Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper

David J. Newman
Department of Computer Science, University of California, Irvine, CA 92697-3100. E-mail: newman@uci.edu

Sharon Block
Department of History, University of California, Irvine, CA 92697-3275. E-mail: sblock@uci.edu

We use a probabilistic mixture decomposition method to determine topics in the *Pennsylvania Gazette*, a major colonial U.S. newspaper from 1728–1800. We assess the value of several topic decomposition techniques for historical research and compare the accuracy and efficacy of various methods. After determining the topics covered by the 80,000 articles and advertisements in the entire 18th century run of the *Gazette*, we calculate how the prevalence of those topics changed over time, and give historically relevant examples of our findings. This approach reveals important information about the content of this colonial newspaper, and suggests the value of such approaches to a more complete understanding of early American print culture and society.

Introduction

With the explosion of the number of pages in the World Wide Web, there is an ever-increasing need for efficient ways to characterize, classify, and index documents. This need has driven recent research in information retrieval and indexing techniques, and made automatic indexing of text documents an essential tool. Such techniques can provide an increasingly important means of identifing and analyzing historical sources, particularly as many sources for historical research are being digitized into full-text documents.

Most indexing techniques are specifically designed for information retrieval, acting like the indices found at the back of books. The quality of such an index is typically measured by its precision/recall characteristics. Yet in this paper, we focus on topics that represent a wide range of related discussions, rather than exact words one might expect in a book's index. We are interested in the quality of the topics, rather than the precision/recall characteristics often important to information retrieval.

The basis for most information retrieval techniques is the vector space model for text data (Salton & McGill, 1983). In this model, each document in a corpus is represented by a term-frequency vector whose elements are the number of occurrences of each word in the vocabulary. Collectively, the set of these term-frequency vectors forms the document–word matrix representation of the corpus. All the methods we consider have this document–word matrix representation as the starting point. The classic information retrieval method, *tf-idf* (term-frequency inverse-document-frequency), is used in many search engines today. Despite tf-idf's popularity, it does not handle synonymy and polysemy. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) devised Latent Semantic Analysis (LSA) to address this deficiency. Their method for detecting relevant documents based on words in queries improved upon simple word matching. Their association of words with documents (what they called *semantic structure*) moves us closer to the notion of *topics*. For example, LSA allows one to compute whether two documents are topically similar, even if the two documents do not have any words in common.

There has been a huge increase in the number of historical primary sources available online.[1] Yet there has been little work done on processing, modeling, or analyzing these recently-available corpora. Previous studies of historic document collections were limited by the number of items a researcher could analyze in a reasonable amount of time. For instance, Clark and Wetherell (1989) analyzed the *Pennsylvania Gazette* by sampling less than 10% of the total number of articles in just a 33 year period. Other authors analyzed a single category of a newspaper's content, such as

---

[1]For just a few examples of such full-text online resources for colonial U.S. history, see Early American Newspapers and Evans Digital Archive http://infoweb.newsbank.com; Jefferson Digital Archive http://etext.lib.virginia.edu/jefferson/texts/; Maryland Archives Online http://www.mdarchives.state.md.us/megafile/msa/speccol/sc2900/sc2908/html/; North American Slave Narratives http://docsouth.unc.edu/neh/neh.html; Virginia Runaways http://people.uvawise.edu/runaways/.

Desrochers' (2002) article on slave advertisements in a New England newspaper. In contrast, we show how one can combine large online document collections with advanced text-processing techniques to produce—in a matter of hours—results that could take years for humans to produce. Topic modeling allows historians to validate well-known trends (such as the political effects of the American Revolution) and to uncover potentially new trends (such as the popularity of various kinds of publications sold during the 18th century).

This paper is organized as follows: In the next section, Topic Decomposition Methods, we outline and compare topic decomposition methods. The third section, Topics in the *Pennsylvania Gazette*, introduces the dataset, discusses preprocessing, and presents baseline results. We then discuss topics in the Gazette computed by different methods, and argue that the probabilistic method for topic decomposition is better for this type of historical research. We conclude this section by discussing one of the practical applications of the method: improved indexing of full-text documents. The fourth section, Extracting Information From pLSA Results, then explores a variety of results derivable from the probabilistic model. We present posterior probabilities, topic hierarchies, and topic trends over time and discuss the relevance of each for historical research. We conclude with a summary of our findings and of the uses of probabilistic topic decomposition for historical research.

Topic Decomposition Methods

Here we discuss several fundamentally different topic decomposition methods. The first is Latent Semantic Analysis. Deerwester et al. (1990) introduced Latent Semantic Analysis, which is based on a singular-value decomposition of the document–word matrix. Deerwester found that LSA was superior to simple term matching for information retrieval. LSA, a type of principal component analysis, goes beyond simple word matching by semantically matching the structure embedded in a collection of documents. Berry, Drmac, and Jessup (1999) built on Deerwester's work by giving a linear algebra framework for information retrieval and addressing LSA's problem with handling dynamic document collections. We point out that LSA was devised for information retrieval, and Deerwester's paper contained no mention of topic decomposition. But we include it here for comparison purposes because the authors for the other methods we use gave detailed comparisons to LSA. In fact, all the methods we use produce analogous quantities, so intercomparisons are appropriate. Furthermore, LSA is mathematically very similar to Lee and Sung's Non-Negative Matrix Factorization (1999), which has been applied to topic decomposition.

Another topic decomposition method is *k-means* (Duda & Hart, 1973), a fundamental technique from the field of clustering and unsupervised learning. Dhillon and Modha (2001) gave some theoretical and experimental results from applying k-means to cluster text data. They compared their concept decompositions to those computed by LSA, and

discussed how k-means could be used for matrix approximation. Another distance-based clustering method is hierarchical agglomerative clustering, which has also been used on text data (Chakrabarti, 2003). Sibson (1973) clustered text data with a single-link hierarchical clustering method called SLINK.

The probabilistic version of k-means is the probabilistic mixture decomposition. The different probability distributions—in this case multinomials—correspond to the clusters. Hofmann (1999) introduced a probabilistic mixture decomposition called probabilistic latent semantic analysis (pLSA), and compared it to LSA. Blei, Ng, and Jordan (2003) devised Latent Dirichlet Allocation (LDA) for topic decomposition, another probabilistic model using a mixture of multinomials. LDA addressed some issues with Hofmann's pLSA relating to the number of parameters to be estimated and how to deal with documents outside the training set. Griffiths and Steyvers (2004) extended LDA by using Markov chain Monte Carlo to estimate parameters, and showed how one could compute, for a series of documents, how the prevalence of topics changes over time. Steyvers, Smyth, Rosen-Zvi, and Griffiths (2004) and Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004). introduced the author–topic model, a further extension of LDA that associates authors with topics, and topics with words.

Of these different topic decomposition methods, some are useful for information retrieval, some work well with dynamic document collections, and some are good for extremely large corpora such as the World Wide Web. In this paper we are interested in a static corpus (a historic newspaper) of modest size (compared to the Web). Also, we are not particularly interested in information retrieval or queries, but rather the identification of topics in the corpus. Thus, the specifics of our study and some experimental results will guide our choice of topic decomposition methods.

Comparison of Three Methods: pLSA, k-means, and LSA

The focus of this paper is the application of probabilistic Latent Semantic Analysis to find topics and topic trends in a series of text documents. We claim that probabilistic models, such as LSA, are well suited to this task. To support this claim we compare pLSA with two other well established methods: k-means and Latent Semantic Analysis. In this section we compare the details of these three methods, and later (Comparison of Results: pLSA, k-means, and LSA) we compare the results from applying these methods to the *Gazette*.

We chose Hofmann's (1999; 2001) pLSA model because of its simple formulation, simple solution, and demonstrated results. The pLSA model also has appealing symmetry: It is symmetric in words and documents. This allows us to recover likelihoods of words or documents given topics *and* likelihoods of topics given words or documents. We illustrate this below (Posterior Probabilities: Single and Multi-Topic Documents) using an example from the *Gazette*. The full details of pLSA are given in Hofmann (1999) and Hofmann (2001).

