8-2012

Significant Themes in 19th-Century Literature

Matthew L. Jockers
University of Nebraska-Lincoln, matthew.jockers@wsu.edu

David Mimno
Princeton University, david.mimno@gmail.com

Authors: Matthew L. Jockers and David Mimno
Title: Significant Themes in 19th-Century Literature

Abstract:

External factors such as author gender, author nationality, and date of publication affect both the choice of literary themes in novels and the expression of those themes, but the extent of this association is difficult to quantify. In this work, we apply statistical methods to identify and extract hundreds of "topics" from a corpus of 3,346 works of 19th-century British, Irish, and American fiction. We use these topics as a measurable, data-driven proxy for literary themes. External factors may predict fluctuations in the use of themes and the individual word choices within themes. We use topics to measure the evidence for these associations and whether that evidence is statistically significant.[1]

Introduction:

It is commonly assumed that novels contain themes. We further assume that the cultural/historical environment of the author plays a role in determining the choice and relative use of different themes.[2] If we can understand what factors influence an author's choice of themes, we will better understand both the novels themselves and the broader context of literary history in which these works are published. But what is a theme, and how do we decide whether a theme exists in a given volume and if so to what extent? Not only are traditional methods prone to individual biases and oversimplification (as in "this book is all about religion") but even in the best case they are limited by the number of novels a scholar can read and are unable to account for a theme's recurrence and prominence over time, across genres, in different national or ethnic milieus. Work, therefore, tends to focus on a standard canon of several hundred books at best, and several thousands texts -- the vast majority of the available archive -- are simply ignored.

To address such problems of scale, humanities scholars have begun using statistical topic models to identify and measure themes in large text collections.[3] Because these models do not require annotated training data and do not attempt to analyze linguistic structures, they are simple to run and robust to variation in language and data quality. Topic models are powerful and scale to large data sets, but their ease of use can be

[1] Parts of this research were supported by a grant from the Mellon Foundation as part of the SEASR Services project. The authors thank Loretta Auvil and Boris Capitanu of the National Center for Supercomputing Applications for their work on the SEASR project where much of this data was initially processed.

[2] A similar point is suggested by Osip Brik's 1929 work *Teaching Writers*: "In every period there is a certain number of artistic methods and devices available for creative use. Changing these methods and devices is not a matter of the individual author's volition, but is the result of the evolution of artistic creativity.

[3] See for example,
Blevins, Cameron. "Topic Modeling Martha Ballard's Diary." *Historying*. April 1, 2010, <http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/>
Block, Sharon. "Doing More with Digitization." *Common Place*, Vol 6, No. 2, 2006. <http://www.common-place.org/vol-06/no-02/tales/>

deceptive. Without testing methodologies, scholars risk reporting results that are statistically insignificant. We present, therefore, a statistical testing methodology for measuring the association between metadata (e.g. publication date, author gender, author nationality) and topics.[4] We provide estimates of the statistical strength of our results as a way to quantitatively contextualize our more qualitative interpretations of what these associations mean in the context of 19th-century literature.[5]

Background:

This work interrogates the results of a thematic modeling of 3,346 works of 19th-century British, American, and Irish fiction undertaken in chapter eight *Macroanalysis: Digital Methods and Literary History* (Jockers 2013).[6] Our objective here is to narrow the focus and assess the extent to which author gender is, or is not, a predictor of thematic attention and to address questions related to how male or female authors write about such themes as "religion," "war," and "fashion." We also investigate the extent to which the topical distributions--the thematic preferences or "general tendencies"--of known male and female authors can be employed to predict the gender of pseudonymous/ anonymous writers. Along similar lines, we compare and contrast the thematic tendencies of anonymous writers with those of known authors in order to explore the extent to which anonymity is itself a predictor of topical distributions. Here we investigate the hypothesis that works written anonymously will tend toward more controversial themes of a religious, political, and/or nationalistic nature.

Methodology:

The simplest approach to statistical analysis of literature is to count words. But if we use word counts to draw conclusions about the thematic tendencies within different classes of authors, we risk making mistakes because words are sparse, variable, and ambiguous. Sparsity arises because vocabularies are large and most words occur infrequently. Variability contributes to this problem; authors often have a choice of several synonyms. In order to make claims about thematic tendencies, we would have to summarize the results of hundreds of word/gender associations, most of which would be poorly estimated due to small sample sizes and because there is no clear way to decide which words out of thousands of words should be considered. Ambiguity adds further complications: if we count the occurrence of a single word, we may inadvertently conflate multiple meanings of that word (i.e. "bank" as a financial institution and "bank" as the side of a river).

[4] Our analysis here is limited to gender, but data about thematic use separated by author-nationality can be found at <http://www.matthewjockers.net/macroanalysisbook/macro-themes/>

[5] Throughout this paper we use the terms "theme" and "topic" as proxies for the same general concept: namely a type of literary content that is semantically unified and recurs with some degree of frequency or regularity throughout and across a corpus. Based on this definition, the word clusters discussed here are self-evidently thematic in nature and even while the matter of what constitutes a theme or topic is a broad area in which some things are black, some white, and some grey, most readers will recognize the larger thematic categories to which these word distributions belong.

[6] Chapter 8 provides detailed overview of the corpus, the methodology, and the modeling parameters.

Statistical topic models, such as latent Dirichlet allocation (LDA), use contextual clues to group related words and distinguish between uses of ambiguous words.[7] They reduce the dimensionality of a corpus to several hundred clusters or *topics*, represented as distributions over the full vocabulary. This level of complexity is rich enough to express much of the variability of the corpus, but small enough to be browsable by humans. Because topics group many words together, they are less vulnerable to small sample sizes than individual word frequencies.

We use the Mallet implementation of Latent Dirichlet Allocation to extract thematic information from a corpus of 19-century novels. When linked with human-coded metadata, topical data offer a way of exploring macro scale thematic trending and tendencies in a faceted manner. Whether these macro scale observations about central tendencies in the usage of particular themes within a particular facet are statistically meaningful and warrant broad conclusions about larger literary historical trends in the 19th century is the open question that this research explores.

Corpus:

Our work begins with a corpus of 3,346 works of fiction from the United States and Great Britain (including at that time Ireland, Scotland, and Wales). The publications dates for these works span a period from 1750-1899 with the major concentrations in the period from 1800 to 1899 (Figure 1).

[7] Blei, David, Ng, Andrew, and Jordan, Michael. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*. 2003.

