Who wrote Ronald Reagan's radio addresses?

Bayesian Analysis (2006) **1**, Number 2, pp. 289–320

Who Wrote Ronald Reagan's Radio Addresses?

Edoardo M. Airoldi*, Annelise G. Anderson†, Stephen E. Fienberg‡, and Kiron K. Skinner§

Abstract. In his campaign for the U.S. presidency from 1975 to 1979, Ronald Reagan delivered over 1000 radio broadcasts. For over 600 of these we have direct evidence of Reagan's authorship. The aim of this study was to determine the authorship of 312 of the broadcasts for which no direct evidence is available. We addressed the prediction problem for speeches delivered in different epochs and we explored a wide range of off-the-shelf classification methods and fully Bayesian generative models. Eventually we produced separate sets of predictions using the most accurate classifiers, based on non-contextual words as well as on semantic features, for the 312 speeches of uncertain authorship. All the predictions agree on 135 of the "unknown" speeches, whereas the fully Bayesian models agree on an additional 154 of them.

The magnitude of the posterior odds of authorship led us to conclude that Ronald Reagan drafted 167 speeches and was aided in the preparation of the remaining 145. Our inferences were not sensitive to "reasonable" variations in the sets of constants underlying the prior distributions, and the cross-validated accuracy of our best fully Bayesian model was above 90 percent in all cases. The agreement of multiple methods for predicting the authorship for the "unknown" speeches reinforced our confidence in the accuracy of our classifications.

Keywords: Ronald Reagan, Radio Addresses, Authorship, Stylometry, Data Mining, Classification, Function Words, Semantic Analysis, Naïve Bayes, Full Bayes, Poisson, Negative-Binomial, Modal Approximation, Mean Approximation

1 Introduction

1.1 Historical Background

Ronald Reagan was elected Governor of the State of California in 1966, and re-elected in 1970 for a second term running through the first days of 1975. When he originally considered running for the 1976 Republican presidential nomination, he could have expected to run for an open seat since Nixon's two terms would be over. But as events developed, Nixon's vice president, Spiro Agnew, resigned over a scandal in 1973. Gerald R. Ford replaced Agnew and then assumed the presidency when Nixon resigned over the Watergate scandal in August 1974. Thus in 1976 Reagan was challenging an incumbent president of his own party. Reagan narrowly lost the 1976 campaign for the

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, http://www.cs.cmu.edu/~eairoldi/

†Hoover Institution, Stanford University, Stanford, CA, http://www.stanford.edu/~andrsn/

‡Department of Statistics and School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, http://www.stat.cmu.edu/~fienberg

§Departments of History and Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, http://www.hss.cmu.edu/departments/sds/faculty/skinner.html

Republican nomination to Gerald Ford, who then lost the presidency to Jimmy Carter in the 1976 general election. But as early as 1975, Reagan had began to focus on the 1980 presidential election.

A series of events caused Jimmy Carter's public approval ratings to decline from 70 percent to a low 28 percent during his term in office. Among them were scandals involving some of his staff members, his brother's alcohol problems, record levels of inflation, the energy crisis, and finally, on November 4, 1979, the seizure of the American embassy in Tehran by Iranian students who captured 63 American citizens and held 50 of them as hostages for 442 days. As the hostage crisis continued throughout 1980—the election year—the television networks opened their evening newscasts with a count of the number of days the hostages had been held. Despite claims by his supporters that Carter was the better prepared candidate, he lost the election to Reagan.

Reagan started promoting his policies and his image in 1975 by means of weekly radio addresses. A Hollywood and television actor through 1964, Reagan had a very good relationship with TV cameras. He presented himself as a leader who would give America back to Americans, who did not want government to intrude in many social concerns because Americans could handle situations themselves, "the American way." Reagan proposed the buildup of military strength as essential to national defense and promised to rebuild a professional, well-paid army of one million soldiers that would be ready to intervene anytime. He promised to make America the "shining city on a hill" that all the world had admired in the "good old days." Overall, the radio addresses were a major vehicle by which Reagan developed and communicated approaches and positions on a wide range of public policy issues and let his listeners know who he was.

1.2 The Data

We accessed a computerized database containing the texts of 1032 radio addresses Ronald Reagan delivered between 1975 and 1979. It was a number so substantial that, provided Reagan actually wrote the words, it contradicts the conventional wisdom that Reagan was merely a communicator of other people's ideas. Reagan's original drafts of 679 radio addresses written "in his own hand" have been found, but there are no drafts for the remaining 353. Of these, archival evidence has confirmed that members of Reagan's staff drafted 39. The authorship of the remaining 314 is uncertain. Of the 1032 commentaries, about 30 percent are on foreign policy and national defense. About 70 percent deal with national domestic issues—the economy (taxes, inflation, unemployment, overregulation), the energy crisis, and other major domestic programs like health, education, social security, and welfare. A few others are more personal, about people—some famous and some unknown—whom Reagan respected for their courage in dealing with life's difficulties and their willingness to help others. Of the speeches we labeled here as "written by Reagan in his own hand," 237 are contained in *Reagan, In His Own Hand*, edited by Skinner et al. (2001a) over 330 in *Reagan's Path to Victory*, edited by Skinner et al. (2004) and a few more in *Stories in His Own Hand*, edited by Skinner et al. (2001b). While reading these books, one can find suggestions on how to distinguish Reagan's style from those of others who worked for him. Reagan was direct and informal when he talked to people; he spoke our language. He did not keep a

distance. Because these speeches are a major key to understanding why Reagan wanted to run for the presidency and what he might do if elected, identifying those of uncertain authorship that are likely to be his own work contributes to understanding the man and the president he became.

We worked with an electronic version of the Reagan speeches prepared as part of a larger ongoing project on the former president. Because of the availability of handwritten drafts, we were able at the outset to attribute 679 speeches to Ronald Reagan and 39 to his collaborators (12 to Peter Hannaford, 26 to John McClaughry, and 1 to Martin Anderson). Authorship of the remaining 312 speeches is uncertain.[1] As Reagan's main collaborator on the radio addresses, Hannaford probably wrote the initial drafts of those Reagan did not write, although a few other people contributed possible drafts. Because of this, we also coded several of Reagan's newspaper columns which are known to have been drafted by Peter Hannaford.

Author	1975	1976	1977	1978	1979	Total
R. Reagan	60	195	52	219	153	679
P. Hannaford (radio)	1	5	2	4	0	12
P. Hannaford (news)	5	0	7	18	0	30
J. McClaughry	0	3	1	15	7	26
M. Anderson	0	0	0	0	1	1
Author uncertain	149	80	4	25	56	314
Total (known author)	66	203	62	256	161	748
Total (all)	215	283	66	281	217	1062

Table 1: Breakdown of the available texts by author and year.

Related Work. Augustus De Morgan in his *Budget of Paradoxes* noticed that "some time somebody will institute a comparison among writers in regard to the average length of words used in composition, and that it may be found possible to identify the author of a book, a poem or a play in this way." In a supplement of *Science* dated March 11, 1887, T.C. Mendenhall followed up De Morgan's idea and showed what he called the characteristic curves of composition. In the same fashion that the spectroscope can be used to assess the presence of a certain element in a solid object, Mendenhall associated characteristic curves, or word-spectra, to different authors under the assumption that each writer makes use of a vocabulary peculiar to himself, the character of which does not change over his productive period. In the long run, he expected that short words, long words, and words of medium length occur with definite relative frequencies. Mendenhall's assumptions were quite strong and soon his conjecture about one word-spectrum for each writer was disproved. However, the fundamental idea that numerical summaries of texts could be used to extract relevant information about authorship was born.

The early approaches to authorship attribution problems stemmed from the studies

[1]Originally 314, two speeches were removed because they essentially contained only quotations, which were excluded from the analysis, leaving 312 for us to classify.

of G. Zipf (1932) on power laws, and of U. Yule (1944) on literary vocabulary. Frederick Mosteller and David Wallace (1964, 1984) defined the problem of authorship attribution in its modern mathematical form in their book on the case of the 12 disputed *Federalist Papers*, and George Miller (1954) presented the linguistic community with several mathematical solutions to the problem. Because of the work of Mosteller and Wallace (1964, 1984), a widespread strategy for authorship attribution problems is to consider high frequency function words (the *filler* words of the language, such as *a, an, by, to* and *that*, not related to the context). Burrows (1992) and others suggested summarizing the information in terms of its principal components, or using some other method suitable for dimensionality reduction. Natural clustering of the texts in the space spanned by these few highly descriptive features is investigated, and an attempt to classify the texts is made.

Outline of the Paper. In this paper, we attempt to resolve in statistical terms the authorship of the radio addresses lacking attribution. As described in section 1, we begin by learning how to discriminate between the writing styles of Reagan and his collaborators in 1975. Then we focus on stylistic differences between Reagan and Hannaford over the years 1976-79. In section 2, we use exploratory methods to identify some features that distinguish Reagan's literary style from that of his collaborators beyond differences we could expect to find in several writings by the same author. Then in section 3, we present a fully Bayesian approach that allows us to estimate posterior odds of authorship. In section 4, we summarize our results: our best "machine learning" classification methods agree in predicting the author of 135 out of 312 speeches Reagan delivered over the years 1975-79, and the more reliable fully Bayesian models agree in predicting 289 of them. The cross-validated accuracies of our methods on about 750 "known" speeches range between 95 percent and 85 percent, with standard deviations of about 3 percent and 9 percent on texts drafted by Reagan and others respectively. A fully Bayesian approach based on the Negative-Binomial model best captured the variability in the data and yielded quite stable predictions across 21 sets of underlying constants.

Notation. The main random quantity in our analyses is the number of times a word appeared in a text, which we denote by X. We generally label probability distributions with the letter p, as in $p(X \mid \theta)$, where θ is a generic vector of parameters. Occasionally we refer to a distribution by its name, as in $Beta(X \mid \theta)$ or $Gamma(\xi \mid \beta)$, where the first argument indicates the random quantity, and the second the parameter vector, which is random itself. Indices appear in a few places: we used X_{nij} for word $n = 1, \ldots, N$ in document $j = 1, \ldots, J_i$ of author $i = 1, 2$; Reagan is always author $i = 1$, whereas we use $i = 2$ to denote Hannaford or the undifferentiated group of Reagan's collaborators, depending on the occasion.