The pLSA model starts with the aspect model, where a latent topic variable $z_k \in \{z_1,\ldots,z_K\}$ is associated with each occurrence of a word in a document. The generative process is: Choose a document d_i with probability $P(d_i)$, then choose a topic z_k with probability $P(z_k|d_i)$, and finally generate a word w_j with probability $P(w_j|z_K)$. This generative process translates into a joint probability model of the form

$$P(d_i, w_j) = P(d_i)P(w_j|d_i), \quad \text{where} \qquad (1)$$

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i). \qquad (2)$$

Note that the joint probability model (1) and (2) can be rearranged using Bayes' formula into the symmetric form

$$P(d_j, w_j) = \sum_{k=1}^{K} P(z_k)P(w_j|z_k)P(d_i|z_k). \qquad (3)$$

Finally, the likelihood to be maximized is given by

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j)\log P(d_i, w_j), \qquad (4)$$

where $n(d_i,w_j)$ denotes the word frequency, i.e., the number of times word w_j occurs in document d_i. Here T is the number of topics, D is the number of documents in the corpus and W is the size of the vocabulary.

Maximum likelihood estimation is done using the Expectation Maximization (EM) algorithm, which is standard for latent variable models (Dempster, Laird, & Rubin, 1977). During the expectation step, the posterior probabilities $P(z|d,w)$ are computed based on the current parameter estimates. Next, the parameter estimates $P(z)$, $P(w|z)$, and $P(d|z)$ are updated given the posterior probabilities. Note that Hofmann (1999) proposed using tempered EM, an annealed version of EM, to improve generalization performance.

The k-means method we use for comparison is a slight variant on the classic k-means clustering method (Duda &

Hart, 1973). Instead of using Euclidian distance between documents and minimizing the intracluster sum of squared distances to the centroid, we use cosine similarity between documents and maximize the intercluster similarity to the centroid. Using cosine similarity as a distance measure is standard for text documents. An important feature of this method is that the real-vector representations of documents are all normalized to have unit length, so words in shorter documents carry more weight than words in longer documents. This normalization conceptually places all documents on the surface of a W-dimensional unit sphere, so the method is sometimes referred to as spherical k-means. This is basically the topic decomposition method used by Dhillon and Modha (2001). For some experiments, they applied inverse-document-frequency weighting to the terms in the document–word matrix, but they did not report any improvements over the unweighted approach.

The other topic decomposition method to which we compare pLSA is Deerwester et al.'s (1990) Latent Semantic Analysis. In LSA, the document–word matrix X is decomposed into its singular values and vectors, i.e., one computes $X = USV'$. A reduced model of the corpus can be computed by retaining the T largest singular values in S, and zeroing the remaining singular values. Calling these truncated matrices U_T, S_T and V_T', we obtain an approximation $X_T = U_T S_T V_T'$ that is optimal in the sense that the Frobenius norm of $||X - X_T||$ is minimized. For information retrieval, a query q is projected into the latent semantic space to generate a pseudodocument $q^* = S^{-1}V'q$. Matching documents are those corresponding to where the columns of U are close to q^*. Including LSA provides a nice comparison because both Hofmann (pLSA) and Dhillon and Modha (2001) (spherical k-means) compare their methods to LSA. We will be able to contrast their comparisons to LSA with our own comparisons to LSA using the *Gazette* data.

Table 1 compares various features of the three topic decomposition methods. Here, *NNZ* is the number of nonzero entries in the document–word matrix X. The most fundamental difference between the methods is the objective function: pLSA maximizes likelihood, k-means maximizes

TABLE 1. Comparison of topic decomposition methods used.

	pLSA	k-means	LSA
Objective function	maximize likelihood	maximize intercluster cosine similarity	minimize Frobenius norm of approximation error
Pseudocode	while (not converged) do - E-step - M-step - compute likelihood	while (centroids moving) do - assign labels - recompute centroids - compute similarity	X = U S V' computed using block power method
Global or local method?	local	local	global
Space complexity	O(T*NNZ)	O(NNZ + T*W)	O(NNZ + T*W + T*D)
Time complexity	O(T*NNZ) per iteration	O(T*NNZ) per iteration	O(T*NNZ) per iteration
Size of topic	P(z)	count of each label	S
Words in topic	P(w|z)	centroid	V
Documents in topic	P(d|z)	label	U

Note. LSA = latent semantic analysis; pLSA = probabilistic latent semantic analysis.

cluster coherence, and LSA minimizes the Frobenius norm of the approximation error of the document–word matrix (i.e., it maximizes the variance explained by T factors). The algorithms for computing the parameters for pLSA and k-means are similar, performing iterations of two separate steps until some convergence criterion is met. Both these methods only compute locally optimal solutions: To improve the results one picks the best result (based on the objective function) from a series of randomly initialized calculations. In practice the LSA calculation is also iterative, typically solved using the Arnoldi or Lanczos method, because only the T-largest singular values are sought. It is, however, deterministic, unlike pLSA and k-means.

The space complexity for k-means and LSA is similar, but significantly smaller than the space complexity for pLSA, which is linear in the number of topics times the number of nonzero elements in X. This $O(T*NNZ)$ storage for $p(z|d,w)$ in pLSA is not trivial: The number of nonzero entries in the document–word matrix for the *Gazette* is nearly 10 million, so computing, say, 100 topics requires an array of nearly 10 GB to store these posterior probabilities. The memory requirements for k-means and LSA are typically orders of magnitude less. The time complexity for all three methods is the same. In practice even the largest computations only took a few hours. We programmed pLSA and k-means using ANSI C, without the need for any specific libraries, software packages, or environments. We implemented LSA using MATLAB.

The results produced by the three methods are somewhat analogous. The topic sizes for pLSA, k-means, and LSA are given, respectively, by $P(z)$, the count of each label, and the T largest singular values, S. All of these quantities are vectors of T real values. For comparison, these quantities can be normalized to sum to unity. Next, the most important words in, say, the tth topic, are given by $P(w|z = t)$, centroid(t), and V_t respectively, each a vector of W real values. Note different normalizations: The conditional probabilities $P(w|z = t)$ sum to unity, whereas the L_2-norm of centroid(t) and V_t sum to unity. Further differences are present: The probabilities and centroids are all nonnegative, whereas the singular vector V_t can have positive or negative values. Coccaro and Jurafsky (1998) suggest an ad hoc way of interpreting these singular vectors as probabilities. In this paper, however, we will simply consider the order of the terms, after sorting by absolute value, to compare results. Lee and Sung (1999) avoid the negative numbers in the singular vectors U and V of $X = USV'$ by computing an approximate factorization $X \approx UV$ that by construction guarantees that U and V are nonnegative. Finally, the most important documents in the tth topic are given by $P(d|z = t)$, where label $= t$, and U_t, respectively, for the three methods. The probability distribution and left singular vector have D terms, whereas the label is just a single number. The fewer degrees of freedom in this parameter for k-means is due to the classification of a document to a single topic: The k-means formulation does not allow a document to be associated with more than one topic.

We finally mention that measures between topics differ for the three methods. The Kullback-Lieber distance to measure the distance between two word distributions $P(w|z = t_1)$ and $P(w|z = t_2)$ is always negative. The inner product or cosine of the angle between two centroids, centroid (k_1) and centroid (k_2), is always positive, whereas the inner product between any two singular vectors is always zero, as singular vectors are always orthogonal.

From the outset, the probabilistic model pLSA has some desirable features for our study. Not only are various likelihoods computed: $P(z)$, $P(w|z)$ and $P(d|z)$, but also the posterior probabilities $P(z|w)$ and $P(z|d)$ are available via Bayes' formula. These quantities are helpful for validation (one can check that likely documents do indeed make up a topic), and also for answering specific questions such as: "What is the most likely topic to generate a given word?" and "What mix of topics is this document made from?" Having pLSA compute these conditional probabilities allows us to formally calculate these posterior distributions, which is not straightforward for k-means or LSA. Blei et al.'s (2003) concern with pLSA was twofold: The number of parameters grows linearly with the size of the corpus, and there is no clear way to assign probability to a document outside of the training set. For our study, neither of these concerns is an issue. Because we are analyzing a historical newspaper, the corpus is unchanging, so the number of parameters to estimate is fixed, and therefore not an issue. Furthermore, we are interested in topics, not queries, so there are no additional documents, or queries posed as pseudodocuments, outside the training set.

One commonality between all three topic decomposition methods is the problem of selecting the free parameter T, the number of topics. Theory tells us that, for all three methods, the objective function will never worsen as T is increased. So how does one select T, the number of topics? A Bayesian approach might ask: What it the most likely number of topics given the data? Other approaches penalize the objective function as T increases. Rather than focusing on the issue of how to select T, in this paper we will take a pragmatic approach, and show how different values of T can give us different levels of details about topics. One noteworthy point is that the LSA computation is independent of the number of topics T. T merely sets the number of (largest) singular values retained. Unlike pLSA or k-means, where increasing T will generally affect all the results, increasing T in LSA does not change any of the singular values or singular vectors already computed.

