ethics. The mean Gunning Fox index was 7.66 and the median was 7.42.

For this project, shallow text characteristics were obtained from two freely available analysis engines, Topicalizer, the free web-based text analysis tool available at topicalizer.com and available at textalyser.net. Textalyser is a web-only service. Topicalizer provides a web-only interface as well as a Python back-end and API. The service will provide all of the shallow text characteristics we used as well as attempt summarization of text. Because our documents were mostly fictional, we did not find the summarization tool to return particularly useful results. Also, it is important to note that because Topicalizer uses language specific characteristics in order to operate most effectively, we coded all of our documents as "English."

From this data we constructed a number of graphs using each text as a node and making an edge between nodes where the shallow text characteristics of the nodes were within a certain threshold of each other. At either extreme, the graph was either a complete graph or a completely unconnected graph.
Of these graphs, a graph that had edges between texts that met (that is, are within a certain threshold of each other) on any of our characteristics gave us our most significant results and thus is the focus in our results section.

In accordance with our goals, we constructed a second network by retrieving purchasing data available at Amazon.com. For this network, we took each of our 88 texts from the Project Gutenberg network and downloaded customer purchasing data from Amazon.com for those titles. Specifically, we gathered the top results in the "Customers Who Bought This Item Also Bought" (herein: customer purchasing data) section of the Amazon page for each item.

To accomplish this, we wrote a Perl script utilizing the WWW::Mechanize module, which is designed to automate web browsing tasks such as clicking buttons and inputting data into forms. This script iterated through our list of texts, searching for each title on Amazon.com, clicking through the first match on the page in the "book" category (specifically those items containing id="book" within the HTML tag for the link) and retrieving the customer purchasing data from the resulting page.

We felt that given the "classic" nature of many of texts we were using, that the Amazon.com customer purchasing data would not provide very useful information. In particular, our thought was that many of these books would be purchased for class studies, and customer data would merely represent commonly assigned texts for class. We also noted several difficulties with our approach. First, because many of our texts are not under copyright, are were available under a number of different editions. Under subsequent examination, we determined that different editions of the same text often had different customer purchasing data. It is our stance that this underscores some of the problems with the Amazon approach to generating "customer preference". Second, customer purchasing data often contained titles not in our original corpus. Because they added no value to our network, we discarded these titles, except in cases where a title occurred repeatedly in several customer purchasing data lists.

A specific instance of problems encountered during our Amazon data retrieval is what we dubbed the “Twilight effect.” Twilight is a new and popular series of young adult books, as well as a recent popular movie. It appeared in customer purchasing data for many of our titles, especially for those in more recent editions. As an example, of the five works purchased by people that bought a 2004 edition of Wuthering Heights, four were Twilight novels. However, Twilight did not appear in the customer purchasing data for an older hardcover edition. We feel that this shows a significant drawback of the Amazon approach to relating works.
Results

In order to get real results from our network, it was necessary to explore thresholding of our edges. This originated from the problem that the original network was simply too dense (see Fig. 1).

An edge is created in the original graph any time that two nodes “meet” on any of the six characteristics as described above. To reduce the density of the graph, we increased the amount of characteristics that two nodes had to meet on in order to have an edge drawn. This eventually produced a good density graph of at a threshold of more than 3 of the characteristics.

Beyond this thresholding we had to discover what level nodes had to be at to be considered “meeting.” Each node had a decimal value between 0 and 1 for the three function words, a value somewhere between 5 and 25 for words per sentence, a value between 2 and 8 for characters per word, and a value between 3 and 15 for the Gunning-Fog index. Due to variance in the specificity of the values we received from our web-based textual analysis tools, the three function word values are to 5 significant digits, while the other three characteristics only have decimal significant values to 2 significant digits. This does not harm the validity of the results, because the significant digits are the same for all texts within a characteristic.
We determined what values to threshold at based on our own subjective perception. We looked at the range of values for a characteristic and decided on a value that ensures that some edges are created while not allowing too many edges. This resulted in the following thresholds:

- % of and: 0.01
- % of the: 0.01
- % of to: 0.005
- words/sent: 1
- char/word: 0.25
- Gunning-Fog: 0.75

All thresholding was accomplished by creating a Python script and running the script in GUESS to both create and threshold the edges.

After thresholding, the network was a lot more manageable and appeared to be developing real structure (see Fig. 2).

At this point, the density of the network is 0.1510847. While this is still generally high, it is much more manageable than the original network. It is also of note that at the threshold-complete level, the graph is not complete any more, as there are two outliers. These two outliers are Marx and Engel’s “Communist Manifesto” and Sir Thomas Malory’s *Le Mort d’Arthur*. These are, indeed, unusual works—“Communist Manifesto” is written more as a speech and less as a work of literature; *Le Mort d’Arthur* is written
in poetic form which may have thrown off some of the metrics such as words per sentence, and its Middle English diction may also have affected the analysis.

The betweenness of the nodes in the graph can be seen expressed in Figure 3. Nodes have been resized in GUESS according to betweenness. As expected, many of these nodes are in the large center cluster.