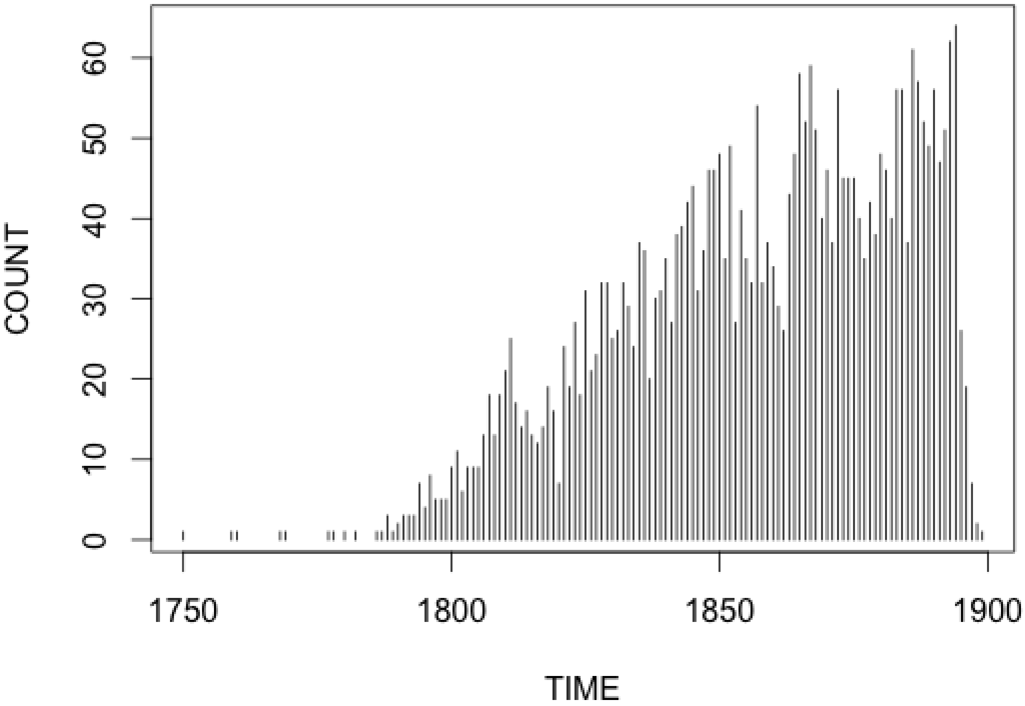


Figure 1: Count of books in corpus by year. Year-to-year variability is high, but there is a clear rising trend over time.

Preprocessing:

1. *Stopword Removal.* Although training a topic model is completely automatic, we can control the output of the model by pre-processing the corpus. It is standard practice to remove common syntactical "stopwords" such as *the*, *and*, and *of*. These words occur so frequently, and with such regularity in all documents, that they overwhelm topical variability. It was determined that standard stopword removal was not sufficient for this corpus of book-length fiction. Novels have properties that differ from the scientific abstracts and news articles typically used in topic modeling. Most obviously different are the use of character names in fiction. Where different news articles will frequently mention the same person, character names are specific only to the text in which they occur.[8] If they are not removed, some topics trained on these novels will tend to form around common names used for totally different characters in totally different books. To avoid this situation, we use the Stanford Named Entity Recognition (NER) software

[8] With the obvious exception being sequels. There are no sequels in this corpus; multivolume novels have been “stitched” together into single documents.

package to pre-process the texts and identify the named entities.[9] Character and personal names identified by the NER software package along with a list of common given names were added to the stop words list along with marks of punctuation and numbers. The stopwords list ultimately totaled 5,631 distinct types.[10]

2. *Segmentation*. Topic models employ what is called a "bag of words" approach to text analysis. The words in a text are treated as if they had been tossed together into an imaginary bag. Because the model looks at which words tend to co-occur, the bigger the bag the more likely it is to have a set of words appear as collocates. For this reason, a modeling of full novels tends to result in topics of a broad and general nature. For topic modeling purposes, the unit of analysis should be a segment of text that is large enough to measure word co-occurrences but small enough that it can reasonably be assumed to contain a small number of themes. Full texts were too large to discover patterns of word use: topics derived from full novels were vague and "washed out," so we split each text into segments of approximately 1000 words, with breaks at the nearest sentence boundary. This division ultimately resulted in a set of highly interpretable and focused topics.[11]

3. *Nouns*. After a series of experiments, it was determined that the thematic information in this corpus could be best captured by modeling only the remaining nouns. Using the Stanford POS tagger, each word in each segment was marked up with a part of speech indicator and all but the nouns were removed.[12]

4. *Modeling*. The Mallet Topic Modeling toolkit was employed to model the corpus and to extract 500 latent topics from a segmented corpus of 631,577 document chunks.[13]

Analysis and Observations:

We begin by asking whether the different thematic values observed for each gender facets (i.e. male, female, unknown) are an accurate representation of the relative usage of those themes in those facet categories. In other words, are women really using certain themes more than men? We can provide a simple answer to this question

[9] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

[10] The full list can be found at http://www.matthewjockers.net/expanded-stopwords-list/

[11] A labeled list of the 500 topics harvested for this experiment, along with a set of graphs charting their distributions across time, gender, and nationality can be viewed at http://www.matthewjockers.net/macroanalysisbook/macro-themes/

[12] Depending on what one wishes to analyze in the topics, the exclusion of certain word classes could be viewed as a controversial step. By eliminating adjectives, for example, the resulting topics may lack information expressive of attitudes or sentiments. Likewise, the elimination of verbs may result in topics that fail to capture the actions associated with the novels. The noun-based approach used here is specific to the type of thematic results desired and is not suggested as a one-size fits all approach.

[13] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

by counting the proportion of all words written by male authors that are assigned to the topic and the proportion of words written by female authors that are assigned to the same topic. A difference between those two proportions is evidence that men and women use this topic at different rates. But how strong is this evidence? It is, indeed, unlikely that we would measure exactly the same proportion even if there were no underlying difference in topic use. Figure 2 offers a corpus wide view of the mean proportion of usage of a theme labeled “female fashion.” Figure 3 provides a “word cloud” representation of the terms the model found for this topic.