How do we compare the results between the three methods? Each method optimizes its respective objective function, but how do we compare a likelihood to a coherence to a Frobenius norm? Research presenting new methods will typically provide experimental results on precision/recall (query performance), perplexity (performance on a hold-out set), or entropy measures. Because we are focusing on the specific application of finding topics, we take a practical approach and see which results appear to make the most sense for our purposes. We also use some special cases to guide our intuition about which methods may be best suited to our topic/topic-trend analysis.

Topics in the *Pennsylvania Gazette*

Newspapers are one of the most important historical print sources in colonial America, and the *Pennsylvania Gazette*, published primarily by Benjamin Franklin, was one of the most important newspapers in the period (Clark and Wetherell, 1989; Aldridge, 1962). Since Accessible Archives made a full-text version of the 18th-century *Gazette* available in the past decade, historians have regularly searched the *Gazette* for specific words or phrases as part of their larger research projects (Block, 2002; Grubb, 1999). Thus, a comprehensive understanding of the content in the *Gazette* is critical to a wide range of historical research.

Preprocessing the Data

We used all the printed text from the *Gazette* from Accessible Archives, resulting in over 80,000 articles and advertisements spanning from its founding in 1728 to the end of the Accessible Archives' content in 1800. Each article is a separate html document that includes the text of the article; a date field with month, day, and year; and occasionally a field indicating a location (e.g., Boston). In addition there is a keywords field that may contain keywords describing the content of the article. The following is an example of a real estate advertisement from the *Gazette*.

March 5, 1754
The Pennsylvania Gazette
To be SOLD,

A large brick house, suitable for a merchant or Tavern keeper, two stories high, 28 feet in length, 18 feet deep, with a large kitchen, stabling, gardening, and an excellent good well; likewise a beautiful orchard, full of grafted trees, 28 yards front, and 27 rod deep, and divers other conveniences not mentioned, situated in Bordentown. Enquire of John Thorn, now in possession of the said house and land, and know the terms.

To get a vector-space representation of the run of the newspaper, we produced a list of the unique words. An initial scan of the 25 million words in the newspaper identified over 100,000 unique words. The size of this vocabulary is relatively large—some of the inflation is due to the introduction of typographical errors based on errors resulting from text transcription. (Much, if not all, of the *Pennsylvania Gazette* was transcribed before widespread usage of OCR software.) We applied standard techniques of excluding stopwords and stemming to reduce the size of this vocabulary. Instead of using a standard stopword list, we chose stopwords from the list of most frequently occurring words in the *Gazette*, to avoid potential removal of words with particular historic significance.[2] Identifying stopwords is

---

[2]For a more limited list of 34 basic stopwords in one early American historical database, see http://infoweb.newsbank.com/iw-search/we/Evans?p_action=help&f_helppage=stopwords. Because scholars use this database to search for exact phrases, rather than to cluster documents, a more limited list of stopwords is appropriate.

TABLE 2. The three-or-more letter stopwords excluded from the *Gazette* before producing the document-word matrix.

Stopwords
about after all also and any are been before being but can could each every for from had have having her here him into last made may more most much must next not now one only other our out part said same several shall should some such than that the their them then there these they this those three time two under upon very were what when where which who will with would yet you your

TABLE 3. Top-10 words in the *Gazette*, after removal of stopwords.

Word	Count
house	59195
good	55409
new	47961
john	46671
day	46543
county	45780
philadelphia	43516
state	40492
street	37582
person	35980

subjective: To facilitate efficient topic decomposition, we aimed to eliminate words that added little discrimination value, but we were also careful to not eliminate common words (i.e., *good*) that might later prove important for historical analysis of a given topic.

Table 2 lists the three-or-more-letter stopwords used. We appended this stopword list with all single-character and two-letter words, and all words that occurred fewer than six times in the entire run of the *Gazette*. We also applied basic stemming to fold-in plurals by removing any trailing *s* characters in words. After stemming and excluding the stopwords, we produced a working vocabulary of about 30,000 words. Table 3 lists the most frequently occurring words in the *Gazette* and their counts, after exclusion of stopwords. This list will be particularly relevant when we compare the different topic decomposition methods (Comparison of Results: pLSA, k-means, and LSA).

This preprocessing resulted in a document–word matrix representation of the *Gazette* with the following parameters: the number of documents, $D = 82,737$; the number of words in the vocabulary, $W = 30,838$; the number of nonzero entries in the document–word matrix, $NNZ = 7.4$ million. This gives a sparsity of 0.3%, i.e., on average 0.3% of the vocabulary appears in each document. Finally, the sum of the counts in the document–word matrix was 10.4 million. The difference between the 10.4 million words in the document–word matrix representation and the 25 million total words in the *Gazette* is due to the removal of stopwords. Note that we do not perform any explicit term-weighting (which is used in information retrieval to improve recall) because we are trying to find topics, not perform queries.
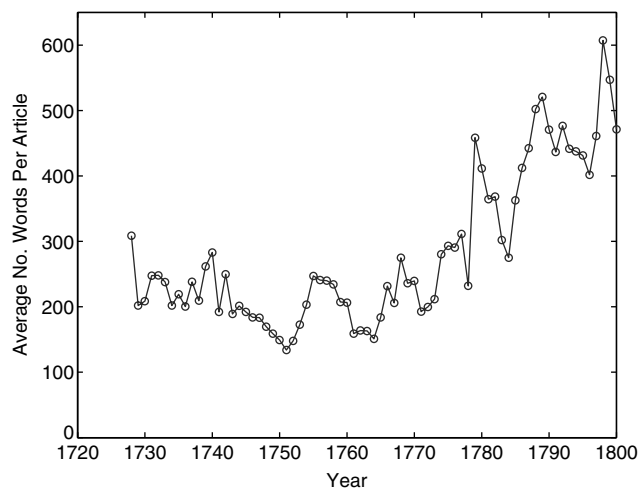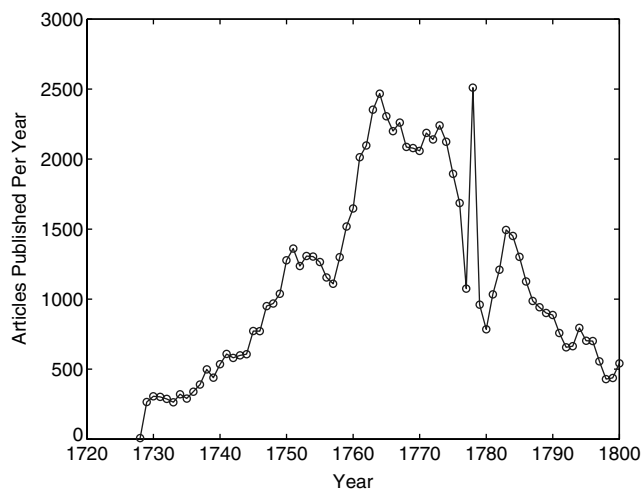
FIG. 1.   Number of articles per year in the *Gazette* (left) and average article length (right).

Having obtained the full run of the *Gazette*, we started to examine the data. The number of articles per year in the *Gazette* increased fivefold, from less than 500 in 1728 to around 2,500 in 1764. After 1764, there was a decline in this number that coincided with Franklin ending his involvement in the *Gazette* and an increasing politicization of the paper with the coming of the American Revolution (Clark & Wetherell, 1989, pp. 280–281, 282). One exception to this trend is a large anomalous spike in 1778 (Figure 1, left), when 2,500 articles were published. This spike corresponds to a similar mid-1770s spike in published colonial imprints generally (Amory and Hall, 2000, p. 507), likely due to the colonies' political upheaval. In the case of the *Gazette*, specifically, the spike was also likely due to the end of the British occupation of Philadelphia in early 1778.

As Figure 1, right, shows, the average article length doubled from about 200 words per article in the first half of the 18th century to more than 400 words per article by the end of the 18th century. This increase in article length is consistent with the increasing overall length of the newspaper and the general expansion in early American print culture, as well as increasing interest in a mainstay of the paper's coverage, politics, with the American Revolution and the formation of the United States (Clark and Wetherell, 1989).

pLSA Results With 40 Topics

We set the number of topics to $T = 40$ for the initial topic decomposition runs. This number was based on some preliminary runs, and was also a reasonably manageable number for intercomparisons with k-means and LSA. We ran the EM algorithm to compute the pLSA probabilities until the relative change in log likelihood was less than 1 in 1 million (i.e., no change in the sixth most significant digit). This took approximately 200 EM iterations. We randomly initialized and repeated this computation 20 times, and selected the run with the highest log likelihood.