What is intriguing is validity of betweenness in this context. As can be seen in the upper right of the giant component, Beyond Good and Evil has rather high betweenness resulting from its connecting two near isolates to the giant component. What, though, does betweenness mean here? What does it mean if Between Good and Evil (BG&E) is between Don Quixote (DQ) and The Odyssey (TO)? What this means is that DQ is unlike most of the works in the group and only finds connection to BG&E and one other work. BG&E, meanwhile, is also unusual and is only connected to DQ and TO. Is it really relevant that there is a one-neighbor-away connection between DQ and TO? Would a person start with TO, then move to BG&E and therefore be more ready to move to DQ because of this?

What may be more useful is to visualize the degree of nodes (see Fig. 4). As also expected, the high degree nodes all cluster to the middle where there are many complete triads.
What this shows is that there are many works of the “classic” genre that are all linguistically similar to one another. While there are some works on the periphery that are not linked to many other works, a large majority of the nodes are interrelated to a large percentage of the other works. This means that most works, regardless of year published, follows similar stylistic patterns. This can be seen even better in Fig. 5. In Figure 5, nodes are again sized by degree. This time, however, they are colored based on date of writing. Purple is anything before 1700, blue is the 1700s, red is the 1800s, green is the 1900s, and orange is the 2000s. As Figure 5 shows, the middle cluster is not one solid time period. Most nodes are red (1800s), but this stems largely from the fact that the set of 88 works skews largely toward the 1800s because many works Western society consider “classics” were written in this time period. The colors are generally well distributed. One outlying sector is the bottom left with the cluster of purples; this is the Shakespeare subgraph. These works may be less integrated than others possibly because of their language use, but also possibly in part because of similar reasons mentioned about Le Mort d’Arthur above, namely that the formatting of the work may have affected the analysis since Shakespeare’s works are plays and formatted as such with speakers’ names and a colon preceding their lines.
After this analysis, we decided to see if any well-defined communities could be resolved out of the network. For this task, we employed Lada Adamic's community finding GUESS script, which is an implementation of the Girvan-Newman betweenness clustering algorithm. What this script does is remove the edge with the highest betweenness consecutively until a community is discovered. While this may then fall prey to the same concerns about the use of betweenness to rank nodes, we find these communities are still useful. We used Lada's algorithm continually until we saw clearly defined and not too cluttered clusters. The result can be seen in Figure 6.

While there are many isolates, there are a good number of clearly defined and useful groups. For example, a number of Jane Austen works formed a cluster, as did Shakespeare’s works. Our outlier works such as the three recent science fiction pieces and the ancient fiction works by Homer and others constitute the isolates. Most of what remains in the clusters is work from the 1800s, as seen in Figure 7 where red nodes are works from the 1800s (and light blue is any other time).
Interestingly, these results show a multitude of cross-author communities. As Figure 8 shows, a large majority of community edges were between different authors. In this figure, an edge is colored red if it spans two authors, and the edge remains black if it is an edge between the same author.
We were surprised by how often works from different authors ended up being in a community together. We assumed going into this research project that an author’s works would tend to form clusters because of inherent writing similarities. Clearly, this is not always the case.

This finding is further supported by Figure 9. Figure 9 is the original graph after thresholds, but with white lines indicating where an edge would be expected because the two nodes have the same author. The dark blue edges represent all edges present in the original graph. Thus, the graph shows all of the missing edges between an author’s works.

This graph generated interesting results about works by the same author. For example, there is a node on the left side of the giant component that has four white lines going of it and only two real (blue) edges. This is Charles Dickens’ Great Expectations. While the other 5 of his books included in the graph form a near complete clique, Great Expectations is not linked to any of Dickens’ other works. For some reason, Great Expectations is not stylistically (in the sense of our metrics) similar to any of his other major works. We also discovered through research that this book was one of the last major book he wrote, and is considered by many his best. Was there a tangible shift in writing style?
The last analysis we completed was a shallow investigation of our network against the current network of recommended books on Amazon. When on an Amazon page, Amazon recommends works that it considers related in a section entitled “Customers Who Bought This Item Also Bought.” As discussed in the Methodology section above, we ran into many flaws with Amazon’s systems including the problem of the Twilight series books taking over the recommendation system because so many people bought the temporarily popular but not necessarily related or relevant books. We created a network where we crawled the recommendation data and extracted only links that were between books in our network (and we included Twilight to show its effects).

The result is Figure 10. As is visually obvious, the graph is much more sparse than our data’s graph, with very few edges between many books. But some works, however, seem to have an almost exaggerated degree. In particular, Pride and Prejudice, Treasure Island, and Twilight. These were very peculiar results, as neither of the two works that we also had in our graph were significant. Amazon data, however, skews these works as extremely relevant and recommendation-worthy.
One potential explanation for these skewed results from Amazon is that many students shop on Amazon for school books. Nearly all of the books in our study were books that would be read in common English classrooms, so many of these books may be bought together not because they are related or similar, but only because a teacher assigns them both for a common year. This data then skews the Amazon results and makes people looking at these recommendations think that two works are related or similar when they are not.

Conclusions

Most of our time and work was spent in figuring out how to collect the data, correctly parse the data, and threshold the data. After this was complete, we did gain some very interesting results.
The community clusters show real promise, especially because many of the edges within the clusters are cross-author edges. Also very interesting was the results showing the missing same-author edges, like the Great Expectations mystery.

While we feel that this is only the tip of the iceberg in exploring stylistic book comparisons, the research done here is a great place to start for future explorations of ways to improve recommendations for discovering related works of literature.

**Works Cited**

Aidan Finn and Nicholas Kushmerick. "Learning to classify documents according to genre." In IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis (2003).