2 Summary of Data Mining EDA Approaches

In this section we introduce the features we consider to capture the literary style of the authors who drafted the Reagan speeches, we discuss the feature selection strategies we adopt and the relevant issues they address, and we identify interesting aspects of

the data using exploratory techniques. Finally, we analyze the performance of several off-the-shelf classification methods in predicting the author of a speech.

2.1 Feature Selection as a Multiple Testing Problem

A preliminary goal of our study was to capture those elements of Ronald Reagan's writing style that would help us differentiate the speeches he drafted from those Hannaford drafted, and from those other collaborators drafted. The classification methods we used take integers as input, hence we looked for features whose frequency of use in the writings of the different authors was diverse enough to support the hypothesis that it was an expression of the differential writing style we were seeking.

In order to perform the classification task, we focused on the use of three types of features: words, n-grams, and semantics. Words are defined as sequences of letters[2] enclosed by non-letters, n-grams are ordered sequences of n adjacent words, and semantic features are defined as sets of patterns of words relevant to a representational theory of composition,[3] as discussed in Collins and Kaufer (2001). In order to quantify the frequencies of use of a feature in the texts of two authors as diverse enough to consider that feature as a good discriminator, several criteria are possible, each of which assigns scores to features and ranks them according to their discriminating power. As selection criteria, we used the information gain and the information gain ratio as defined in Mitchell (1997), a stepwise procedure described in Mosteller and Tukey (1968), the Kolmogorov-Smirnov statistic described in Wasserman (2004), the Welch approximate t-tests proposed by Welch (1938), and the Δ^2 statistic, originally proposed as a useful heuristic by Mosteller and Wallace (1964, 1984), and for which Airoldi (2003) and Airoldi et al. (2005) performed a full distributional study.

Briefly, if X_{nij} denotes the number of times the n^{th} word in the dictionary appears in the j^{th} document written by the i^{th} author, and $\{x_{n1j} : j = 1, ..., J_1\}$ and $\{x_{n2j} : j = 1, ..., J_2\}$ denote the observed counts in the texts, the value of Δ^2 statistic for feature n is computed according to the following formula:

$$\Delta_n^2 \quad = \quad \frac{\left(\sum_{j=1}^{J_1} x_{n1j} - \sum_{j=1}^{J_2} x_{n2j}\right)^2}{\sum_{j=1}^{J_1} x_{n1j} + \sum_{j=1}^{J_2} x_{n2j}}. \tag{1}$$

We use Δ^2 to test whether the two sets of observed counts come from different Poisson or Negative-Binomial distributions. The other selection criteria intuitively perform similar tests but assume different frameworks; the Kolmogorov-Smirnov test does not assume a parametric form for the distributions of X_{n1}, X_{n2}, the Welch t-test compares the means of X_{n1}, X_{n2} without assumption on the variances, and so on.

[2]Numbers and dates were transformed into keywords (DG and DT, respectively) in a preliminary data preparation and cleaning stage; preprocessing steps are fully documented in Airoldi (2003).

[3]This theory entails a method of characterizing language choice which is drawn from rhetoricians' long-standing interest in the patterns of language that provide interactive experiences for an audience. For example, the simplest way for readers to feel the elapse of time is through strings containing the simple past tense, and the advantage of the simple past for writers is that event starts and stops can be signaled to the reader. The feature *Past Events* captures all n-grams that contain simple past tenses.

The issues we must consider to assess the goodness of a feature selection strategy are: (a) whether or not a selection criterion entails a measure of the degree of certainty about the discriminative power of each feature; (b) if statistical tests are performed to select features, we need to take extra care to deal with the sharply different number of texts written by the various authors in our sample; (c) if several statistical tests are performed, we need a correction for multiple tests; and (d) we want to select features whose discriminatory power is high and stable over non-overlapping batches of texts to be able to make robust predictions for the unknown texts. With regard to (a) and (c), the scores based on information gain arguments provide a ranking of the features, whereas the Kolmogorov-Smirnov statistic, the Welch approximate t-statistic, and the Δ^2 statistic allow for p-values to be computed and for a correction for multiple tests to be applied. With regard to (b), both the Welch t-test and the Δ^2 statistic allow for corrections for unbalanced sample sizes. For point (d), we divided the speeches in two batches (those broadcast in 1975-77, and those broadcast in 1978-79) and performed a two-stage selection process where we pruned the words that passed the first selection using the second batch of documents.

In our experiments we started from six different pools of features, we scored features in each pool using the six criteria above, and we used the two-stage selection procedure and the correction for multiple tests when appropriate, for both the case of Reagan versus Hannaford and the case of Reagan versus his collaborators. Below, we summarize the results of five selected strategies, each applied to a different pool of features[4] we considered, for the case of Reagan versus Hannaford.

High-frequency words: We obtained the first pool of features by handpicking function words (we found 267 of them) among the most frequent 3000 words in Reagan and Hannaford vocabularies. We performed feature selection by using both the Kolmogorov-Smirnov statistic and the Welch approximate t-statistic, and we corrected the p-values to take into account the fact that multiple tests were performed using the false discovery rate[5] (FDR) proposed by Benjamini and Hochberg (1995). Using this strategy we ended up with a list of 55 discriminating words.

SMART list of words: The second pool of features that we considered as a starting point was the list of 523 words in the SMART text categorization system by Salton and Buckley; their system would remove these words from the analysis as considered not useful in the classification of texts by topic. Yet, this list of words fits our purpose of classifying texts by author since authors write about multiple topics. Therefore, we performed feature selection by using both the Kolmogorov-Smirnov statistic and the Welch approximate t-statistic, and we corrected the p-values to take into account the

[4]We do not present here any of the results regarding the pool of common, weakly discriminating words {*and, in, the, or, of*}, used by Mosteller and Tukey (1968). For such results, and for a complete account of other strategies we applied to the five pools of features described in the text, we refer to Airoldi (2003).

[5]The FDR correction proposed by Benjamini and Hochberg (1995) assumes independent tests. In section 3, we show that our data supports the independence hypothesis for some words; however, we do not expect independence to hold. The FDR correction has been shown to hold for dependent data in Storey et al. (2004), under mild conditions. In any case, the correction can be regarded a practical approximation.

fact that multiple tests were performed using the false discovery rate, ending up with a list of 62 discriminating words.

Semantic features: As a third pool of features, we retrieved the 21 semantic features discussed in Collins and Kaufer (2001). Again, we performed feature selection by using both the Kolmogorov-Smirnov statistic and the Welch approximate t-statistic, and we corrected the p-values to take into account the fact that multiple tests were performed using the false discovery rate. Further, we explored the relevance of these features by using a double jackknife procedure on a linear discriminant function following Mosteller and Tukey (1968) and ended up with six weakly discriminating features.

Information gain: The fourth pool we considered consisted of all the words in Reagan and Hannaford vocabularies. One of the most widely used methods to perform feature selection in the computer science approaches to text classification is based on the scores derived from information theoretic arguments. Our goal was to find a small, robust set of discriminating features. Hence we selected 100 words with highest information gain and information gain ratio scores, and we then handpicked the non-contextual words among them. We ended up with about 30 discriminating words in both cases.

Two-stage selection on 4-grams: The fifth pool of features we started with consisted of all the unique 4-grams: $69,000$ in the Hannaford dictionary and $729,000$ in the Reagan dictionary. The strategy we adopted was a two-stage selection procedure: the first stage consisted in using the statistic Δ^2 to perform feature selection on all the 4-grams using two batches of texts in sequence, as discussed above, to mitigate selection effects. In the second stage, we performed feature selection by using both the Kolmogorov-Smirnov statistic and the Welch approximate t-statistic, and we corrected the p-values to take into account the fact that multiple tests were performed using the false discovery rate. This strategy returned a list of 41 discriminating words and bi-grams, for $\alpha = 0.01$. As an alternative two-stage strategy: in the first stage we used the statistic Δ^2 to perform feature selection on all the 4-grams using two batches of texts in sequence, as discussed above, to mitigate selection effects; in the second stage we used the statistic Δ^2 again, and we corrected the p-values to take into account the fact that multiple tests were performed using the false discovery rate. This strategy returned a list of 142 discriminating words and bi-grams, for $\alpha = 0.05$.

Some of our preliminary explorations led to dead ends. For example, in some of the early analyses it appeared possible to perfectly classify Reagan speeches using abbreviations, punctuation, and the American and Canadian spellings of words such as *theatre, theater*. Further inspection of the original documents revealed that Reagan spelled such words both ways, and that he was very inconsistent about spelling and punctuation. In addition, the spelling was often changed from an original manuscript during retyping or editing. The original drafts were typed, both in the offices of Deaver & Hannaford and in the offices of Harry O'Connor (where they were recorded) by many different people who had different views of spelling, capitalization, and so forth. Thus, these differences were not indicators of authorship, and we did not use these features in our subsequent analyses.

2.2 Exploring Feature Spaces with Stylometry Tools

It is possible to form a preliminary assessment about which selection strategy finds the most discriminating features through visual inspection by using two exploratory tools widely adopted by the literary community in authorship attribution studies: principal component analysis and unsupervised clustering of the texts.

Principal Components: Given a candidate pool containing V discriminating features, we represent documents as vectors containing in the n^{th} position the number of times the n^{th} feature in the pool appears in that document; i.e. for the j^{th} document of author i we write $[\ X_{1ij},\ X_{2ij},\ ...,\ X_{Vij}\ ]$. The set of texts is thus represented as a *matrix of counts* with as many rows as the number of texts in the data set, and as many columns as the number of words in the pool. A document can be thought of as a point in a V-dimensional Euclidean space, or the feature space.

Following the practice in authorship attribution studies, we built the matrix of counts for the pool containing the 267 highest frequency function words. We then performed the principal analysis decomposition of such a matrix as discussed in Burrows (1992) in order to visualize the texts written by the various authors as points in the lower dimensional space spanned by the first few principal components. We performed the same operations for each of the pools of discriminating features we obtained by combining the six starting pools of words with the selection strategies we considered. In figure 1 we present the results for four example pools of features; in each panel we display the texts in the space spanned by the first and second principal components of the corresponding matrix of counts.