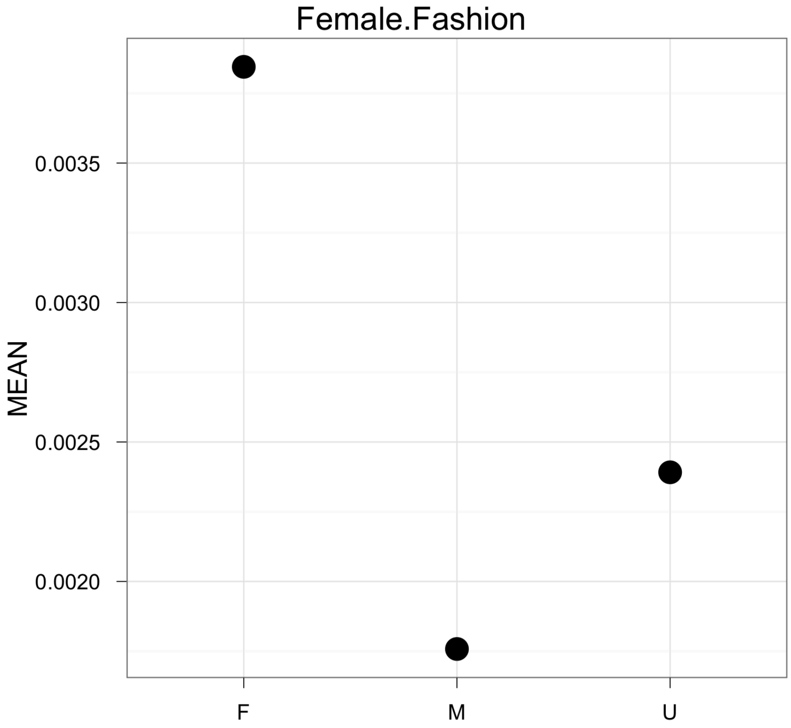


Figure 2: Mean proportion of “female fashion” theme distributed by gender

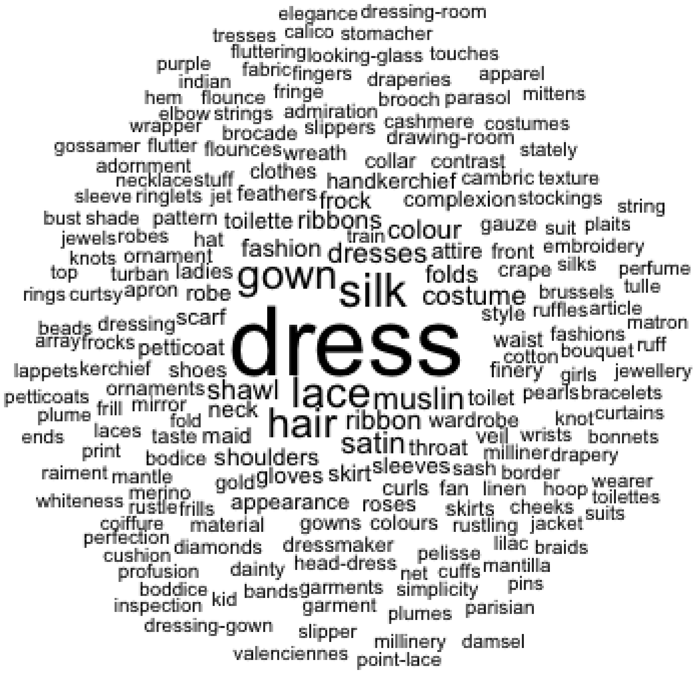


Figure 3: Word cloud of topic labeled “Female Fashion.”

Figure 2 suggests that female authors are more than twice as likely to write about women’s clothing or “fashion” than men. However, we can only say that an observed difference in proportions is significant if we know something about the range of proportions we might expect if there were no real distinction between two groups of authors. This assumption--that there is no distinction between groups--is called the *null hypothesis*. Estimating the probability of an observed difference in proportion under the null hypothesis is the central task of statistical testing. We cannot recognize pattern without an understanding of randomness.

To simulate the distribution of proportions under the null hypothesis we use a randomization test.[14] Our randomization test has two parts. First, it has a function of interest: in this case, the difference between topic proportions. Second, it has a rule for generating a fake dataset under the null hypothesis. Our real data set consists of a table with two columns. Each row represents one novel. The first column is the gender of the novel's author. The second column is the proportion of the words in the novel that are assigned to a particular topic *k*. We generate a fake data set by holding the second column fixed and randomly shuffling the values in the first column. In other words, we create a hypothetical world, in which *Persuasion* might have been written by a man and *Bleak House* by a woman, but the ratio of male- to female-authored works remains constant. For each randomized data set, we calculate the average proportion of topic *k* in each group by dividing the rows into those with *Male* in the first column those with *Female*, and calculating the average of the values in the second column for both groups separately.

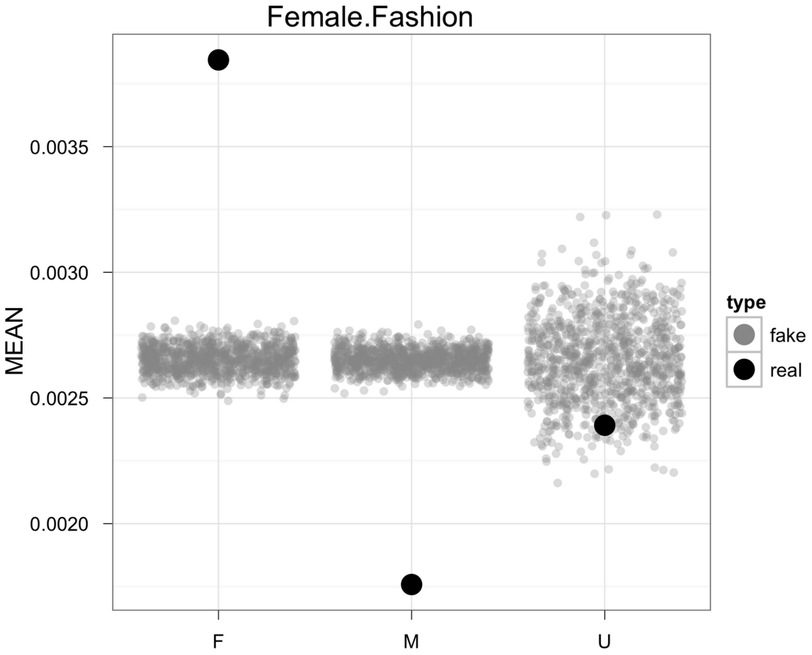


Figure 4: Permutation plot of Female Fashion

Figure 4 provides a graphical representation of running such a test over the data pertaining to the theme of “female fashion.” The “real” data is represented by three large black dots, as before. For each replication, we sample an alternative universe of author genders and add three small gray dots (one in each vertical column of the graph), representing where the mean values would be in that universe. We have

[14] Fisher, Ronald. *The Design of Experiments*. Oliver and Boyd, 1935. We use a finite approximation to a permutation test, as presented in Mark D. Smucker, James Allan, and Ben Carterette. “A Comparison of Statistical Significance Tests for Information Retrieval Evaluation.” Conference on Information and Knowledge Management, Lisboa, Portugal, 2007.

added random "jitter" to the position of the gray dots on the x-axis in order to reduce overlapping of points and make the plot easier to read.

The dense regions of gray dots indicate where we would expect the observed mean proportion of this topic to be found if there were no consistent difference in its use. As we would expect, the region of highest density is around 0.00265 for all three author categories, which is the overall mean proportion of this topic in the corpus. The difference between categories is the variability: the dense region is smallest for male authors, slightly larger for female authors, and largest for unknown authors. This difference in density is the result of sample size. Male-authored volumes make up the largest category, there are fewer female authored-volumes, and only 145 unknown volumes. With fewer novels, more extreme values are more likely.