The 40 topics, in order of size, are described in Table 4. Each row shows the rank of a topic, the topic size in

percentage, the most likely words in the topic, and our label to describe the topic. Note that the topic size is simply the mixing proportion or topic probability, $P(z)$, converted to a percentage. Our labeling of a topic is a subjective task, and requires a subject expert—in this case a historian—to correctly interpret the semantic meaning described by the list of most likely words. It is worth re-stating that these lists of words create topics that are not a priori prescribed by a human: They are purely the outcome of the method.

Classifying some of these topics can be obvious (i.e., laypeople would likely identify topic #2 as related to the U.S. government), whereas identifying the focus of others can require a more in-depth examination. For example, topic #3 is, perhaps, the least thematic: The individual articles in this topic are largely a series of extremely lengthy political letters written during various political conflicts by individuals and government officials, such as a letter from the Governor to the Assembly published on August 26, 1742 (ITEM #5293), or the letters, commenting on *Common Sense*, published in the December 21, 1778 issue (ITEM #63962).[3] Because these public letters were regularly upwards of 4,000 to 5,000 words (an order of magnitude larger than the average article length), they tended to have a higher number of words that related to argument-making (i.e., *think*, *reason*, *say*), rather than to one specific topic. We discuss the combination of topics in such lengthy articles below (Posterior Probabilities: Single and Multi-Topic Documents. Some topics that seem immediately incomprehensible become more understandable when we examine the individual articles that contribute to the topic. For instance, topic #15 of NAMES seems unlikely until one finds articles consisting of various published lists of Pennsylvania residents, such as a list of people for whom there were letters at the Philadelphia Post Office in the July 24, 1766 issue (ITEM #38438), or the list of elected officials in the October 7, 1742 issue (ITEM #5368).

---

[3]ITEM # is Accessible Archive's numbering scheme for articles.

TABLE 4. Most likely words in the 40-topic decomposition of the *Gazette*.

rank	Size (%)	Most likely words in topic	Our topic label
1	5.6	away reward servant old whoever named year feet jacket high paid hair pair secure coat run inches	RUNAWAY
2	5.1	state government constitution law united power citizen people public congress right legislature	GOVT–U.S.
3	4.9	say thing might think own did know against without make reason man men give people never good	LONG ARGUMT
4	4.6	good house acre sold land meadow well mile premise plantation stone containing mill dwelling orchard	REAL ESTATE
5	4.0	country america war great liberty nation people american men let cause peace enemy present state she	GOVT–REVOLT
6	3.9	silk cotton ditto white black linen cloth women blue worsted men fine thread plain coloured	CLOTH
7	3.6	governor province excellency assembly house majesty gentlemen general honour government council	GOVT-IMPERIAL
8	3.4	act person aforesaid state within enacted authority thereof county further day law hereby officer	GOVT-LEGISLAT.
9	3.3	capt ship privateer taken vessel gun french she men arrived port board sloop bound prize sail captain	PRIVATEER
10	3.2	general officer enemy army troop men regiment major colonel col soldier lieutenant militia wounded	MILITARY
11	3.0	house fire off night down got went near great came boat morning people found took wind man over	DISASTER
12	3.0	sloop brig ship schooner jamaica snow carolina boston island barbados john antigua new brigt mary	SHIPPING
13	3.0	state united france french treaty war letter nation vessel american between article king minister peace	GOVT-TREATIES
14	2.8	colony act great britain america parliament province trade right duty british liberty good subject lord	MERCANTALISM
15	2.5	john william james thomas robert samuel joseph george philadelphia alexander henry richard smith	NAMES
16	2.4	money pound per sum hundred bill year thousand interest state paid dollar shilling credit five pay debt	MONEY
17	2.4	horse reward whoever mare hand high stolen old paid year near shilling white subscriber reasonable	HORSE THEFT
18	2.4	captain day spoke long arrived lat new port london ship brig bound capt vessel york virginia	SHIP CAPTAIN
19	2.3	land sale late acre vendue containing day public sold situate sheriff tract virtue directed taken execution	LAND SALE
20	2.3	french arrived war men capt england ship day letter island fleet place new hear boston admiral advice	WAR
21	2.3	committee house read state bill report petition ordered resolved congress motion appointed united act	GOVT-ACTS
22	2.2	resolution consideration indian fort party men killed nation town day people letter creek sent cherokee	INDIAN
23	2.2	street feet lot ground house front side brick story city philadelphia north alley depth south breadth	CITY
24	2.2	many great year use well water make person method cure quantity disease place without kind health	HEALTH
25	2.0	new town york city day letter morning philadelphia boston monday evening week saturday place	CITIES
26	2.0	person estate desired account indebted against demand pay deceased make late bring payment notice	DEBT
27	1.9	church life god society great friend christian year college day good virtue religion minister character rev	RELIGION
28	1.9	land river mile tract road creek delaware water town new great acre country county near side west	LAND
29	1.8	within branch thence philadelphia city day house meeting election notice pennsylvania ticket general	PHILA. BUSINESS
30	1.7	street market sold good store second imported assortment door front just house london between	MARKET
31	1.6	county township chester lancaster buck john west james pennsylvania castle montgomery cumberland	COUNTIES
32	1.4	silver public favour watch business horse best waggon house philadelphia work sort manner depend	ARTISAN/WARES
33	1.4	owner charge property away take come came pay desired plantation old prove white subscriber year	PROPERTY
34	1.4	book published vol new price school history printing sold paper english work just office writing	BOOKS
35	1.3	year negroe she sold enquire man age printer servant likely country well woman young busines good	SERVANT/SLAVE
36	1.2	court person justice committed goal trial jury taken found called against brought murder prisoner guilty	CRIME
37	1.1	ditto sugar wine rum barrel good hogshead flour quantity tea sold salt west cask choice coffee india	FOOD & DRINK
38	1.1	thy like sun fire day round form head joy piece air thou light see great thee figure seen curious	LITERARY
39	0.9	oil glass medicine sort pot best size large powder salt lead bottle white london sold case boxes stone	MEDICINE
40	0.8	board master ship vessel sail passage commander apply freight passenger wharff good carolina lying	SAILING

Overall, we can see that the main topics covered in the *Gazette* related to economics and politics. As the newspaper became more politically focused in the mid-1760s, many of the political topics related to the coming of the American Revolution and the debates over the form of the new government of the United States. Many of the commercial topics are generated by advertisements for products such as cloth, books, or foodstuffs (topics #6, 32, 37); notices for sales of land, homes, or livestock (topics #4, 17, 19, 28); and discussions of markets and various businesses (topics #29, 30). The proliferation of these topics reveals that the *Gazette* was a crucial enabler of local and trans-Atlantic commerce. Philadelphia's economic importance as a port city is also revealed in the many topics related to shipping and ships (#9, 12, 40).

Some of the smaller topics include legal discussions of civil and criminal issues (topics #26, 36) and Native Americans (topics #22), reminding us that Pennsylvanians lived in a world bound by institutional governmental processes, but also subject to the reality of being a colonial outpost. Contrary to images of Colonial America that focus on New England's expansive religiously focused print culture, religious commentaries account for only a rather small topic (topic #27), suggesting perhaps Franklin's lesser interest in religious rhetoric, as well as Pennsylvania's religious diversity.

The largest individual topic relates to advertisements for runaway and indentured servants, revealing both the centrality of servants to Pennsylvania life, and the frequent difficulties servants had with their position as bound or indentured laborers (Salinger, 1987). Pennsylvanians also held significant numbers of slaves in this period, as evidenced by topic #35's inclusion of multiple words related to sales of slaves.