Visual inspection of the information summarized by the principal components of both high-frequency words (top-left panel) and semantic features (bottom-left panel) in the texts by Reagan and Hannaford suggests that these features are weakly discriminating overall. On the other hand, the information summarized by the principal components of the words we selected using the Δ^2 statistic (top-right panel) and Information Gain (bottom-right panel) suggests that these features have potential for discrimination, and thus may capture some of those elements of differential writing style we are looking for. It is up to the selection strategy to provide a measure of confidence that such elements of style are not the outcome of pure chance. To this extent information gain lacks a measure of confidence, whereas the Δ^2 statistic provides us with such a measure: the p-values corresponding to each word. What principal component analysis lacks is an overall judgment about whether the degree of separation of the texts we observed is likely to be the outcome of pure chance.

In conclusion, visual inspection of the texts by principal component analysis suggested the features we selected using information gain and the statistic Δ^2—which include few medium and low frequency non-contextual words—entailed a stronger discriminating potential than the features we selected using other strategies. Although principal component analysis does not allow for an overall probabilistic judgment about the separability of the texts, the selection strategy based on the statistic Δ^2 provides us with features that are likely to capture elements of differential writing style.

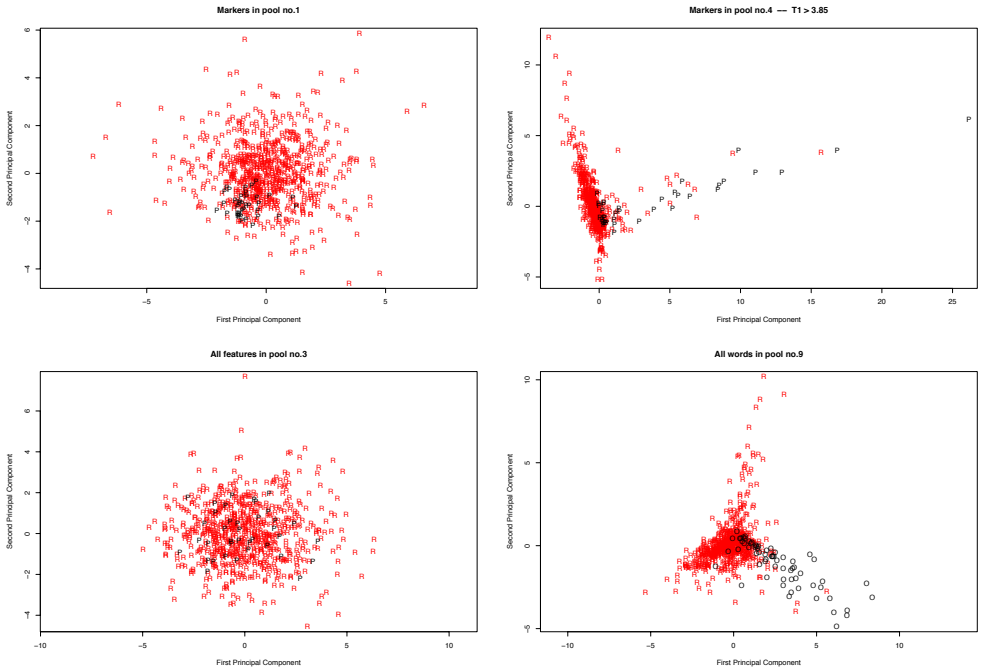

Figure 1: The information contained in the high frequency words (top-left panel) is compared to the information contained in the 41 markers we found using the statistic Δ^2 (top-right panel), to the information contained in the 18 semantic features (bottom-left) and in the 30 words selected using Information Gain (bottom-right).

Unsupervised Clustering: The starting point for unsupervised clustering methods is the matrix of counts, where all the texts in our data set are represented as row vectors $[\ X_{1ij},\ X_{2ij},\ ...,\ X_{Vij}\ ]$. In such a representation, a text is described in terms of the number of times words (or features) were used by its author. A widely adopted practice in authorship attribution studies is to compute the distances between pairs of texts and then use these distances to form a *dendrogram*—that is, a visual tool to analyze the clustering behavior of the texts.

Briefly, a dendrogram is a tree whose leaves correspond to clusters formed of single texts and whose root corresponds to a cluster containing all the texts. In a tree oriented with root at the top and leaves at the bottom, we measure the distance according to a specific metric on the Y axis; the X axis can be ignored. Nodes in the tree denote clusters that include all their corresponding leaf-nodes. Last, clusters are formed at different distances by aggregating leaf-nodes via *hierarchical clustering algorithms*,[6] which intuitively iterate through all texts and assign a text to a cluster at each pass.[7]

[6] Example algorithms used are: single linkage, average linkage, complete linkage, median, Ward, and centroid. For details on all of these see Hastie et al. (2001).

[7] The specific heuristics used to decide which texts are assigned to which cluster at each pass differ-

For each pool of discriminating features we analyzed the dendrogram formed by average linkage using euclidean distances, the common choice in authorship attribution studies. We noted a lack of natural clustering for the texts of the same author. We extended our analysis to other metrics (city-block, Mahalanobis, cosine, correlation and Minkowski) and to other aggregation algorithms (single-linkage, complete-linkage, median, Ward and centroid), and eventually we performed clustering on the texts, as described by few principal components instead of word counts, again to discover a lack of natural clustering of texts written by the same author. The observed lack of clustering may be a by-product of the fact that in our data set there are few texts written by authors other than Reagan. In conclusion, we anticipated that classification methods that borrow their strength from some notion of *geometric closeness* of the texts written by the same author, e.g., nearest neighbor, would be ineffective in this problem.

2.3 Off-the-Shelf Data Mining Methods

An empirical assessment of the discriminatory power of the features selected by the various strategies relies on the *balanced cross-validated* accuracy of a set of standard classifiers. This approach allows us to gain further insights about the texts, described as a bag of words, by analyzing the hypotheses underlying successful methods.

According to our balanced cross-validation scheme, 80 percent of the texts written by each author were used to train a classifier, and the remaining 20 percent were used to test it. We estimated the accuracy of a classifier as the average accuracy the classifier achieved over 1000 such 80/20 experiments; notice that only out-of-sample texts were used during the estimation to reduce bias. Hastie et al. (2001) discuss alternatives.

Naïve Bayes: In our experiments naïve Bayes classifiers performed best. Recall that we represent the j^{th} text written by the i^{th} author as a vector, $X_{ij} := [\ X_{1ij},\ X_{2ij},\ ...,\ X_{Vij}\ ]$, whose coordinates are random variables encoding information about words 1 through V in a certain pool. Each classifier in this family assumes texts are generated according to a certain parametric model, $p(X_{ij}\,|\,\theta_i)$, where θ_i is a generic vector of parameters describing the writing style of the i^{th} author. The training data is used to estimate θ_i for each author. These estimates are then composed with prior probabilities of authorship, π_i, via the Bayes rule to classify a new text, $[X_{1\,new}, ..., X_{V\,new}]$, according to the posterior probability that an author would have generated it, or, equivalently, according to its *odds* of authorship—that is, ratio of posterior probabilities:

$$odds(X_{new}) \quad := \quad \frac{p(\,[X_{1\,new}, ..., X_{V\,new}]\,|\,\hat{\theta}_1)\cdot\pi_1}{p(\,[X_{1\,new}, ..., X_{V\,new}]\,|\,\hat{\theta}_2)\cdot\pi_2}.$$

In the *multivariate Bernoulli* model $p(X_{ij}|\theta_i) = \prod_n Bernoulli(X_{nij}|p_{ni})$—that is, a product of binary random variables which encode presence or absence of the corresponding words. In the *multinomial* model $p(X_{ij}|\theta_i) = Multinomial(X_{ij}|p_{1i}, ..., p_{Vi})$—that is, a random vector whose coordinates encode the number of occurrences of the corresponding words. In both cases, $\theta_i = [p_{1i}, ..., p_{Vi}]$. These models assume independence:

entiate the various algorithms.

(I_1) among occurrences of pairs of words, i.e. $X_{nij} \perp X_{mij}$ for $n \neq m$, and (I_2) among occurrences of the same word, i.e. $p(X_{nij}|\theta_i)$ does not depend on previous occurrences of the word n. Further, (I_3) these models are independent of the position of words in the text. We discuss these assumptions in section 3; for now we note that assumption (I_2) is the major cause for extreme log-odds which harm the credibility of these models (Mosteller and Wallace 1964, 1984; Church 1995).

In order to correct possible inaccuracies of the maximum likelihood estimates for θ_i due to the sparseness of the data, we tried various adjustments (Dirichlet, M-estimate, Good-Turing, Witten-Bell) proposed in the literature. For example, the Dirichlet adjustment boils down to estimating θ_i with the mean (vector) of the posterior distribution obtained by combining a multinomial model for the word counts, $Multinomial(X_{ij}|p_{1i}, ..., p_{Vi})$, with a symmetric Dirichlet prior for its underlying parameters, $Dirichlet(p_{1i}, ..., p_{Vi}|\eta)$, for a fixed underlying constant η. This increases (by η) the observed counts of all words, and forces the estimates for those words that passed the selection process but do not appear in the texts used for training to be greater than zero[8]. The highest accuracy was achieved by a multinomial model with symmetric Dirichlet prior for the words selected with the Δ^2 statistic.

More Classifiers: Our explorations included three versions of linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) described in Ripley (1996), logistic regression described in Bishop et al. (1975) and a variant of it, a double jackknife procedure, described in Mosteller and Tukey (1968), unit-weight models described in Dawes and Corrigan (1976), support vector machines (SVM) described in Joachims (1998), κ-Nearest Neighbor described in Hastie et al. (2001), classification and regression trees (CART) as described in Breiman et al. (1984), and random forests described in Breiman (2001). We also crafted a collection of simple non-parametric classifiers, each based on the information about a single feature, and aggregated their predictions via majority voting or, alternatively, using maximum likelihood, as described in Airoldi (2003).

For each pool of words we estimated the accuracy of these classifiers on the "known" texts and recorded their accuracies. For details about these experiments, see Airoldi (2003). The accuracy corresponding to the best version of each classifier is presented in table 2 along with its standard deviation (in brackets). The notes specify which version of the classifier and which pool of words the figures correspond to.

In our experiments the major issues were: (1) the variable length of the sampling units (the texts of the speeches[9]), and (2) the sharply different number of Reagan's texts as opposed to the texts of Hannaford and others. In order to deal with the first issue we simply normalized the feature counts to the number we would have observed in a reference text of 1000 words. We applied variance-stabilizing transformations to the counts, such as $\log(X_{nij} + 0.5)$, to improve the accuracy of those algorithms that

[8] See Zhai and Lafferty (2001) for details about alternative correction methods for text data.