Figure 4 demonstrates that the observed mean proportions shown in figure 1 are a reliable representation of the extent to which the theme of female fashion is "gendered" in this corpus. The difference we measure is the result of a consistent pattern in use, not random bad luck. The "real" values for the male and female usage proportions are far outside the "expected" range of values represented by the fake data. If the results plotted in figure 1 were the result of mere chance, we would expect to see the real data plot somewhere inside the cluster of fake data. Indeed, that is what we observe for the unknown category, which we believe to be a mix of male and female authors.[15] In this case, we can visually determine an approximate *p*-value: the proportion of gray dots that are below the observed value is 60 out of 1000, or 0.06. In the case of the male and female categories, all replications fall on one side or the other of the real values, so a *p*-value is not informative. We can instead calculate the mean and standard deviation of the replicated values. The observed values for the male and female categories are 21 and 22 standard deviations below and above their replicated means, respectively. In other words, the real values are far beyond what we might expect by mere chance.

Randomized permutation tests tell us whether there is evidence for a difference in the use of topics *between* categories. If the observed value (the large black dot) falls outside the band of gray dots, we may wish to pursue a closer reading and deeper interpretation of the results. Before that, however, we might also want to know how confident we can be that the observed mean of novels in a category actually represents the category. We would like to measure the variability of use of a topic *within* a category. Is the mean topic use of male-authored novels definitely 0.00158, or could it just as easily have been 0.00211 or 0.00136? We don't expect all the novels to contain exactly the same proportion of any one topic, but if a small number of novels contain very large proportions, while others have none, they could have a disproportionate effect on our results, skewing the mean in ways that misrepresent the realities of the distribution: if one or two of those outlier novels were left out, the mean proportion would change greatly.

[15] A discussion of the unknown texts follows below.

A technique that estimates the variance of a function of a data set is the bootstrap.[16] Similar to a randomization test, the bootstrap works by creating fake data sets based on the real data set, and saving values of our function for each fake data set. Unlike the randomization test, we sample *with* replacement. In the example above, the data set is the topic use proportions of all the male-authored novels, and the function is the mean of those proportions. Consider a bookshelf with 100 novels. We select one at random, write down its proportion of topic *k*, and return it to the shelf. We do this 99 more times. It's possible --- in fact, almost certain --- that some books are selected more than once, and others not at all. Once we are done, we calculate the mean of those 100 values, record it, and start again. In the extreme case where the topic makes up a large percentage of one novel, but occurs nowhere else, the mean of each fake data set will be highly variable. If we sample that one novel zero times, the mean will be zero. If we sample it once or twice or three times, the mean will increase proportionally. If, on the other hand, the topic is spread evenly through most of the books, sampling one book or another will make little difference.

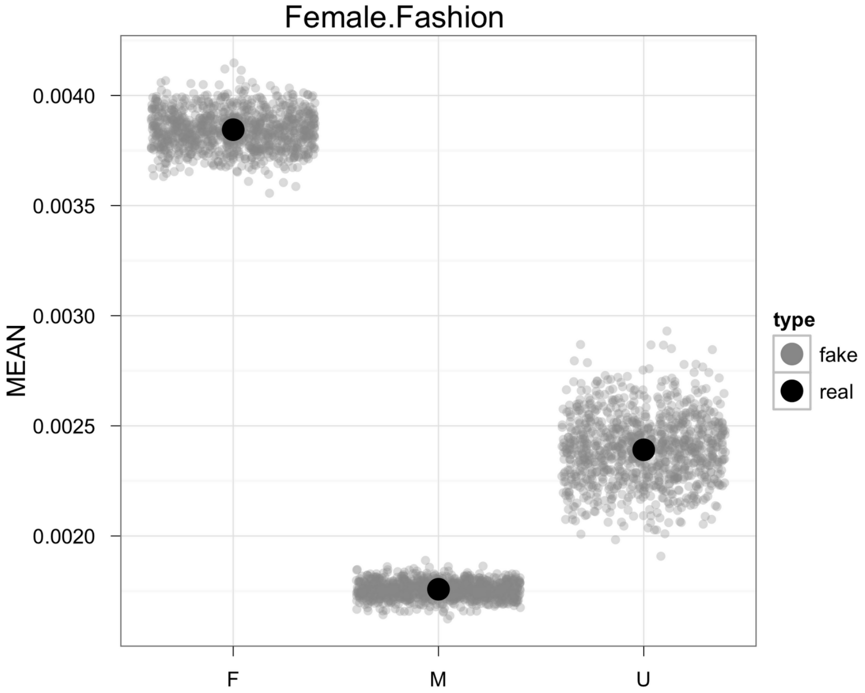


Figure 5: Bootstrap plot of "Female Fashion."

Figure 5 shows the results of a bootstrap experiment on each of the three categories for the topic "Female Fashion." Unlike the randomization test, the real values are in the exact middle of the fake data values. Because we are only sampling from within the actual category, and not mixing between categories, the regions of greatest density are not aligned at the overall mean value. Again, we find that there is relatively little variability in the male and female categories, but more variability in the smaller unknown category. For this topic, this difference can be explained by the smaller sample

[16] Efron, Bradley, and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1994.

size of this category rather than the presence of outliers (topic proportions vary between 0.0 and 0.011, with many values in between).

Figures 6-9 provide a similar series of charts for a theme labeled “Enemies.” As the figures make clear, this is a theme significantly overrepresented in male authors.