We can also look across the topics for particular words of interest. For example, we see few topics that relate primarily to women: References to *women* or *woman* occur in topic #6, in relation to clothes for sale, and in topic #35, in reference to

TABLE 5. Comparison of words in the top-6 topics computed by pLSA, k-means, and LSA. Bolded words are words that are in the 10 most frequent words in the document–word matrix.

topic #	pLSA	k-means	LSA
1	away reward servant old whoever named year feet jacket high paid hair pair secure coat run inche master	away reward servant whoever year old paid named secure run master jacket high county feet charge age reasonable	**state house day** general act **person** great **county new good** united government law people public **john** men power
2	**state** government constitution law united power citizen people public congres right legislature general convention principle	good acre **house** land sold mile plantation well meadow premise **county** containing stone **philadelphia** orchard barn mill water	**john state** william james **county** thomas sloop captain **philadelphia** ship united capt robert government **new** acre brig
3	say thing might think own did know against without make reason man men give people never **good** how	capt ship arrived **day** vessel french taken men privateer **new** she sloop board port gun captain war bound	**county john** captain ship **state** sloop aforesaid act men **person** french william arrived brig jame acre capt vessel
4	**good house** acre sold land meadow well mile premise plantation stone containing mill dwelling large orchard **philadelphia**	street sold **house** second market **philadelphia good** front between door near john best arch corner store water william	**state john** aforesaid united william **person street** jame **good** act **day** land **house** government thoma acre feet enacted
5	country america war great liberty nation people american men let cause peace enemy present **state** she power	great country people **good state** public **new** men many without well present general power government america year **person**	acre sloop aforesaid act **good** ship **street house** captain **person** land brig feet enacted **day state** authority county
6	silk cotton ditto white black linen cloth women blue worsted men fine thread plain coloured handkerchief striped lawn	land acre **county** sold tract containing **good** mile **house philadelphia** meadow sale situate township well term near **john**	**state** united acre sloop **street** governor province ship **john** brig feet colony land assembly people lot **county** schooner **good**

slave labor. The few references to *she* (topics #5, 9), refer to the feminized personification of liberty or ships, rather than to actual colonial women. Thus, discussions of women appear in the newspaper largely as political symbolism or in reference to economic commodities (including the women themselves!) for sale. In contrast, references to *man* or *men* occur in 10 places in the most frequently appearing words in these topics. The overall absence of female-focused topics suggests that even though social history scholars traditionally use newspapers such as the *Gazette* to talk about an array of topics related to women, such articles may have been a relative rarity in a paper focused far more on men's political and economic concerns.

There are a multitude of ways to examine and interpret the topics produced by the 40-word topic decomposition. Besides providing an overall view of the content of this important colonial newspaper, this method allows historians to examine particular areas of interest in greater depth. Before doing so, however, we turn to a comparison of pLSA with two other well-known topic decomposition methods.

Comparison of Results: pLSA, k-means, and LSA

We now compare the results from the three different topic decomposition methods when applied to the *Pennsylvania Gazette*. The number of topics was kept at $T = 40$, and each of the three methods was run on the identical document–word matrix representation of the *Gazette*. We have seen in some detail the results of the pLSA 40-topic decomposition in the previous section. Here we supplement our theoretical comparison of the three methods (see above, Comparison of Three Methods:pLSA, k-means, and LSA) with some experimental results computed using the *Gazette* data.

A comparison of the topic decompositions of the *Gazette* computed by pLSA, k-means, and LSA is given in Table 5. The table lists, in order of topic size, the most important words in each topic. As discussed in the Comparison of Three Methods section, we equate importance with the absolute value of the measure of words in a topic: For pLSA this is the sorted list of word likelihoods; for k-means this is the topic centroid vector, sorted by element magnitude; and for LSA this is the word (right) singular vector, sorted by element absolute magnitude.

We see some similarity between the topics computed by pLSA and those computed by k-means. The top topic for both pLSA and k-means is about runaway servants, and the top words in these two topics are similar. The second k-means topic about real estate clearly corresponds to the fourth pLSA topic. The higher k-means rank of this topic is due to k-means' normalization: Real estate advertisements are typically shorter in word length, so their relative importance is boosted. This increased importance is similar to measuring a topic's size by counting the number of documents about a topic, as opposed to counting the total number of words written about a topic. In some cases the word order within the topic differs. We attribute this to a combination of two factors: First, the k-means does not allow mixtures of topics within a document, and second, k-means' normalization weights the words. Thus, despite differences in the methods' weighting practices, we can still see several similarities between the pLSA topics and the k-means topics.

LSA topics diverge strikingly from the overlapping results of pLSA and k-means methods. For instance, the top topic for LSA is some mix of concepts, and difficult to interpret as a single topic. In fact, it is more recognizable as a partial list of the most frequently occurring words (comparing

to Table 3). The fact that the LSA topics are in general saturated with mixes of frequently occurring words is not surprising: If the largest singular value and singular vectors are to best explain the variance in the corpus, they must include a significant mean component, i.e., the most frequently occurring words. What is, perhaps, surprising are the LSA topics following the first topic. We see more mixes of concepts, and many frequently occurring words (bolded). In fact the LSA top-6 topics' top words have 36 counts of the most frequently occurring words in the corpus, compared to 6 for pLSA and 21 for k-means. Furthermore, it is difficult to explain how the words in the second LSA topic, in combination with the first LSA topic, best account for—using only two factors—the variance in the corpus. Mathematically nothing is awry: We are just seeing how the LSA gives a decomposition that is not easily interpreted as topics in text data.

The distribution of topics' sizes from the three topic decomposition methods is shown in Figure 2. With the exception of the first topic (corresponding to the largest singular value), LSA gives the flattest distribution. The disproportionately large first LSA topic, accounting for 29% of the corpus, is merely a reflection that most of the variance in the corpus is explained by the mean (i.e., the most frequently occurring words). The largest topic for pLSA (5.6%) is significantly smaller than the largest topic in k-means (9.3%). This is partly due to the total attribution of documents to topics, rather than allowing documents to be made from a mixture of topics. Of the three methods, pLSA seems to have the smoothest distribution of topics, particularly at the large end of the size spectrum.

The pLSA method appears to separate the topics more cleanly than k-means. We see the words *good*, *house*, and *philadelphia* repeated in three of the top-6 topics from k-means, whereas for pLSA the same words only occur once (topic #4). The word *county* occurs in three of the top-6 topics from k-means but does not occur in any pLSA topic. The LSA topics are the least intuitive, and the least useful for the particular problem of finding understandable topics in
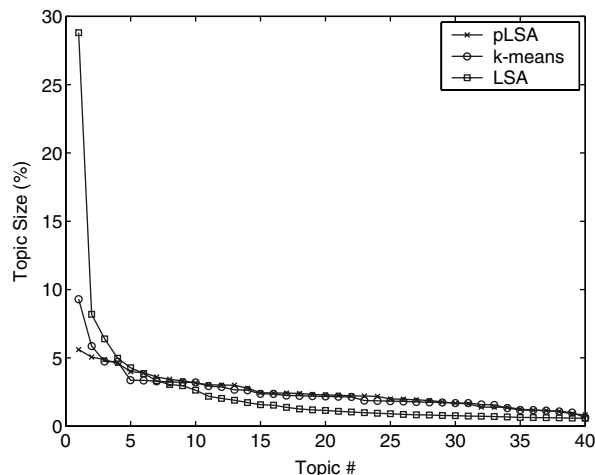
a collection of text documents. We make no claims against LSA's usefulness in performing other types of tasks such as information retrieval. In fact, LSA generally has better precision/recall characteristics than tf-idf retrieval (Chakrabarti, 2003). A possible interpretation is that LSA's topics exist in frequency space, because higher topic numbers correspond to higher frequency components of variance, and lower topic numbers correspond to lower frequency components (e.g., the first topic has large overlap with the list of frequently occurring words).

In our opinion, pLSA's topics are more useful than the topics computed by k-means, and both are significantly more useful than LSA's topics. We also believe that pLSA's ability to model a document made from a mixture of topics is the most appropriate and realistic for historical research. First, its mixing feature ultimately results in more coherent topics. Second, pLSA has an advantage over the other two methods in its consistent formulation: If one runs all three algorithms with $T = 1$ topic, pLSA is the only method to correctly return the list of most frequently used words. Because of k-means' discounting of words that occur in longer documents, it produces a slightly different list of most important words. Finally, we see LSA's invariance: Its list of most important words is the same as what is shown for LSA topic #1 in Table 5. LSA's topics do not change with the selection of T.