[9] Although the speeches were tailored for three- to five-minute broadcasts, we removed quotations and words of others to isolate the author's style, thus increasing the variability in the lengths of the texts.

Method	Cross-Validated Accuracies		Notes
	Reagan	Hannaford	
Naïve Bayes	91% (2%)	92% (7%)	Multinomial. Words selected by Δ^2.
LDA and QDA	88% (5%)	90% (17%)	Population version. SMART words.
Logistic Regression	97% (2%)	81% (15%)	Words selected by Δ^2.
Unit-Weight Models	$\approx 40\%$	$\approx 40\%$	
SVM	98% (10%)	61% (18%)	Linear kernel. Words by Δ^2.
κ-Nearest Neighbor	$\approx 40\%$	$\approx 20\%$	
CART	85% (15%)	49% (25%)	Words selected by Δ^2.
Random Forests	95% (5%)	75% (15%)	Words selected by Δ^2.
Majority Voting	85% (6%)	70% (19%)	Histogram estimate. Words by Δ^2.
Maximum Likelihood	96% (2%)	66% (16%)	Kernel estimate. Words by Δ^2.

Table 2: Summary of data mining methods: we quoted the cross-validated accuracies and the corresponding standard deviations (in brackets) for the best classifiers. In the notes we specify which version of the classifier performed best on which pool of words.

did not need integer data. Because of unbalanced sample size, several classifiers were very accurate on Reagan texts, but performed poorly on Hannaford texts. In order to mitigate this phenomenon, for example, we artificially augmented Hannaford texts using the bootstrap, a resampling method described in Efron and Tibshirani (1998), and we used this new augmented sample to compute weights for the *real* texts, for those methods that could use weights such as the logistic regression, since using both real and re-sampled texts would introduce repeated observations and might not be desirable.[10] Last, in order to obtain more robust estimates of the cross-validated accuracies corresponding to each pool of words, and to the outcome of a selection strategy, we further pruned features on the training data of each experiment using several heuristics: we removed collinear features from log-linear models, performed step-wise selections using Akaike information criterion (AIC) in log-linear models for semantic features, and removed variables based on Welch approximate t-tests with FDR correction, based on information gain arguments, and based on the statistic Δ^2 with FDR correction.

In conclusion, normalizing the feature counts to what would have been observed in a reference text of 1000 words seems a reasonable solution to deal with texts of different lengths; however, the problem of unbalanced sample sizes for the different authors remains and appears to be unavoidable. Ways around it are available for performing specific tasks; for example, the Welch approximate t-test and the Δ^2 statistic naturally allow for different sample sizes. In general, unbalanced sample sizes (recall also the lack of clustering observed above) seem to damage the performance of those classifiers that borrow their strength from some notion of geometric closeness of the texts in a high-dimensional euclidean space (κ-Nearest Neighbor, SVM) whereas those classifiers based on generative statistical models for the texts perform better (naïve Bayes, logistic regression). Further, because of the few texts available for authors other than Reagan, it is wise not to use methods that learn dependencies among words, both in terms of

[10] The figures in table 2 were obtained without any data augmentation.

the variance-covariance matrix of their corresponding vector representations, (QDA) and in terms of cascading conditional rules (random forests). In light of this, assuming pairwise independence of words (I_1) seems more than a practical approximation to the inference problem with sparse data (Mosteller and Wallace 1964, 1984; Blei et al. 2003; Erosheva et al. 2004; Erosheva and Fienberg 2005; Airoldi et al. 2005); it is also a reasonable way around the lack of data that prevents us from fitting more complex models and believing their inferences. Finally, we note that very few (20 to 40) features, most of them high-frequency, non-contextual words, were used to perform the classification task.

3 A Fully Bayesian Model for the Radio Addresses

In this section we propose a fully Bayesian model to deal with the problem of authorship attribution. It is rooted in the pioneering work of Frederick Mosteller and David Wallace but departs from it by integrating recent developments in the data-mining literature, thus portraying a more mature analysis. We explore the extent to which the model captures relevant characteristics of the data and allows for prior information to be expressed in a natural way. We conclude with a discussion of the strategies that we adopted in order to estimate the underlying parameters on the texts of known authorship, which we used to produce predictions for those texts of uncertain authorship.

3.1 In the Footsteps of Mosteller and Wallace

Herbert Simon (1955) argued that as a text progresses, it creates a meaningful context within which words that have been used already are more likely to appear than others. Such recurrences are not independent as the multivariate Bernoulli and Multinomial models assume; they are captured by *contagious distributions*, among which is the Negative-Binomial. In the same spirit, Mosteller and Wallace (1964, 1984) modeled the choices of an author about non-contextual words not as independent of previous occurrences, but rather as indicative of his or her own personal writing style. We adapt and extend their methodology here.

Our data consists of the number of times words appeared in a set of texts. Specifically, for each of two authors ($i = 1, 2$) we have a collection of texts ($j = 1, ..., J_i$), and we represent each one as a bag of words (a random vector $X_{ij} := [\, X_{1ij}, X_{2ij}, ..., X_{Vij} \,]$), where the words (indexed by $n = 1, ..., V$) belong to one of the pools discussed above. In the following discussion we denote the observed word counts with lowercase xs.

The Likelihood. In our experiments we considered both Poisson and Negative-Binomial models for the word counts. According to the Negative-Binomial model,[11] the likelihood

[11] The Poisson model can be seen as the limit for Negative-Binomial model as $\delta \to 0$ (for fixed μ). See Johnson et al. (1992) for details.

of the set of all texts, denoted as $\ell(\{x_{ij}\}|\{\theta_{ni}\})$, is written as

$$\ell(\{x_{ij}\}|\{\theta_{ni}\}) = \prod_{n=1}^{V}\prod_{i=1}^{2}\prod_{j=1}^{J_i} Neg\text{-}Bin(x_{nij}|\omega_{ij}\mu_{ni},\, \kappa_{ni},\, \omega_{ij}\delta_{ni}), \tag{2}$$

where $\{\theta_{ni}\}$ denotes the entire set of parameters, and

$$\begin{aligned} Neg\text{-}Bin(x|\omega\mu,\, \kappa,\, \omega\delta) &= \tfrac{\Gamma(x+\kappa)}{x!\,\Gamma(\kappa)}\,(\omega\delta)^x\,(1+\omega\delta)^{-(x+\kappa)}, \quad x = 0, 1, 2, \ldots \\ &\quad \text{s.t.} \quad \omega > 0, \quad \mu > 0, \quad \kappa > 0, \quad \delta > 0, \quad \kappa\delta = \mu \end{aligned}$$

is the density of a Negative-Binomial random variable. We indexed the parameters consistently so that μ_{ni} is the Poisson rate for word n and author i, that is, the number of such words we would expect to see in any 1000 consecutive words of text; δ_{ni} is the non-Poissonness rate; $\kappa_{ni} := \frac{\mu_{ni}}{\delta_{ni}}$ is a redundant parameter that will be useful for some derivations; and ω_{ij} is the word length of a document expressed in thousands of words.

In the parameterization in terms of $(\mu_{ni},\, \delta_{ni},\, \kappa_{ni})$ we used for the Negative-Binomial model, δ seemed stable across words and authors—mostly $\delta_{ni} \in [0,\, 0.75]$ with some heavy tails. Such heavy tails in the non-Poissonness parameter δ are mostly due to personal pronouns, but we included them in the analysis nonetheless since they make good discriminators. In order to use a simple prior for δ_i we used a variance stabilizing transformation to reduce the heavy tails as in $\zeta_{ni} = \log(1+\delta_{ni})$. Assume that $\delta_{n1} = \delta_{n2}$ is satisfactory for most function words but not for low frequency markers. Even though differential non-Poissonness is potentially discriminating, our actual motivation for the choice of modeling possibly distinct δ_{ni} was to avoid upsetting the analysis.

Reparameterization and Prior Study. Eventually we introduced a different parameterization that allows for prior information about the differential use of words by two authors to be expressed in a natural way, and for this differential information to be captured by priors with a simple functional form. From $\theta_n = (\mu_{n1}, \mu_{n2}, \delta_{n1}, \delta_{n2})$ we switch to $\theta_n = (\sigma_n, \tau_n, \xi_n, \eta_n)$. In order to separate the average rate of use of a word n from a comparison between the rates themselves for Reagan and the alternative author, we introduced the parameters $(\sigma_n,\, \tau_n)$, where

$$\sigma_n = \mu_{n1} + \mu_{n2}, \quad \text{and} \quad \tau_n = \frac{\mu_{n1}}{\mu_{n1} + \mu_{n2}}.$$

Recall that we defined $\zeta_{ni} = \log(1+\delta_{ni})$ to reduce the heavy tails of the non-Poissonness parameters δ_i. We eventually transformed $\zeta_{n1},\, \zeta_{n2}$ into $(\xi_n,\, \eta_n)$, so

$$\xi_n = \zeta_{n1} + \zeta_{n2}, \quad \text{and} \quad \eta_n = \frac{\zeta_{n1}}{\zeta_{n1} + \zeta_{n2}},$$

where ξ_n and η_n measure combined and differential non-Poissonness respectively.

As we discussed above, estimates for the parameters underlying models for word counts suffer from the sparsity of the data. If a word did not appear in the training data, a natural, non-informative assumption is that both authors may be using such

a word in a similar fashion; the new parameterization in terms of $(\sigma, \tau, \xi, \eta)$ leads to expected values of $(0.5, 0.5)$ for (τ, η), with equal chance of deviating towards zero or one, whereas it does not lead to expectations about (σ, ξ). Mosteller and Wallace (1964, 1984) performed a study on 90 words to gather prior information about $(\sigma, \tau, \xi, \eta)$ that led to the assumption of specific functional forms for the prior distributions of these parameters, $\pi(\sigma, \tau, \xi, \eta|\beta)$, given a set of *underlying constants* β. We performed a similar study on the same set of 90 words plus an extra 30 to gather more information on what sensible prior distributions for $(\sigma, \tau, \xi, \eta)$ should look like. The 120 words we considered ranged from high to medium and low frequency, and some of them were weakly discriminating. The words we used in the experiments to study possible priors were then set aside and never used again.