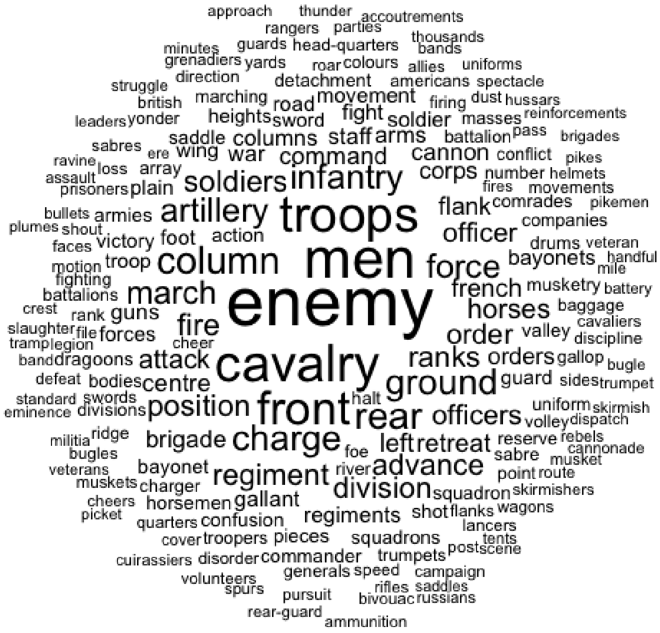


Figure 6: Enemies word cloud

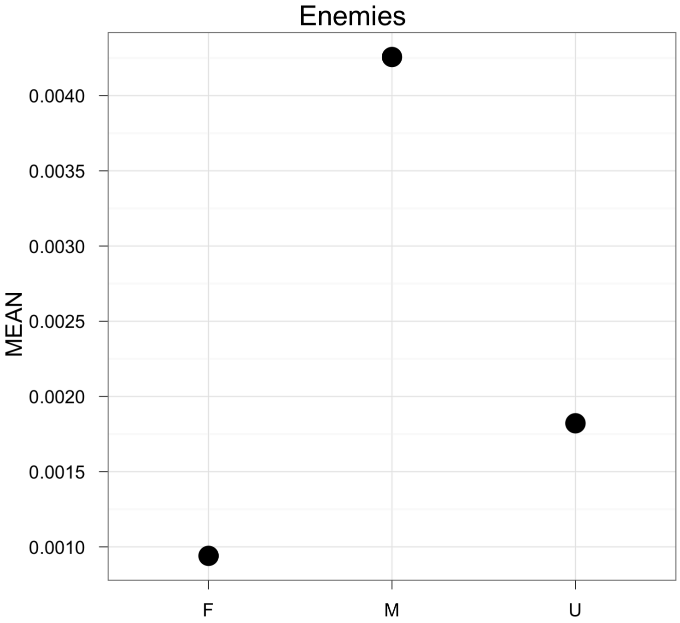


Figure 7: Gender means for Enemies

The “Enemies” topic is male dominated, but also highly variable. As with the previous example, we are not surprised by the initial result in Figure 7: we expect men to write more about war.

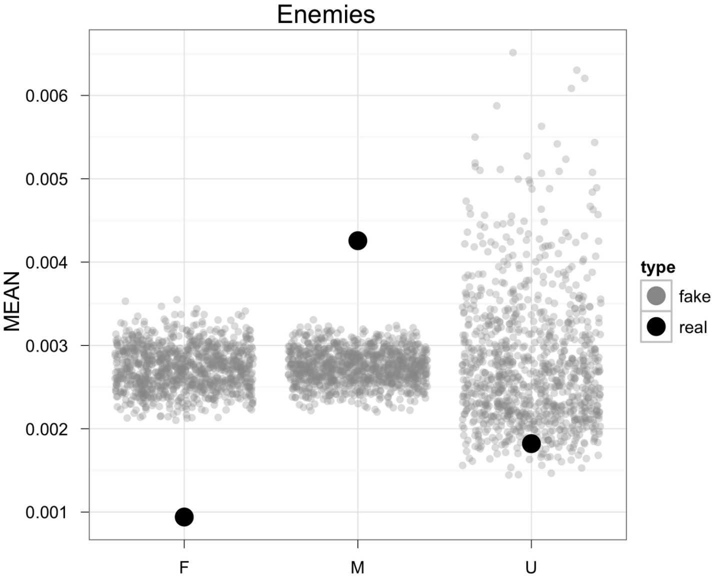


Figure 8: Permutation plot

The randomization test in Figure 8 confirms that this observed difference is unlikely to be due to chance. We have sufficient samples in both the male and female categories to get a good estimate of what the proportions should be if gender had no association to the topic. Our observations for male and female authors lie well outside those regions.

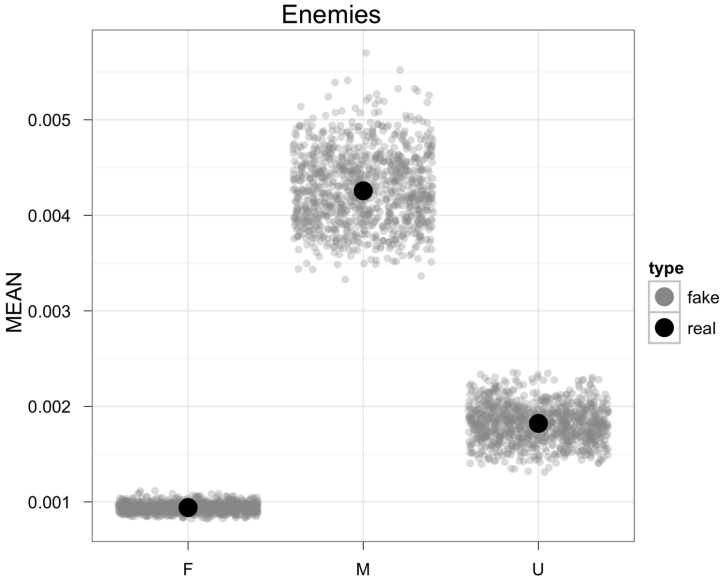


Figure 9: Bootstrap plot

The bootstrap plot in Figure 9, however, tells a slightly different story than the similar plot for "Female Fashion" in Figure 5. The range of bootstrap values for female authors is very small: almost all novels by females contain the same very small proportion of this "Enemies" topic. The range for males, in contrast, is wide--even wider than the range for the unknown authors. This variability cannot therefore be explained by sample size, since we have many more male-authored novels than unknown novels. Looking at individual novels, we find that the mean proportion of the 20 novels with the greatest representation of this topic (all male) is 0.119. The presence of these outlier novels explains the asymmetry seen for the unknown author category in Figure 8, where there are numerous values well above the mean, but few that are far below it. The class of unknowns is small enough that if one of the male "Enemies" novels is randomly assigned to the unknown class, the mean of that class will be significantly increased.

These plots provide a visualization of the different characteristics of these two topics. "Female Fashion" occurs regularly, but only in small proportions. "Enemies" occur more rarely, but can account for large proportions of the words of a novel.

Gender Classification:

It was observed that many topics in this model appeared to be "gendered" (i.e. used to greater or lesser extents by either male or female authors). To explore this observation more deeply, a classification experiment was constructed to test the degree to which the topical proportions of any given text could be employed to predict author gender. For this test, we applied the Nearest Shrunken Centroid (NSC) classifier to the data using the "pamr" (Prediction Analysis for Microarrays) package that is freely available on the R-statistical software website.[17] In order to determine the success rate of NSC at classifying texts of known author-gender, we performed cross-validation. Roughly speaking, cross-validation was performed as follows:

1. Randomly split the samples of known author-gender into two sets: a "training set," containing 2/3 of the samples and a held out "test set" containing a smaller portion, 1/3 of the samples.
2. Perform the classification training on the training set and testing on the test set.
3. Compute the error fraction from the number of misclassified test set samples.

The above process was repeated multiple times, and an average misclassification error rate of 19% was recorded across these iterations. In other words, on average, the probability of a correct classification using the topical data was seen to be 81%. In addition to providing a measure of model accuracy, the classifier also returns a ranked list of features that were found to be most useful in distinguishing between the two classes. The top twenty-five most useful features (in ranked order) are shown in Table 1. A positive value indicates overrepresentation of the topic in the given gender; a negative number under-representation.