Both Hofmann (2001) and Dhillon and Modha (2001) compare their results with LSA results. We come to a conclusion similar to Hofmann's: that LSA appears functionally similar to pLSA with analogous parameters, but the topics produced are very different. Hofmann also points out that LSA's methodological foundation and application to count data are somewhat ad hoc. Dhillon and Modha take the opposite side, arguing that their k-means results are close to LSA in the sense that the angle between their k-means centroids and the singular vectors is small. Despite this claim, their experimental results gave LSA topics that looked quite different from their k-means topics (as we saw using the *Gazette* data). Not surprisingly, their LSA topics were, like ours, an uninterpretable mix of concepts. Their final tie-in with LSA was their suggestion that k-means can be used as a matrix approximation scheme, which is mathematically what is computed by the truncated singular value decomposition in LSA. This may be interesting, but it is not relevant to our task of finding topics.

Value of Mixture Model

Using a mixture model such as pLSA also allows flexibility in the ability to describe a document. We also see in practice that pLSA results in a cleaner delineation of topics when compared to k-means. We can compute, on average, what mix of topics a *Gazette* article is made from, and validate the need for a mixture model. For every document in the *Gazette*, we computed $P(z|d)$. We then sorted each vector of T conditional probabilities from largest to smallest, and averaged over D (the entire collection of documents) of
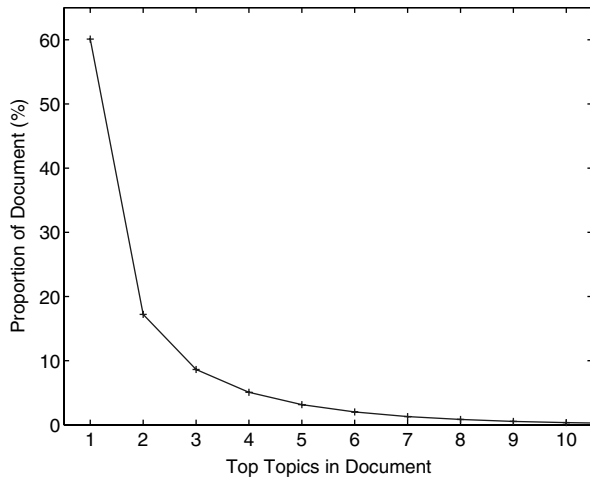

FIG. 2. Distribution of topic sizes from three topic decomposition methods.

FIG. 3.   Average mixing proportion of topics in single document.

FIG. 4.   Inconsistent use of keywords for indexing the *Gazette*.

these sorted vectors. Figure 3 shows the average mixing proportions of a *Gazette* article. For the 40-topic decomposition, on average an article consists of 60% one topic, 17% a second topic, 8% a third topic, and 15% some combination of the remaining 37 topics. This fairly broad combination of topics further highlights the argument for using a mixture model. If we restricted the document model to allow only a single topic per document, the sharpness of the topic definitions would be blurred.

Automatic Metadata Enhancement

Berry et al. (1999, p. 336) talked about the consistency problem of human-generated indexes and how "the extraction of concepts and key words from documentation can depend on the experiences and opinions of the indexer." We include the following related anecdote from our study of the *Gazette*. We decided to examine advertisements in more detail. Luckily, the keyword *adv* was the most frequently used keyword, so our search for articles with this keyword returned more than 15,000 articles. To double check the completeness of our search, we computed the most frequently used keywords. The top-3 keywords used were *adv* (15,792 articles), *real-estate* (7,296 articles), and *runaways* (4,957 articles)— 32,000 articles contained no keyword descriptions at all. This gave us an inkling of the consistency problem of keywords, because both *real-estate* and *runaways* are advertisements (*adv*). Plotting the histogram of the number of articles with these keywords over the century illustrated an additional problem (Figure 4). We see that the indexer started with the keyword *adv*, switched to the more specific categories of *real-estate* and *runaways* after 12,000 articles, then switched back to the keyword *adv* after 70,000 articles. This is perhaps a more extreme example of an inconsistent index, but it does serve as a cautionary tale.

One obvious application of computing topic decompositions is automatic metadata enhancement of text documents. To enhance the metadata for a series of text documents one
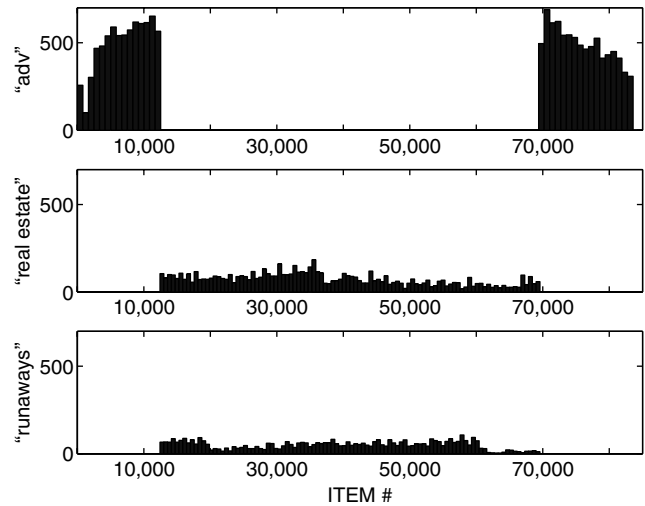
could first compute topic decompositions at some level of detail (i.e., setting *T*), assign labels to these topics, and then go back and tag the documents with these labels according to their topic membership probabilities. This would at least result in consistently generated topic headings. For historical research, which usually focuses on big analytic issues, topical indices that address general ideas and themes (e.g., consumption) would be of great potential value, rather than, say, a specific index that emphasizes a precise object or activity (e.g., book). Such a method could be used to greatly improve topical indices. For instance, Readex has recently released digitized versions of Early American Newspapers online (available by subscription at: http://infoweb.newsbank.com/). Although users can do keyword searches for specific words in the full text, there is no topical index that might allow for a more comprehensive search for articles related to a central theme. Thus, pLSA could have particular applicability to historical research.

Extracting Information From pLSA Results

After running pLSA on a corpus, several further analyses can enhance the value of the results for historical inquiry. In this section we examine the likely documents to contain a given word; study the differential results from increasing the numbers of topics; and analyze and compare trends in topics over time.

Posterior Probabilities: Single and Multi-Topic Documents

For the pLSA results we have looked at the likely words in a topic, $P(w|z)$ and the topic sizes, $P(z)$, in detail. Now, what can we learn from the likely documents given a topic, $P(d|z)$? Furthermore, what can we learn from the posterior probabilities $P(z|w)$ and $P(z|d)$? A closer look at two case studies allows us to see the usefulness of such probabilities for articles that are largely about a single topic, and in articles that are likely to be generated by multiple topics. To

begin, the 1,600 occurrences of the word medicine in the *Gazette* allow us look at the relationship between the topic of MEDICINE, the actual word *medicine*, and documents relating to medicine as a case study of the advantages and flexibility of this topic decomposition method.

From the 40 topics listed in Table 4 we see that topic #39 is about MEDICINE. We will use this topic to demonstrate one advantage of the probabilistic clustering method: the ability, via Bayes' formula, to formally compute specific likelihoods and posterior probabilities. The most likely words $P(w|z =$ MEDICINE) in the topic are: *oil, glass, medicine, sort, pot, best, size, large, powder, salt,* etc. Now let us take the *word medicine* (as opposed to the topic MEDICINE), and see which are the most likely topics to generate this word. Computing $P(z|w = medicine)$ tells us that the topic MEDICINE (topic #39 in the Table 4) has a probability of 0.77 of producing the word medicine, and the topic HEALTH (topic #24) has a probability of 0.23 of producing the word medicine. The probability that any of the other 38 topics generated this word is a billion times smaller.

With the symmetric parameterization of pLSA, we can also investigate medicine from the document perspective. The four most likely documents to produce the topic of MEDICINE are ITEM #33114, #59255, #11098, and #10619. Here the distribution of likelihoods is fairly flat: The most likely document is less than twice as likely as the 20th most likely document for the topic of MEDICINE. This is natural— even with a MEDICINE topic size of 0.9%, we expect 700 articles that talk about medicine (a search for the word medicine on Accessible Archive's Web site returns 1,146 matching articles). The following is an excerpt from the most likely document in the *Gazette* to generate the word medicine (ITEM #33114)

April 26, 1764

Now OPENING by NATHANIEL TWEEDY, DRUGGIST, At the Golden Eagle, a few Doors above Franklin and Hall Printing office, near the Courthouse, Philadelphia, just arrived in the King of Russia, Captain Robinson, from London, AS large a quality of DRUGS and MEDICINES as ever has been imported by any one person, at one time, in Philadelphia, which he will dispose of on the best terms; together with a fine assortment of Shop Furniture, Surgeons Instruments, Painters Colours, Medicine chests.