In figure 2 below we show estimates[12] for $(\sigma_n, \tau_n, \xi_n, \eta_n)$ using our data for the same 90 words used by Mosteller and Wallace. The panels show that both τ and η appear approximately symmetric about 0.5, which is the value for no differential use of words, and more analysis yielded a set of Beta distributions that brackets reasonable priors for both. Due to the small variability of ξ we assumed that the prior on ξ (for which a gamma distribution turned out to be a reasonable choice) is independent of the prior from η. It was not safe to make the same assumption about (σ, τ) because of the wide range of σ, and we assumed that the variability of τ decreases as σ increases, as the left panel of figure 2 suggests. We then assumed a constant prior for σ.

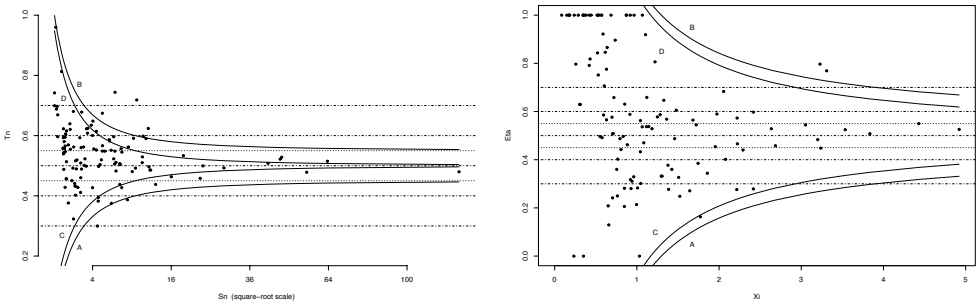

Figure 2: **Left:** Sample estimates of the parameters (σ_n, τ_n) for 90 function words used by Mosteller and Wallace (1964, 1984), with high and low frequency. Curves C and D show two-standard-error bands for t_n when $\tau_n = 0.5$. Curve A shows a two-standard-error band below $\tau_n = 0.45$. Curve B shows a two-standard-error band above $\tau_n = 0.55$. **Right:** Sample estimates of the parameters (ξ_n, η_n) for 90 function words, with high and low frequency. Curves C and D show two-standard-error bands for t_n when $\eta_n = 0.5$. Curve A shows a two-standard-error band below $\eta_n = 0.45$. Curve B shows a two-standard-error band above $\eta_n = 0.55$.

As a result of the prior study we were able to specify a set of prior distributions for $(\sigma, \tau, \xi, \eta)$ in terms of a set of underlying constants B, such that the tails of the prior distribution $\pi(\sigma, \tau, \xi, \eta|\beta)$ would be within the confidence bands around the tails

[12]Estimators for these quantities are discussed below.

of the empirical distributions of the estimates of $(\sigma_n, \tau_n, \xi_n, \eta_n)$ corresponding to the 120 words in the study, for some $\beta \in B$.

Prior Distributions. For each random variable that encodes the occurrences of the n^{th} word in a certain pool we introduced the parameters $(\sigma_n, \tau_n, \xi_n, \eta_n)$. For each set of underlying constants $\beta := (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ we assume that for all words in the pool from which the V words were selected:

(P_1) the vectors $(\sigma_n, \tau_n, \xi_n, \eta_n)$ are independent across words,

(P_2) ξ_n, η_n and the pair (σ_n, τ_n) are independent from each other for each word n,

(P_3) σ_n has a χ^2 density that can be approximated by a constant,

(P_4) conditional on $\tau_n|\sigma_n$ has symmetric Beta density with parameter $(\beta_1 + \beta_2\,\sigma_n)$,

(P_5) η_n has symmetric Beta density with parameter (β_3),

(P_6) ξ_n has Gamma density with parameters $(\beta_5, \frac{\beta_4}{\beta_5})$.

We mainly used 21 sets of underlying constants β in the posterior computations for the features in all pools, for both Poisson and Negative-Binomial models, in order to perform sensitivity analysis of our predictions, and we explored up to 40 sets of underlying constants on several occasions.

Parameter Estimation. Following Mosteller and Wallace (1964, 1984) we used method of moment estimators that make use of weights to deal with different word length of texts, and their choice of weights is "optimal" at the Poisson limit.[13] The estimators we used are:

$$\begin{cases} \hat{\mu}_{ni} &= m_{ni}, \\ \hat{\delta}_{ni} &= d_{ni} = \max\left\{0, \frac{v_{ni} - m_{ni}}{m_{ni} r_i}\right\}, \end{cases}$$

where,

$$m_{ni} = \frac{\sum_j x_{nij}}{\sum_j \omega_{ij}},\ v_{ni} = \frac{1}{J_i - 1} \sum_j \omega_{ij} \left(\frac{x_{nij}}{\omega_{ij}} - m_{ni}\right)^2,\ r_i = \frac{1}{J_i - 1} \left(\sum_j \omega_{ij} - \frac{\sum_j \omega_{ij}^2}{\sum_j \omega_{ij}}\right),$$

and the estimators for $(\sigma_n, \tau_n, \xi_n, \eta_n)$ are derived in a straightforward manner by transforming the estimators above.

3.2 Checking the Assumptions

Here we briefly discuss the various independence assumptions, the goodness of fit of the Negative-Binomial and Poisson Models, and an alternative way of calibrating the prior for $(\sigma, \tau, \xi, \eta)$ based on the data.

[13] For the Poisson model, maximum likelihood estimators are available.

Independence of Words. One of the assumptions of our model (I_1) is that words are pairwise independent, i.e. $X_{nij} \perp X_{mij}$ for $n \neq m$. Since we considered words like *thus*, *that*, and *till*, that are not related to the context and tend to be separated by multiple other words in the text, this type of independence seemed reasonable at the onset. We then considered pairs of function words and further explored their independence by means of χ^2 tests; in most of the cases independence held. For the most dependent pairs like *if we* and *that it*, independence did not hold on the speeches drafted by Reagan, mainly as a result of the high number of speeches. As we considered subsamples of 250 speeches, however, independence held on average. More generally, this assumption was adopted by Mosteller and Wallace (1964, 1984) and, in several recent statistical models for textual data, Blei et al. (2003), Erosheva et al. (2004), and Airoldi et al. (2005) and is a convenient approximation given the sparseness of the data in this type of application. Finally, our aim was to produce reliable predictions for the "unknown" texts and considerations about the stability of our predictions as much as about how their out-of-sample cross-validated accuracies supported our modeling choices, and to convey the real strength of the model with respect to the standard approaches presented above.

Non-Binomiality. The assumption (I_2) is that occurrences of the same word in a text do not depend on its previous occurrences, i.e. $p(X_{nij}|\theta_i)$ does not depend on previous occurrences of the word n (as in a Bernoulli process). We considered the high-frequency words, the semantic features, the words we found with the statistics Δ^2, and the words we identified as having high information gain in discriminating Reagan from Hannaford and from other authors in general. In order to test this assumption we considered blocks of 4 adjacent sets of 200 words each; Reagan texts provided about 1600 such blocks, Hannaford texts provided about 120 such blocks, and the texts drafted by other collaborators in general provided about 180 such blocks. For each block we computed the observed frequency of occurrence and compared it to the frequency of occurrence prescribed by a Binomial distribution with a constant parameter across blocks by looking at the Binomial dispersion index discussed in Hoel (1954) that compares block-to-block variations to theoretical variability of the Binomial. In order to account for the large amount of blocks in Reagan texts, enough to make significant even relatively small differences between observed and expected counts, we sampled subsets of about 250 blocks.[14] In conclusion, the independence of occurrences of the same word (I_2) did not hold in general; the sampling scheme underlying the Negative-Binomial distribution is more appropriate (Mosteller and Wallace 1964, 1984; Church 1995).

Goodness of Fit. Overall the Negative-Binomial model captures the variability in the data better than the Poisson model. Docu-Scope features also fit the Negative-Binomial profile. We explored further the goodness of fit of these two distributions using Kolmogorov-Smirnov tests, and we summarize the results in table 3 below. We note, however, that the Kolmogorov-Smirnov two-sample tests we performed have been proposed for *continuous distributions*, and the computation of the corresponding p-values assume that the probability of repeated observation is negligible. This assumption

[14] Mosteller and Wallace (1964, 1984) used 247 blocks of 200 words each in their analysis of the *Federalist Papers*.

does not hold in our case since we have discrete data, and the result of the tests must be taken as an indication.

	Poisson Model		Negative-Binomial Model		
Pool of words	Hannaford (38 texts)	Reagan (75 texts)	Hannaford (38 texts)	Reagan (679 texts)	Reagan (75 texts)
50 highest frequency words	12 (50)	3 (50)	31 (50)	0 (50)	49 (50)
54 words in pool no.1	4 (15)	0 (17)	14 (15)	2 (17)	13 (17)
21 features in pool no.3	3 (21)	1 (21)	21 (21)	0 (21)	20 (21)
49 n-grams in pool no.4	1 (12)	0 (14)	12 (12)	2 (14)	14 (14)
27 words in pool no.6	1 (11)	0 (11)	10 (11)	1 (11)	11 (11)
31 words in pool no.7	1 (5)	0 (3)	5 (5)	0 (3)	1 (3)
27 words in pool no.9	0 (7)	0 (8)	7 (7)	2 (8)	8 (8)

Table 3: Goodness of fit of Poisson and Negative-Binomial models for various pools of words. In brackets we quote the actual number of words compared using the corresponding p-values obtained using a two-sample Kolmogorov-Smirnov test. The rightmost columns of each distribution titled Reagan (75 texts) contain the results of our tests over 100 samples of 75 texts each. We freely discarded low-frequency words—less than 8 per 10,000 words.