TABLE 1

	Label	Male-Authors	Female-Authors
1	Female Fashion	-0.2015	0.2614
2	Flowers And Natural Beauty	-0.1698	0.2203
3	Tears And Sorrow	-0.1619	0.2101
4	Drawing Rooms	-0.16	0.2076
5	Drink As In Liquor And Beer And Tobacco	0.1489	-0.1932
6	Governesses And Education of Children	-0.1469	0.1906
7	Nurses For Children	-0.1467	0.1904

[17] See: Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G.(2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proceedings of the National Academy of Sciences, 99: 6567–72.
Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). "Class prediction by nearest shrunken centroids, with applications to DNA microarrays." Statistical Science, 18: 104–17. For the efficacy of the NSC algorithm to linguistic analysis, see: Jockers, M. L., and D. M. Witten. (2010). "A Comparative Study of Machine Learning Methods for Authorship Attribution." *Literary and Linguistic Computing* 25 (2):215-223 and Jockers, M. L., D. M. Witten, and C. S. Criddle. (2008). "Reassessing Authorship in the Book of Mormon Using Nearest Shrunken Centroid Classification." *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* 23:465-91.

8	Pistols And Other Guns	0.144	-0.1869
9	Children Girls	-0.1374	0.1783
10	Pity	-0.1341	0.174
11	Children	-0.1333	0.173
12	Facial Features	-0.1324	0.1719
13	Affection	-0.132	0.1712
14	Health And Illness	-0.1314	0.1705
15	Landlords	-0.1301	0.1688
16	Men With Guns	0.1298	-0.1684
17	Moments Of Confusion In Battle	0.1292	-0.1677
18	Grief And Sorrow	-0.1269	0.1646
19	Happiness	-0.1253	0.1627
20	Afternoon And Tea Time	-0.1243	0.1613
21	Swords And Weapons	0.1241	-0.161
22	Male Clothing	0.1234	-0.1601
23	Tea And Coffee	-0.1232	0.1599
24	Soldiers	0.121	-0.157
25	Dear Girls Children Creatures	-0.1198	0.1554

The thematic data derived from the topic model proved to be effective in classifying authors into gender categories and these results were consistent with similar work in gender attribution using more traditional word-frequency data (e.g. Koppel et. al.

2002).[18] Where previous work has shown that males and females have detectable stylistic affinities, this present work suggests that in addition to different stylistic habits, male and female authors, at least in this corpus of 19th-century fiction, tend to write about different things, or more precisely, to write about the same things but to very different degrees. Males are far more likely, for example, to write about guns and battles than women, and when men write about these themes they tend to do so with more intensity than female authors who, though they may touch upon these themes as a matter of course, devote far more space to, for example, the care and education of children.

It is worth emphasizing here that while these themes do tend to be highly gendered, they are not entirely so. Remember that 20% of the texts in this corpus were misclassified in the experiment, a fact that tells us that there are some male authors using what are predominantly female themes and vice versa. Nevertheless, 80% accuracy is impressive and high enough to justify taking our analysis a step further and focusing attention upon the 145 books in this corpus for which the authors and, by extension, the authors' genders, are unknown.

Employing the model discussed above, we classified the 145 anonymous texts into the categories of male or female.[19] Of the 145, 71 were classified as being most similar to the thematic habits of known female authors and 74 were identified as being most like the known males of the corpus. It was interesting, and at least a little bit surprising, to see the subset of unknown authors so evenly split between male and female. Given the time period of the corpus, we hypothesized that women were probably more likely than men to write anonymously, and so we expected to see many more of these unknown works classed as being of likely female authorship. We will, of course, never know the truth. It is entirely possible, if very unlikely, that every one of these unknown authors was in fact a woman. Indeed, it is possible that these were all women who were very consciously writing about things that were more obviously male in nature and that this conscious digression from the gender norm compelled them to conceal their true identities. Provocative and tantalizing, yes. We will never know the truth.

Considering the possible motives for anonymity in this corpus, Jockers observes in related work (2013) that

[18] To cross-check and corroborate our topic driven result, we ran a parallel experiment using simple "stylistic" data--i.e. high-frequency word data--in place of the topical data. The cross-validation results for a style-based classification of gender showed an average probability of correct classification of 77%. The use of high-frequency word and punctuation features in author and gender attribution research is well established. See specifically Koppel, M., Argamon, Shlomo, and Shimoni, Anat Rachel. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17 (4):401-412. More generally, see Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* 22 (3):251-270.

[19] We also classified the 145 unknown texts using the more traditional word-frequency model discussed in the previous footnote. The two tests (thematic and stylistic) agreed on all but 13 texts (91% agreement).

> several themes in the corpus are over represented in works of anonymous authorship, and these themes often relate to socio-political institutions such as the monarchy, as seen in topics 97, 108, and 239, or to religious institutions as in the “Convents and Abbeys” theme found in topic 31 or the theme of “Religion” found more generally in topic 448. At the same time, these anonymous works are also very high on a theme dealing with the expression of opinions, topic 28. All of this makes perfect sense if we believe that these authors felt the need to conceal their identities in order to present a more candid portrait of politics and/or religion . . . Taken together this evidence suggests a class of writers generating thinly veiled narratives that express opinions about religious and nationalistic matters that would have likely been awkward or impossible to express without the use of a pseudonym.”

Using a combination of the permutation and bootstrap plots along with some further classification experiments, we tested these hypotheses and in the end found them lacking. We began by studying the themes identified in Jockers’s previous work, themes that were, or appeared to be, over-represented in the class of unknown authors. The theme labeled “Convents and Abbeys” (figure 10) seemed to be an obvious candidate: it was over represented in the unknowns, and it could be considered a controversial topic given the anti-Catholic (and frequently gothic) tradition of convent confessions in which anti-Catholic authors penned sensationalized tales detailing horror, abuse, depravity, and corruption within church institutions.[20] Such tales were frequently purported to be the first hand accounts of “escaped” nuns and thus provided a perfectly valid, if thinly veiled, excuse for anonymity.

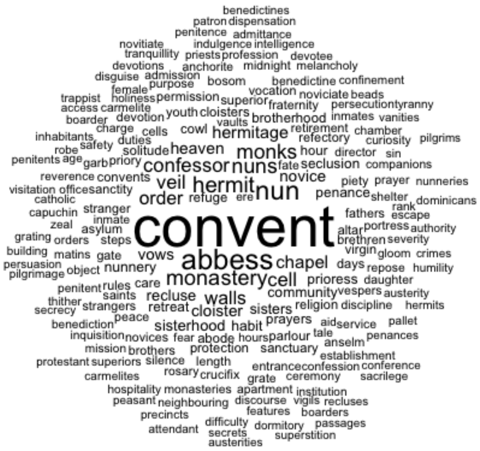


Figure 10: The “Convents and Abbeys” word cloud.

[20] Rebecca Reed’s *Six Months in a Convent* (1835) is an obvious example

Figures 11-12 show the corpus means for the Convents and Abbeys theme in the context of both the randomized data and the bootstrapped data. Once we examined the corpus means within the context of random chance, the smell of a smoking gun quickly dissipated. What became clear was that the high value of the corpus mean for Convents and Abbeys was largely the result of two outlier texts that were pulling the mean in an artificially high direction. This can be seen in the very wide distribution of the fake data in the unknown columns of figure 11 and figures 12. The permutation plot (figure 11) shows that the mean proportion of Convents and Abbeys in books with unknown authors is at the extreme end of what might be expected by random chance.