Obviously, this article about the opening of a drugstore and importation of medicines and medical supplies clearly epitomizes the topic we have labeled MEDICINE.

Finally, we can determine the mix of topics that make up this article. Computing $P(z|d = $ ITEM#33114), we find that this document is 92% about MEDICINE (topic #39), 3% about FABRIC (topic #6), 2% about ARTISAN/WARES (topic #32) and 1% about FOOD & DRINK (topic #37). The final 2% is from a combination of the remaining 36 topics. This suggests the connections between colonial medicine (which would not be professionalized until the 19th century) and a variety of other trade goods. More broadly, such a technique allows

scholars to find connections, between topics, that might not be otherwise apparent.

One may ask, from an information retrieval perspective, how the articles that contain the word *medicine* are distributed in the topic MEDICINE. It does not make sense to ask whether all documents containing the word *medicine* are in the topic MEDICINE because, even to infinitesimal extents, every document is contained in every topic. So we restrict this examination to the more likely documents in the topic. Of the 831 articles that contain the word medicine, 561 are in the 2,000 most likely documents in the topic MEDICINE, and the remaining 270 appear after the 2000th most likely document in the topic MEDICINE. But we caution the reader to not interpret this as a measure of precision/recall: Our method properly did not strongly identify some documents that contained the word *medicine* as rating highly in the topic MEDICINE. For example, one article selected from the post-2,000 most likely documents in the topic MEDICINE is a 6,000 word report on the "Operations of the Allied Armies of France and America, under the Command of His Excellency General Washington" (ITEM # 66676). This long article is primarily about army operations, but does contain a small amount of text about "Hospitals to be furnished . . . with medicines." Thus, although it contains the word *medicine*, it is correctly not identified as primarily being about the topic MEDICINE. In addition, there are articles in the top-2,000 that closely relate to the topic of MEDICINE, but don't contain the actual word *medicine* (e.g., ITEM #12422, which discusses various chemicals used medicinally).

Unlike previous quantitative analyses of colonial newspapers that assign each article to a single categorical topic, pLSA encourages the more realistic identification of multiple topics in a single article or advertisement (Clark and Wetherell, 1989, p. 292). Posterior probabilities are particularly useful for those general topics produced by lengthy and multitopic articles, as we can see in our second case study. As discussed, topic #3 is one of the less specific overall topics. Articles such as exchanges of letters result in this rather generic topic having to do with reasoning and making an argument. Unlike topic #39 MEDICINE (and the overall average of an article consisting 60% of one topic), the most likely documents given topic #3 LONG ARGUMENT are less than 50% about topic #3. For example, a message from the Governor to the Assembly from August 21, 1755 (ITEM #18603) is only 35% about topic #3. Other topics contributing to this article are 25% GOV'T - IMPERIAL (topic #7); 15% MERCANTALISM (topic #14); 11% MONEY (topic #16); and less than 5% each, GOV'T U.S.; GOV'T LEGISLATIVE; LAND; DEBT (topics #2, 8, 28, 26). Such complex articles with varied foci particularly benefit from a topic decomposition method that allows for more than a single topical categorization. Such a method allows historians to see crucial connections between thematic discussions in colonial newspapers, revealing, in this case, how closely early Americans connected politics, economics, and imperial concerns.

TABLE 6. The topic of books splits into three subtopics.

Number of topics	Most likely words in topic BOOKS		
80	BOOKS (1.0%) book vol published history price sold just new work volume english bible edition art author letter dictionary containing law collection		
160	BOOKS-RELIGIOUS/REFERENCE (0.3%) book vol bible testament history dictionary ink grammar paper english spelling life sort latin prayer small watt cole common work	BOOKS-NON-FICTION (0.3%) book vol history volume collection new work map edition complete law curious art price author bound dictionary neatly dollar page	ALMANACS (0.3%) published just price sold printed containing sun rising almanack author table account observation press american added franklin moon printing year

Topic Hierarchies

The problem of how to choose the number of clusters (here, topics) is well known in clustering. For our three decomposition techniques, pLSA, k-means, and LSA, the objective function always improves as T, the number of topics, increases. There are several techniques for finding the optimal value of T. One simple approach is to plot the objective function—in our case log likelihood—versus T, and look for a *knee* in the curve, that is, where there is a rapid flattening of the slope of the curve. We ran pLSA for $T = 10, 20, 40, 80, 120$, and 160, and plotted the log likelihood versus T in Figure 5. We tentatively suggest there is a *knee* in this curve at $T = 40$, indicating that this may be an optimal number of topics. Nonetheless, we more strongly advocate that for such a large corpus, any of these values of T are valid, and the appropriate value of T depends on how one plans to use the results. Furthermore, we believe that computing topics at different values of T gives us an interesting and useful hierarchy. Note that this hierarchy is not formal, like in hierarchical clustering—it would not be possible to construct a dendrogram (or tree) representing the hierarchy. Nevertheless, being able to generate topic descriptions at any level of detail is particularly useful for historical research.

In Table 7, we show an example of a topic that splits into separable subtopics. Going from 80 to 160 topics, we see a split of the BOOKS topic into topics we have named
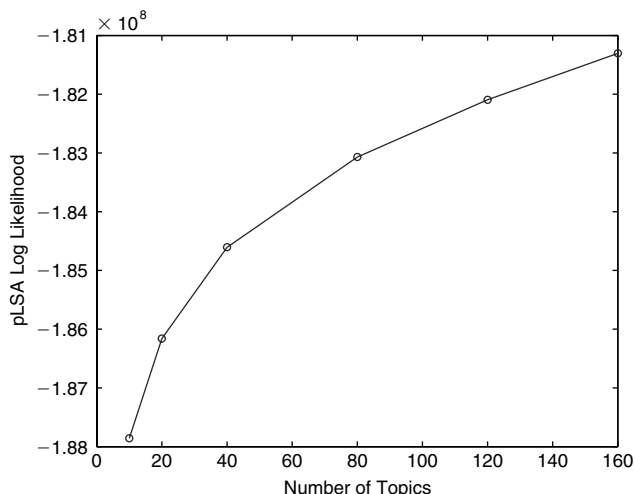
BOOKS-RELIGIOUS/REFERENCE, BOOKS-NON-FICTION, and ALMANACS. These topic splits reveal important information about the depth of the Philadelphia publishing world. With the emergence of the subtopic ALMANACS, we can see the importance of such inexpensive, locally-produced publications; the inclusion of the word *franklin* in the top-20 words reflects Benjamin Franklin's importance in this print genre (as the author of the well-known *Poor Richard's Almanac*). Although the other two topics overlap somewhat, the larger number of descriptors in the second topic (*new, edition, complete, curious, neatly*, etc.) suggests that these might account for new domestic publications rather than imported reprints of standard literary works (*bible, common law, dictionary*, etc.). Some scholars might be interested in the overall place of book advertisements in the *Gazette* at 80 topics, but the increased refinement at 160 topics would allow individuals more interested in popular print to trace the specificities of almanacs or imports. Thus, the precise levels of refinement may be as much a question of historical inquiry as computational expertise.

Returning to our MEDICINE example, we see a different structural feature in the topics in Table 7. At 40 topics, the topic of MEDICINE generates 0.9% of the *Gazette*. As we increase the number of topics to 80, this topic shrinks to 0.4% and gets more refined, with words relating precisely to medicine (e.g., *pill, balsam, elixir*) replacing more general words that might apply to a variety of sale goods (sort, best, size). However, as we further increase the number of topics to 160, the size of this topic stays at 0.4%, with 14 of the top 20 words identical to those at the 80 topic-size resolution.

FIG. 5. Likelihood versus number of topics.