3.3 Bayesian Strategies for Classification

In order to attribute a text of uncertain authorship, $X_{new} := [X_{1\,new}, ..., X_{V\,new}]$, to Reagan (author 1) or to the alternative author (author 2), given the set of all texts of certain authorship, $\{X_{ij}\}$, we needed to compute the posterior odds of authorship as:

$$\begin{aligned} odds(X_{new}) &= \frac{\int Neg\text{-}Bin(X_{new} \mid \{\theta_{n1}\})\, p(\{\theta_{n1}\}|\{X_{ij}\})\, d\{\theta_{n1}\}}{\int Neg\text{-}Bin(X_{new} \mid \{\theta_{n2}\})\, p(\{\theta_{n2}\}|\{X_{ij}\})\, d\{\theta_{n2}\}} \times \frac{\pi_1}{\pi_2} \\ \text{(final odds)} &= \text{(likelihood ratio)} \times \text{(initial odds)} \end{aligned} \tag{3}$$

for any parameterization of the Negative-Binomial, $\{\theta_{ni}\}$. The initial odds in (3) allow us to introduce historical or expert beliefs about the author of the text X_{new}. The two integrations in (3) must be carried out with respect to the posterior densities $p(\{\theta_{ni}\}|\{X_{ij}\})$. This study is based on two simple approximations for these integrals of the form $\int Neg\text{-}Bin(X_{new} \mid \{\theta_{ni}\})\, p(\{\theta_{ni}\}|\{X_{ij}\})\, d\{\theta_{ni}\} \approx Neg\text{-}Bin(X_{new} \mid \widehat{\{\theta_{ni}\}})$, where $\widehat{\{\theta_{ni}\}}$ is some central value of the posterior distribution $p(\{\theta_{ni}\}|\{X_{ij}\})$. In particular, we consider the approximation at the *mode*, computed using first and second order derivatives, and the approximation at the *mean*, computed using MCMC according to a Metropolis in Gibbs sampling scheme.

Assumptions $P_1, ..., P_6$ entail that $p(\{\theta_{ni}\}|\{X_{ij}\}) = \prod_n p(\theta_{ni}|\{X_{ij}\})$; that is, we can compute the posterior approximations above for the parameters corresponding to each word n independently. Notice that $p(\theta_{ni}|\{X_{ij}\}) = \prod_{ij} Neg\text{-}Bin(X_{nij}|\theta_{ni})\, \pi(\theta_{ni}|\beta)$ depends on a specific (vector) value $\beta \in B$. We used two Bayesian strategies in order to decide which (vector) value $\beta \in B$ to use for producing the final predictions: a fully Bayesian strategy that involves sensitivity analysis, and an empirical Bayes strategy.

Full Bayes. Briefly, for each given set of underlying constants β, we used the set of texts with known author $\{X_{ij}\}$ to compute the posterior distributions of the parameters θ_{ni} for each word, and then used the posterior modes or, alternatively, the posterior means in order to approximate the posterior log-odds of authorship for the "unknown" texts, along with neutral initial odds. Out-of-sample cross-validated accuracies were computed on the "known" for 41 sets of underlying constants to address both the problem of distinguishing Reagan and Hannaford texts and that of distinguishing Reagan and other collaborators' texts. The 41 sets of β contained both the 21 "more extreme" sets of constants[15] used by Mosteller and Wallace (1964, 1984) and 20 more we chose to explore the B space. The cross-validated accuracies were stably above 90%. We note that the approximation of the log-odds at the posterior modes was not available for all $\beta \in B$ because of the vanishing of a certain matrix of second derivatives. On the contrary, the approximation of the log-odds at the posterior means was always available.

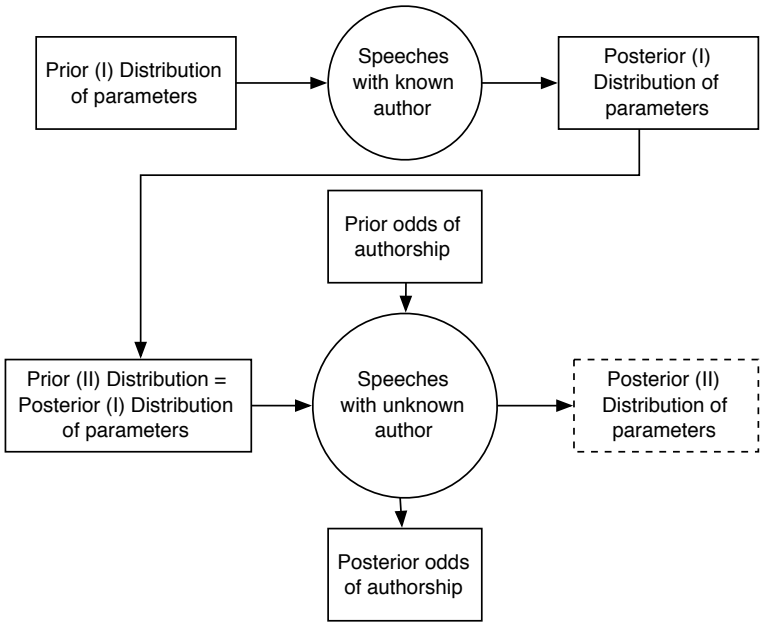

Figure 3: Graphical description of the fully Bayesian strategy for resolving the attribution of the texts of uncertain authorship. For each given set of underlying constants β, we used the set of texts with known author $\{X_{ij}\}$ to compute the posterior distributions of the parameters θ_{ni} for each word, and then used the posterior modes or, alternatively, the posterior means in order to approximate the posterior log-odds of authorship for the "unknown" texts.

Empirical Bayes for Non-Believers. Following an empirical Bayes approach we also

[15] In terms of the strength of the evidence supporting differential writing styles needed to change such prior beliefs of a certain amount.

looked for the set of constant β that is "most likely" given the model, in the sense of Mosteller and Wallace (1964, 1984). Preliminary calculations indicated the values $\beta_1 = 7$, $\beta_2 = .2$, and for the Negative-Binomial model the additional values $\beta_3 = 7$, $\beta_4 = .9$, and $\beta_5 = 1.1$. This value $\beta_* = (7, 0.2, 7, 0.9, 1.1)$ is in the middle of B, the set of underlying constants we used in the fully Bayesian analysis.

A Note on Word Selection. We also considered a different road to final word selection. For a final aggregate pool of about 170 features, we looked for combinations of words with high cross-validated accuracy disregarding concerns about computational costs. Finding the combination of words that yields the highest cross-validated accuracy among all the possible combinations is an NP-hard problem.[16] We explored the space of combinations of features using several sampling strategies based on random sampling, local search, and asymmetric random walk to find good sets of words. After running random walks for a week we found sets of words (30 to 50) that had a cross-validated accuracy in predicting the author of the "known" texts above 95 percent for several values of the underlying constants, and never below 90 percent.

4 Odds of Authorship and Predictions

In this section we empirically validate our model on the "known" texts and pursue a fully Bayesian strategy to explore how the accuracy changes across several sets of underlying constants that encode information in terms of possible scenarios about the writing styles of two authors. Then we present multiple predictions for the "unknown" texts, we explore their sensitivity to different scenarios, and we discuss validation methods to boost the confidence we have in our predictions. We conclude this section with some illustrative drafts of uncertain authorship along with our predictions for them.

4.1 Out-of-Sample Odds of Authorship

The real test for a sound model lies in its out-of-sample cross-validated accuracy.

Poisson Predictions. For the Poisson model $\beta = (\beta_1, \beta_2)$, so the number of different sets $\beta \in B$ of underlying constants in the sensitivity analysis reduces to 20 in total. In table 4 we present the cross-validated accuracies and standard deviations we obtained using this model in over 1000 experiments. The best predictions were obtained using the words selected with the Δ^2 statistic for the speeches delivered in 1975, whereas for the speeches delivered in 1976-79, the predictions obtained using words selected with information gain and the Δ^2 statistic were comparable. We also produced aggregate predictions composing by majority voting the predictions corresponding to single betas. The predictive accuracy of the Poisson model was sensible to the set of underlying constants that we used, dropping below 90 percent in some cases. Last, we compared the predictive accuracy of the modal approximation of the posterior odds of authorship to

[16] Intuitively, the drawback of such approaches is that the probability of adding or removing the right feature to the current set, at each step, decreases exponentially in the distance between the current set of features and the optimal set of features we are trying to reach by randomly moving.

True Author	Set of underlying constants (β_1, β_2) used							Voting
1975	no.1	no.2	no.5	no.6	no.12	no.16	no.20	(all sets)
Reagan (136)	117.0	120.8	112.5	111.5	114.1	111.6	118.6	112.5
Std. Dev.	3.6	2.9	3.8	4.3	3.9	4.2	3.3	3.9
Others (14)	12.2	12.4	12.5	12.5	11.8	11.6	12.0	12.5
Std. Dev.	1.6	1.5	1.4	1.5	1.6	1.7	1.8	1.4

True Author	Set of underlying constants (β_1, β_2) used							Voting
1976-79	no.1	no.2	no.5	no.6	no.12	no.16	no.20	(all sets)
Reagan (136)	117.8	122.0	112.7	112.5	115.0	111.8	119.3	115.0
Std. Dev.	2.79	2.90	2.87	2.16	2.49	3.66	2.49	2.49
Hannaford (8)	6.9	6.9	7.0	7.1	6.4	6.5	6.6	6.4
Std. Dev.	0.70	0.54	0.63	0.54	0.92	0.92	0.66	0.92

Table 4: Out-of-sample cross-validated number of texts correctly predicted using the Poisson model in over 1000 experiments. We quote the average accuracies and standard deviations on the "known" speeches in 1975 (top) and in 1976-79 (bottom) for seven sets of underlying constants $\beta = (\beta_1, \beta_2)$, and for all sets $\beta \in B$ (rightmost column), aggregating the predictions by majority voting. The predictions for 1975 were obtained using the words selected with the Δ^2 statistic, whereas those for 1976-79 used the words selected with information gain.

the approximation at the posterior mean for all sets $\beta \in B$, and for all pools of features, to discover that they were comparable in the Poisson case. The approximations at the posterior means were obtained using a Gibbs sampler with Metropolis steps.

Negative-Binomial Predictions. The Negative-Binomial model was more accurate than the Poisson model, thus justifying the increase in the complexity of the analysis. Mosteller and Wallace (1964, 1984) did not use out-of-sample cross-validation to assess the predictive accuracy; instead, they trained and tested their models on the same set of texts. The accuracy measured in such a way is referred to as *apparent* in the recent statistical literature to stress the fact that it is optimistically biased. However, such bias depends on how fast the model learns the correct values of the parameters given increasing sizes of training data. The remarkable fit of the Negative-Binomial model keeps the apparent accuracy in the same ballpark as the out-of-sample cross-validated accuracy. In table 5 we show that the apparent accuracy drops by half to 3 points of a percent if cross-validation is performed; the drop is less than a point when the odds of authorship are approximated at the posterior means and the predictions are accurate.