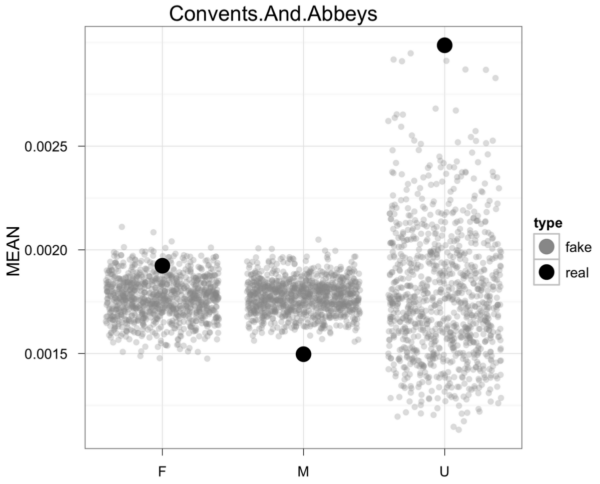


Figure 11: Convents and Abbeys Randomization Plot.

The bootstrap plot (figure 12) indicates that the use of this topic by unknown authors is variable, and could easily be closer to the range expected for female authors.

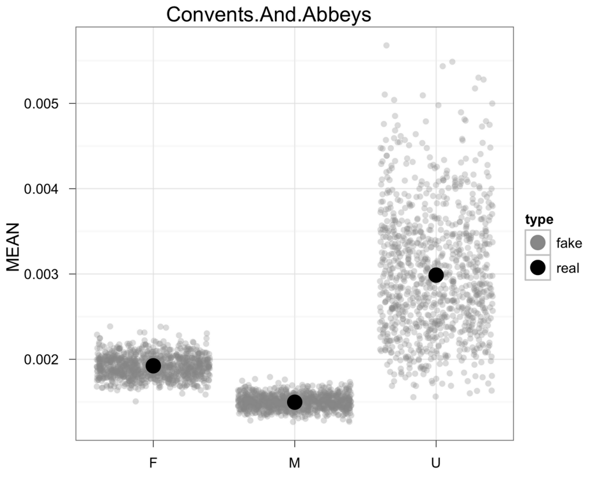


Figure 12: Convents and Abbeys Bootstrap Plot

Having abandoned the Convents and Abbeys theme as a reliable indicator of anonymity, we continued to interroge the data and Jockers's initial hypothesis about a possible relationship between anonymity and subversive or controversial themes. Using the NSC algorithm, we ran a classification experiment in which a model was trained to use topical proportions to differentiate between male-, female-, and unknown-authored texts. Classification accuracy hovered around 42%. We ran another test in which males and females were merged into a single class of "known" authors. The classifier had only to distinguish between authors of known gender and unknown gender. Here model accuracy rose to 67%, but the features the model found most useful in separating the two classes were a hodgepodge of themes with no obviously unifying thread.[21] In the end, no unifying thread to tie up the unknowns could be found, but our investigation of the unknowns did lead us to a final exploration of how gender may not only predict fluctuations in the use of themes but in the individual word choices within the themes.

Our model assumes that themes are represented by a single weighted combination of words that is the same regardless of context. This assumption is implicit in our representation of themes as word clouds. The word cloud in Figure 10 presents the theme of Convents and Abbeys as a *thing*, essentially a Platonic concept. "Convent" is the most likely word, "hermit" somewhat less likely, and "crucifix" one of many words that are probable, but infrequent. With this word cloud we are stating that if this theme appears in a novel, we will see these words, in roughly these proportions.

We know, however, that this assumption is not true: different authors emphasize different words within a theme. There is no way such a simple probabilistic model could account for the full variability of word choice in novels. But how good of a fit is it?

To explore this question, we can use a method similar to the permutation test presented earlier.[22] Our topic model implicitly tells us what our observed data should look like if the Platonic themes assumption is true. If a word in a novel is assigned to the Convents and Abbeys topic, it should have the same (large) probability of being "convent" and the same (smaller) probability of being "hermit" regardless of who the author is. We can therefore compare the real words that we observe to samples from this model distribution. If the assumption holds, real data should fall within the range of variability of fake data. If, however, there are correlations between certain authors and certain words *within a theme*, we should see observed values outside that range.

As before, we define a measurement of interest, and then a method for generating fake data. In this case, we are interested in whether certain words are used disproportionately within a theme by male or by female authors. Topic model inference involves assigning each word token to one topic. We estimate the probability of a word (say, "convent") in a topic by gathering all of the words assigned to the topic, counting how many of them are "convent," and dividing by the total number. We can similarly estimate the probability of the same word in novels by male authors by collecting all the words assigned to the topic that appear in books by male authors, counting how

[21] Topping the list were "Affections Passions Feelings Of Attachment," "Situations Of Female Heroines," "Marriage 1," "Accounts And Opinions," "Affection And Happiness," "Hearth Fires," "Parties And Conversation," "Apartments And Chambers," "Us Dollars And Us Cities," "Genius And Talent"

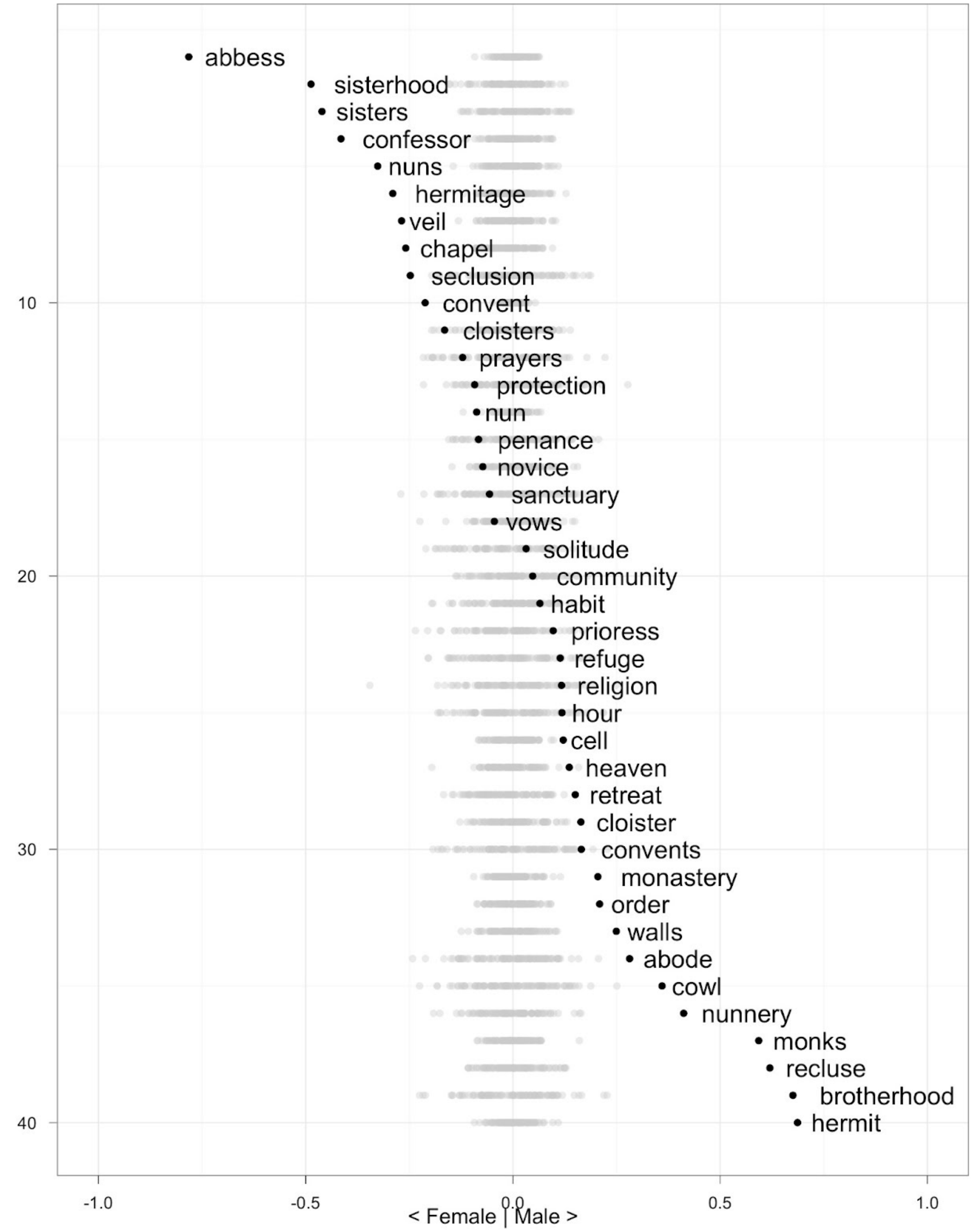
[22] Mimno, David and Blei, David. (2011) Bayesian Checking for Topic Models. In proceedings of *Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland.