In general, analyzing the predictions for the speeches delivered in 1975-79 with known authorship over all sets of underlying constants, we recorded apparent accuracies between 92 percent and 99 percent, and out-of-sample cross-validated accuracies between as high as 91 percent and 95 percent, with a low 89 percent corresponding to the set β no.2 (see table 5 above). We note that the approximations of the odds at the posterior means led to more accurate predictions, from 1 to 4 percent, than the corresponding approximations at the posterior modes. Moreover, the posterior modes could not be computed for some combinations of words and underlying constants because of rank

True Author 1976-79	Type	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used no.2	no.20
Reagan (136)	apparent	125	130
	cross-val.	121.1	129.3
Std. Dev.		7	9.4
Others (14)	apparent	14	13
	cross-val.	12.5	12.8
Std. Dev.		1.9	1.3

Table 5: Apparent and cross-validated number of texts correctly predicted using the Negative-Binomial model for the "known" speeches in 1976-79. Averages over 1000 experiments.

deficiencies of a certain matrix of second derivatives, thus introducing non-controllable instabilities into a fully automated process. In table 6, we present the apparent accuracies we obtained using the Negative Binomial model in over 1000 experiments. The best predictions were obtained using the words selected with the Δ^2 statistic for all speeches delivered over the years 1975-79. We also produced aggregate predictions (in the rightmost column) composing by majority voting the predictions corresponding to single betas.

True Author 1975	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used no.5	no.6	no.8	no.9	no.10	no.11	no.12	no.14	Voting (all sets)
Reagan (679)	624	631	633	630	630	648	613	621	628
Others (69)	68	68	66	67	66	56	67	66	68

True Author 1976-79	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used no.1	no.2	no.6	no.8	no.11	no.12	no.17	no.20	Voting (all sets)
Reagan (679)	640	663*	633	648*	649	624	630*	651*	633
Hannaford (42)	39	39*	41	40*	40	40	40*	40*	40

Table 6: Number of texts correctly predicted using the Negative-Binomial model in over 1000 experiments. We quote the average accuracy on the "known" speeches in 1975 (top) and in 1976-79 (bottom) for eight sets of underlying constants $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$, and for all sets $\beta \in B$ (rightmost column), aggregating the predictions by majority voting. All predictions, 1975-79, were obtained using the words selected with the Δ^2 statistic. An asterisk indicates that the predictions were obtained using the mean approximations of the posterior odds.

4.2 Predictions for the Speeches of Uncertain Authorship

We produced Negative-Binomial and Poisson predictions to resolve the attributions for those speeches of uncertain authorship. Further, we produced predictions for the best off-the-shelf classifiers: the logistic regression and the naïve Bayes based on the Multinomial distribution. The parameter values were estimated at the posterior means using the information in the "known" texts for all sets of underlying constants, for the

words selected with the Δ^2 statistic, and for the semantic features.

Multiple Predictions. As a measure of the goodness of our predictions, we present a three-way table that displays the degree of agreement among our classifiers on the speeches of uncertain authorship. In table 7 we compose the predictions obtained from the fully Bayesian models using several sets of underlying constants with the predictions of the Multinomial naïve Bayes and the logistic regression classifiers. Notice that more speeches are assigned to Ronald Reagan by the logistic regression classifier than by the Bayesian models. Nonetheless the three classifiers all agree on 207 out of 312 speeches (66.3 percent).

	Poisson full Bayes (β no.1,4,8)			
	Hannaford		Reagan	
Multinomial	Logistic Regression		Logistic Regression	
naïve Bayes	Hannaford	Reagan	Hannaford	Reagan
Hannaford	53	31	26	8
Reagan	8	10	21	154

Table 7: Agreement of unweighted logistic regression, Multinomial naïve Bayes classifier with uniform prior for authorship and Dirichlet smoothing for unseen words, and Poisson Bayesian model using sets of underlying constants β no.1 to no.4, and no.8. These predictions were obtained using the words selected with the statistic Δ^2 on the "known" speeches.

4.3 Reading the Texts

To validate our findings, we read the texts of the speeches of uncertain authorship to look for more subjective hints that could verify the correctness of our predictions. We present two examples here that are representative of the speeches in 1975 and in 1976. We present the full text of these examples in appendices at the end of the paper.

The first example is speech number 75-02-A5 called "Rocky's Story," delivered in 1975. This commentary tells about people who helped an airline passenger who was taking his son to a hospital; he lost his wallet and the crew and passengers collected funds to help. Although Reagan usually addressed policy issues, he did quite a few commentaries on stories that simply demonstrated the goodness of people—and this sounds like quintessential Reagan to us. The odds are in favor of Reagan for all sets $\beta \in B$: the Negative-Binomial model gives Reagan 37 to 1 using the words, and 1.3 to 1 using the semantic features; the Poisson model gives Reagan 1.6 million to 1. Further, there is complete agreement among the remaining classifiers that Reagan is the most probable author of this draft.

The second example is speech number 76-01-A2 called "Platforms A," delivered in 1976. This is the second of a series of four commentaries Reagan recorded on September 1, 1976, just after he lost the 1976 nomination to Gerald Ford. In the first handwritten series he talks about his campaign for the nomination and party platforms: "... they

make specific proposals as they should and I'm going to tell you about them." "Platforms A" is an overview of the Democratic and Republican party platforms. It concludes with a preview of the next radio address: "Tomorrow I'll start with welfare." In "Platforms B" and "Platforms C"—both in Reagan's own hand—Reagan compares the platforms of the two parties on specific domestic and foreign issues, including welfare. The substance of the four radio addresses supports the finding in this paper that "Platforms A" was drafted by Reagan himself. The odds are in favor of Reagan for all sets $\beta \in B$ with some exceptions for the semantic features: the Negative-Binomial model gives Reagan 400,000 to 1 using the words; the Poisson model gives Reagan 3 million to 1. The Negative-Binomial odds for the semantic features range from 2 to 1 in favor of Hannaford to 1.8 to 1 in favor of Reagan, depending on the set β we consider, with average odds 1.1 to 1 in favor of Hannaford. Further, there is complete agreement among the remaining classifiers that Reagan is the most probable author of the draft for this speech.

5 Conclusions

The aims of this study were to determine the authorship of 312 of Ronald Reagan's 1970s radio broadcasts for which no direct evidence of authorship is available, and to provide an assessment of the confidence we have in the predictions of authorship. We used the study of *The Federalist* papers by Frederick Mosteller and David Wallace (1964, 1984) as a starting point for our modeling choices of word count data. From them we learned about heuristics for selecting features based on Δ^2, about possible parameterizations and related estimation issues for Negative-Binomial counts when the sampling units (the texts) have different lengths, and we learned how to "bracket" the prior distributions, using several sets of underlying constants. Then we fully explored the distributions of Δ^2 based on the Poisson and Negative-Binomial models to properly address the selection of features as a multiple testing problem, and we used both an ad hoc word counts analysis and a semantic decomposition of the speeches to create features able to capture elements of literary style beyond those affected by the frequency of function words, thus adding robustness to our predictions. Finally, we cross-validated the accuracies of the fully Bayesian models and assessed the goodness of the approximations of the log-odds at the posterior mode and at the posterior mean. We also compared our results with standard solutions to authorship attribution problems from both the linguistic and the computer science communities, and we concluded that in 1975, Ronald Reagan drafted 77 speeches and his collaborators drafted 71, whereas over the years 1976-79, Reagan drafted 90 speeches and Hannaford drafted 74.

Some highlights of our analyses and assessments are:

1. The goodness-of-fit study indicated that the Negative-Binomial model is appropriate for word counts and semantic features counts data, and we based both our best word selection scheme, through thresholds for the statistic Δ^2, and the likelihood of the data upon it.

2. We chose the constants underlying the prior distributions with the aim of miti-

gating the variations in the use of words that would play a role in the attribution of authorship. We ran our experiments for 21 sets of constants, entailing possible scenarios that we identified as "reasonable" with two small studies on 90 and 120 words, on speeches drafted by Ronald Reagan and other collaborators.

3. The remarkable descriptive power of the Negative-Binomial model fully translated into predictive power. The predictions we obtained with the fully Bayesian Negative-Binomial model were very stable, both in terms of cross-validated accuracy across 21 sets of constants and in terms of predicted authorship for the 312 "unknown" speeches.

4. We provided separate models for the speeches in 1975 and those in 1976-1979 and obtained stable and accurate predictions on speeches given in different years about various topics.

5. The magnitude of the log-odds of authorship entailed clear-cut predictions for the authorship of many of the "unknown" speeches. Further, the bold agreement of several accurate classification methods, based on both the analysis of words and a semantic decomposition of the speeches, reinforced our confidence.

A major shortcut that we used in our models was the assumption of the independence of words, one from another. While this presumes the absence of syntax and cannot be true in general, it produced a reasonable first-order approximation because we focused only on high frequency, non-contextual words. Nonetheless, however approximate the assumption, because of our reliance on out-of-sample cross-validation, the results of its application are not overstatements or misrepresentations. Rather, the assumptions relating to independence only result at worst in poorer accuracy than that we might achieve if we had captured dependence appropriately. In particular, a more desirable model would account for some functional form of dependence by assuming, for example, "attraction and repulsion" among words along the lines of Beeferman et al. (1997) or, alternatively, it would learn those dependencies among words relevant for predicting the author directly from the texts, along the lines of Airoldi et al. (2006).

The Poisson and Negative-Binomial models along with the fully Bayesian analysis we carried out led to cross-validated accuracies above 90 percent in all cases when tested on the "known" speeches, and predicted authors for the "unknown" speeches stable across many possible scenarios. Our confidence in our predictions was strengthened by measuring a strong agreement with several exploratory models and by our reading of the texts of the drafts for a more subjective assessment. In conclusion, out of the 312 speeches of uncertain authorship, we predicted that 167 had been drafted by Reagan in his own hand.

When Ronald Reagan died last year following a prolonged illness with Alz-heimer's disease, nearly every commentator described him as "the great communicator." But he did more than communicate. Our study shows that Reagan wrote the vast majority of his radio addresses were written by him in a form designed to convey deeply held beliefs and convictions. In this sense, our study contributes to the understanding of the man and the president he became.

Appendix A: 75-02-A2: "Rocky's Story"

I wonder if you are as fed up as I am with all the prophets of doom who have us living in a sick society, wrapped up in selfish materialism? They sound like a line from the poem "The Shooting of Dan McGrew." You know the line:

QQQ[17] "half-dead things in a half-dead world clean mad for the muck called gold."

Well, as the song says, "it ain't necessarily so."

There are any number of stories tucked away in the middle pages of our daily papers, and even more that don't get printed at all, about an America that is all around us every day—a different America than the one described by the doom criers.

A journalism student learns that news stories are based on "who, what, or where." For example, a "who" story is one that is news because the person involved is famous or notorious. A "where" story doesn't have to involve a famous person to be news—it's the location that makes it newsworthy. A young man drives his car through a gate. He wasn't a well-known figure and hitting a gate isn't all that unusual, but it's a story if it happened to be the White House gate. The "what" story, of course, is about something that's happened. Some weeks ago, the A.P. carried a "what" story. Larry Stewart, a disabled construction worker, took a plane from Detroit to Chicago with his three-year old son, Rocky. He was taking him to a hospital there. Rocky is unable to control the muscles in his legs.

When they arrived at O'Hare Airport, Larry discovered his billfold with $80.00 was missing. He didn't have money to get to the hospital. Stewardess Marsha Greiger was the first to learn of his plight. A hasty search of the plane failed to turn up the wallet. Before the plane continued on its way to Phoenix, Arizona, the stewardess gave Larry Stewart $27 and a half the crew had contributed for cab fare to the hospital. When she gave Larry the money he didn't want to take it, and only did so if he could have the names and addresses of the crew in order to repay them.

After the plane took off for Arizona, Miss Grieger asked some of the passengers to look around their seats for the missing wallet. They learned the whole story about Larry and Rocky and the missing money. One of the passengers suggested a collection. Almost instantly Marsha had $150.00 in her hand. Then a man on the other side of the plane wanted to know what was going on. When he heard the story, he started collecting. A nine-year-old boy drew a get well card, others asked for Rocky's address so they could write to him. The passengers were clapping and cheering. As Marsha said,

QQQ "It was beautiful but crazy."

On the following Saturday, stewardess Marsha Grieger was back in the Midwest. She presented Rocky's father with a set of stewardess wings, a chief pilot's wings, a letter to Rocky that said,

[17] A line that starts with QQQ in the text of a speech denotes a quotation. Such lines were excluded in our analyses.

QQQ "Someday have your daddy tell you how eleven of us (the crew) and 240 passengers fell in love with a little guy most have not even seen. In the meantime, if you ever get the feeling that the whole world is bad and no one cares, give these wings a long look."

There was $426 in the letter.

Rocky's father, holding back his tears, said it all:

QQQ "I just forgot that people are like that."

This is Ronald Reagan. Thanks for listening.

Appendix B: 76-01-A2: "Platforms A"

We've all tended to be cynical about party platforms and with good reason. All too often they represented nothing more than generalities expressing in a way overall party philosophy, but watered-down to compromise the differences between the factions within each party so as to present something of a show of unity.

Maybe we're improving or perhaps the two major parties are polarizing as to philosophy. At any rate, the voters in this election should look closely at the platforms, for they give a distinct choice as to methods for resolving our problems.

There is no question but that the Democratic leadership tried to express its true philosophy in its platform. Curiously enough the Republican platform reflects the grass roots sentiment of Republicans. The national committee's original draft appeared to be the old-fashioned idea I've mentioned of platitudinous generalities. Then the convention committee made up of rank and file members from throughout the country had its say and the changes were drastic, to say the least.

Obviously time won't permit a reading of the complete platforms. That would require about 20 of these sessions so I'll do some summarizing, no editorializing, and I'll do my best to honestly report the facts.

On the economy, the Democratic platform says,

QQQ "the Democratic Party is committed to the right of all adult Americans willing, able, and seeking work to have opportunities for useful jobs at living wages."

The platform then advocates government employment plans and

QQQ "direct government involvement in wage and price decisions,"

QQQ "which may be required to insure price stability."

It also calls for making the Federal Reserve a full partner in national economy decisions.

The Republican platform says,

QQQ "If we are permanently to eliminate high unemployment, it is essential to

protect the integrity of our money. This means putting an end to deficit spending."

It opposes wage and price controls, supports the independence of the Federal reserve system and rejects government jobs as an answer to unemployment.

On labor, the Democratic platform seeks repeal of Section 14-B of Taft, Hartley canceling out the right of states to pass "right-to-work laws." The platform also supports common site picketing.

The Republican platform favors keeping 14-B and opposes common site picketing.

On taxes, the Democrats pledge a complete overhaul of the tax system to ensure that

QQQ "all special tax provisions are distributed equally."

They pledge also to reduce the use of unjustified tax shelters in such areas as oil and gas, tax loss farming and real estate.

The Republican platform says,

QQQ "The best tax reform is tax reduction."

It then supports policies to ensure

QQQ "job producing expansion of our economy,"

more capital investment and an end to double taxation of dividends.

The Democratic platform urges breaking up the oil companies and barring them from owning other kinds of energy such as coal. It advocates a minimal dependence on nuclear energy. The Republicans oppose breaking up the oil companies and urge elimination of price controls on oil and newly discovered natural gas in order to increase supplies. Their platform also favors increased use of nuclear energy through processes that have proven safe.

Tomorrow, I'll start with welfare.

This is Ronald Reagan. Thanks for listening.

References

Airoldi, E. M. (2003). "Who wrote Ronald Reagan's radio addresses?" Technical Report CMU-STAT-03-789, Carnegie Mellon University. 293, 294, 299

Airoldi, E. M., Bai, X., and Padman, R. (2006). "Sentiment extraction from unstructured texts: Markov blankets and meta-heuristic search." In *Lecture Notes in Computer Science*. Springer-Verlag. To appear. 313

Airoldi, E. M., Cohen, W. W., and Fienberg, S. E. (2005). "Bayesian methods for frequent terms in text: Models of contagion and the Δ^2 statistic." In *Proceedings of the Classification Society of North America and INTERFACE Annual Meetings*. 293, 301, 305

Beeferman, D., Berger, A., and Lafferty, J. (1997). "A model of lexical attraction and repulsion." In Cohen, P. R. and Wahlster, W. (eds.), *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 373–380. 313

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B*, 57: 289–300. 294

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis. Theory and practice*. MIT Press. 299

Blei, D., Ng, A., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993–1022. 301, 305

Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1): 5–32. 299

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group. 299

Burrows, J. F. (1992). "Not unless you ask it nicely: The interpretative nexus between analysis and information." *Literary and Lingustic Computing*, 7: 91–109. 292, 296

Church, K. W. (1995). "One term or two?" In *Proceedings of ACM SIGIR, 18th Conference on Research and Development in Information Retrieval*, 310–318. 299, 305

Collins, J. and Kaufer, D. F. (2001). "Docu-Scope: a Java application for statistical literary style modeling." Technical report, Carnegie Mellon University. 293, 295

Dawes, R. and Corrigan, B. (1976). "A unit weighted model for discriminating the *Federalist Papers*." Technical Report 5, Oregon Research Institute. 299

De Morgan, A. (1872). *Budget of Paradoxes*. London Longmans, Green. 291

Efron, B. and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Chapman & Hall. 300

Erosheva, E. A. and Fienberg, S. E. (2005). "Bayesian mixed membership models for soft clustering and classification." In Weihs, C. and Gaul, W. (eds.), *Classification–The Ubiquitous Challenge*, 11–26. Springer-Verlag. 301

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. (2004). "Mixed-membership models of scientific publications." *Proceedings of the National Academy of Sciences*, 101(Suppl.1): 5220–5227. 301, 305

Hastie, T., Tibshirani, R. J., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag. 297, 298, 299

Hoel, P. G. (1954). *Introduction to Mathematical Statistics*. John Wiley. 305

Joachims, T. (1998). "Text categorization with support vector machines: learning with many relevant features." In Nédellec, C. and Rouveirol, C. (eds.), *Proceedings of the 10th European Conference on Machine Learning*, 137–142. 299

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). *Univariate Discrete Distributions*. John Wiley. 301

Mendenhall, T. C. (1887). "The characteristic curves of composition." *Science*, 11: 237–249. 291

Miller, G. A. (1954). "Communication." In Stone, C. (ed.), *Annual Review of Psychology*, 137–142. Banta Publishing Company. 292

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill. 293

Mosteller, F. and Tukey, J. W. (1968). "Data analysis including statistics." In Lindsey, G. (ed.), *Handbook of Social Psychology*, volume 1, 80–203. Knopf. 293, 294, 295, 299

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley. 292, 299, 301, 305, 312

— (1984). *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer-Verlag. 292, 299, 301, 305, 312

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press. 299

Simon, H. A. (1955). "On a class of skew distribution functions." *Biometrika*, 42: 425–440. 301

Skinner, K. K., Anderson, A. G., and Anderson, M. (eds.) (2001a). *Reagan, in His Own Hand: The Writings of Ronald Reagan that Reveal his Revolutionary Vision for America*. Free Press. 290

— (2001b). *Stories in His Own Hand: The Everyday Wisdom of Ronald Reagan*. Free Press. 290

— (2004). *Reagan's Path to Victory: The Shaping of Ronald Reagan's Vision: Selected Writings*. Free Press. 290

Storey, J., Taylor, J., and Siegmund, D. (2004). "Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach." *Journal of the Royal Statistical Society, Series B*, 66: 187–205. 294

Wasserman, L. (2004). *All of Statistics*. Springer-Verlag. 293

Welch, B. L. (1938). "The significance of the difference between two means when the population variances are unequal." *Biometrika*, 29: 350–362. 293

Yule, U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press. 292

Zhai, C. and Lafferty, J. (2001). "A study of smoothing methods for language models applied to ad hoc information retrieval." In *Proceedings of ACM SIGIR, 24th Conference on Research and Development in Information Retrieval*, 334–342. 299

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press. 292

Acknowledgments

The authors would like to thank Jared Cohon, president of Carnegie Mellon University, for suggesting this study, and Robyn Dawes, Charles J. Queenan Jr. University Professor in Social and Decision Sciences at Carnegie Mellon, for advice and analysis throughout this project. We also thank anonymous reviewers for useful comments. This work was carried out when the first author was affiliated with the Department of Statistics at Carnegie Mellon University.