many of those words are “convent,” and dividing by the total number of words in the topic in the male-authored subset of the corpus. Note that that total number could differ substantially between sub-corpora. We know from Figure 11 that there will be fewer words about Convents and Abbeys in male-authored novels than in female-authored novels. We want to know whether the number of instances of “convent” in contexts about Convents and Abbeys in novels by male authors is higher or lower than we expect. When we subdivide the data so finely, we must be concerned with small sample sizes.

In order to control for the variability caused by small samples, we compare our actual measurements to fake data. Rather than shuffling the labels of whole books, we shuffle the specific words used in books. In each fake data set, the total number of words assigned to the Convents and Abbeys topic in male- and female-authored novels will be constant, but the specific words will be randomly assigned to novels.

Figure 13 provides a visualization of the top 40 words in the Convents and Abbeys topic cluster, sorted by their relative use by female authors. The position of the dots represents the relative use of a word by female and male authors.[23] Dots further to the left (in the negative direction) show relatively greater density in works by female authors, while dots further to the right (in the positive direction) show greater density in works by male authors. The black dots record the actual values. The lighter grey dots hovering around either side of the center of the vertical “0.0” axis represent the expected range if gender were not an influence.

[23] The position on the x-axis represents the log of the probability of the word in male-authored works divided by the probability of the same word in female-authored works. Thus, if the word has greater probability in female-authored works, the ratio will be less than one, so the log will be negative.



Beginning at the top and reading down, one finds ten words (abbess, sisterhood, sisters, confessor, nuns, hermitage, veil, chapel, seclusion, convent) from the cluster that are highly associated with works of female authorship. The last ten (monastery, order, walls, abode, cowl, nunnery, monks, recluse, brotherhood, hermit) are far more indicative of male-authored texts. The words closer to the center (i.e. sanctuary, vows, community) tend to be less gendered. The width of the expected range under the static-topic hypothesis is a function of the frequency of the word. The most common word, "convent," has the narrowest range. Even though the actual difference in word density for “convent” is not as great as “brotherhood,” we can nevertheless be confident that this difference is not due to random chance.

Having now visualized the topical data in this highly gendered manner, it seems, in retrospect, that a more appropriate label for this topic might have been “Convents and Monasteries.” With a few exceptions (i.e. “nunnery” in the male dominated half and “confessor” in the female half), the words used by the male and female authors within this topic split into word groups closely associated with the respective, and highly segregated, institutions that are convents on the one hand and monasteries on the other; or, sisterhood vs. brotherhood.

In ways that are at once revealing and fascinating this word distribution data opens new avenues for investigation and closer analysis. After all, what we are really examining here are 19th century gender norms of thematic attention. While it may be the case that the norms revealed here conform to expectation--males write about monks and women about nuns--it is revealing nevertheless to measure how biased the perspectives really are. In the context of this macroscale picture, we must now return to those specific portraits of 19th century Catholicism and examine, for example, how the monks and sisters of two prototypical gothic novels in this vein--Radcliffe’s *The Italian* and Matthew Gregory Lewis’s *The Monk*--may be represented not only through a very stereotypical anti-Catholic perspective, but through the uniquely gendered perspective of the novel’s authors.[24]

Conclusions

Topic modeling is useful in analyzing literature, but, as our work here suggests, it must be applied with care and within the context of statistical tests that can measure confidence in results. Topic models can identify broad themes in literature. We can use such models as a means of detecting and measuring differences in the concentration of themes from one section of a corpus to another. These measurements by themselves, however, do not provide a full picture: some measured differences are meaningful, others are not. Getting quantified results from models can be deceptively, seductively easy.

With this in mind, we tested the significance of the model’s measurements in three separate ways. First, we compared the observed differences to the expected range of random variation given the size of our samples using a permutation test. Second, we checked whether observed differences are sensitive to outliers using a bootstrap test. Third, we measured the predictive power of themes using a classification test. And then, in addition to measuring the concentration of topics as a whole, we measured patterns in the use of specific words within each topic, using a similar battery of significance tests.

[24] Such a reinvestigation seems especially promising with regard to *The Monk* where existing scholarship on gender roles in the novel has already been explored in some detail. See, for example, Blakemore, Steven (1998). "Matthew Lewis's black mass: sexual, religious inversion in *The Monk*." *Studies in the Novel* **30** (4): 521–39 also DeRochi, Jack (2007). A Feminine Spectacle: The Novelistic Aesthetic of Matthew Lewis's The Captive.” In *Prologues, Epilogues, Curtain-Raisers, and Afterpieces*. Edited by Daniel James Ennis and Judith Bailey Slagle. Newark: University of Delaware. pp. 238–252.

The approach we present here scales to collections containing of thousands of books. It is an approach far beyond the reach of traditional scholarly methods, which depend in large part upon the careful, close reading of individual texts. The models we present here cannot represent the full meaning of individual books any more than satellite photos can show the details of individual trees. Like the satellite view, however, these macro-, or "distant-," scale perspectives on literature offer scholars a necessary context for and complement to closer readings.