TABLE 7. The topic of medicine gets resolved.

| Number of topics	Most likely words in topic MEDICINE
40	MEDICINE (0.9%) oil glass medicine sort pot best size large powder salt lead bottle white london sold case boxes stone assortment drug
80	MEDICINE (0.4%) medicine oil flower drug ball golden london marshall balsam white powder pill drop water sold gum assortment elixir bottle brushes
160	MEDICINE (0.4%) oil medicine drug powder golden lead glass balsam pill london gum water elixir sold best drop assortment patent imported large
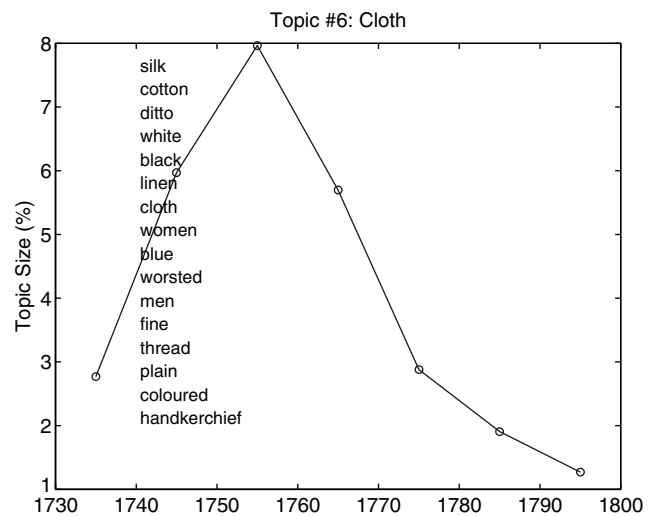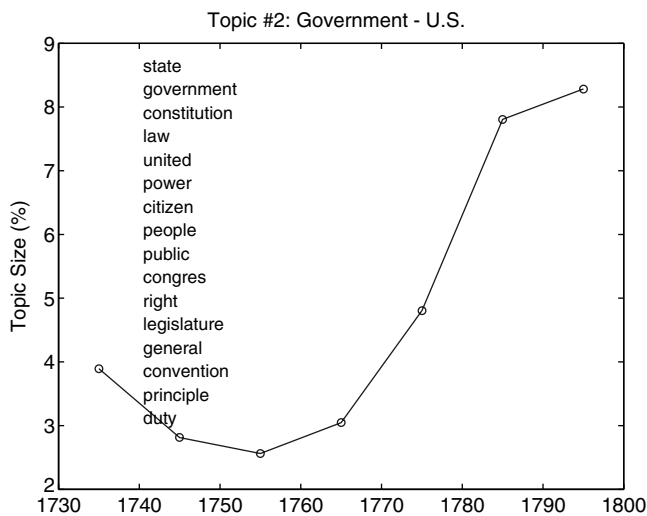
FIG. 6. Trends in the topics of GOVT'-U.S. (left) and CLOTH (right) over the century.

Topic Trends

For a historian, identifying topics, rather than single keywords, in a large corpus such as the *Gazette* is an extraordinary accomplishment. But it is even more useful when the topic trends are charted over time. Griffiths and Steyvers (2004) demonstrated how to use probabilities (in their case, produced by the LDA probabilistic mixture model) to analyze the temporal dynamics of the topics. After computing topics, a simple analysis indicates that the prevalence of a particular topic in any given time period is proportional to the number of words generated by that topic (as per the aspect model in the Comparison of Three Methods section). To compute the trends for the *Gazette*, we split the time period by decade. For simplicity we included articles from 1728 and 1729 in the 1730s decade, and articles from 1800 in the 1790s decade. Then, for each decade, we estimated the topic mixing ratios $P(z)$ using the word counts for that decade, and the previously computed topic posterior probabilities $P(z|w)$ for the entire run.

The prevalence of a given topic in any particular time period is given by a weighted allocation of each word in that time period to the given topic. We verified that all the topic trends shown in the figures are significant well beyond the 99.9% level. This is not particularly surprising, given the large number of observations: approximately 10 million words divided into seven time periods. Griffiths and Steyvers (2004) point out that if finding topic trends is the primary goal (above the goal of finding topics), then more sophisticated generative models may be more appropriate. Such models would incorporate parameters that describe the importance of a topic at any given point in time. Nevertheless, charts from our model show significant shifts that reveal important aspects of early American history.

This consistency indicates that this topic is already relatively coherent, and likelihood increases are instead found in other topics.

We see a variety of decadal trends for individual topics. Some, such as topic #2 (Figure 6), about the United States government, follow obvious trends in political history: discussion of a national government near-triples from the pre-Revolutionary (1760s) to the Early National (1790s) period. Topic #6, CLOTH, first shows an increase as luxury goods became more readily available and the *Gazette* increasingly advertised for imported fabrics. However, a marked decline from the 1760s on, likely relates to the colonists' growing emphasis on homespun fabrics as part of their boycott of British goods during the years leading up to Independence. The trends in CLOTH also suggest the changing place of advertisements in the *Gazette*, as it became a more expressly political newspaper in the Revolutionary era (Clark & Wetherell, 1989). Our qualitative analysis of a variety of trends identified through our method shows that our topic trends match historians' findings about early American history and early American print culture. For example, Clark and Wetherell found that advertisements (such as those for CLOTH shown in Figure 6) peaked in the 1750s, which agrees with the timing of our peak in the topic CLOTH. Similarly, the significant increase of documents related to the topic GOVERNMENT-U.S. in the 1780s and 1790s obviously parallels the well established public political discussions taking place during the formation of a new United States Government (Figure 6).

Figure 7 shows how charts of temporal shifts can be used to reveal cultural, rather than political or economic, histories. Topic #36, on CRIME, reached a high in the 1730s, and then declined precipitously until a Revolutionary era climb brought it nearly back to original levels. A somewhat similar pattern occured in the *Gazette*'s discussion of religion (figure 7, right). Other scholars have pointed to the decline in religious content in stand-alone criminal narratives in the early Republic, as trial transcripts replaced the publication of sermons given at executions (Cohen, 1993). But the shifting content in the *Gazette* suggests that similar cultural ideologies may have driven interest in both earthly and divine justice, even if criminal narratives no longer had an
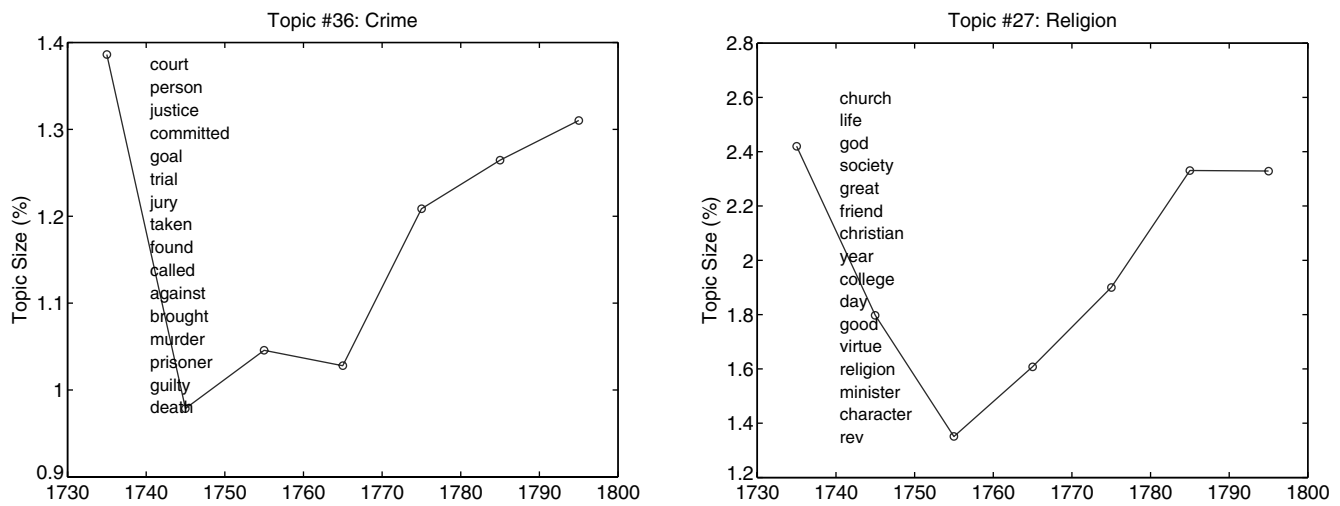
FIG. 7. Trends in the topics of CRIME (left) and RELIGION (right) over the century.

expressly religious focus, and raises questions about the widespread applicability of shifts in single print genres to all print sources. Thus, the use of topic trends can be applied to a wide variety of historical analytic questions in political, social, and cultural history.

Conclusions

We have applied a new method—probabilistic Latent Semantic Analysis (pLSA)—for finding topics in the *Pennsylvania Gazette*, one of the most important newspapers in early America. After reviewing several topic decomposition methods, we argue, using various experimental results, that pLSA is a good method for this type of historical research. The pLSA method is able to efficiently compute meaningful topics in a large corpus (10^5 documents, 10^7 words), and allows a mix of topics within a single document. Also, because of pLSA's probabilistic formulation, we are able to compute interesting probabilities that tell us about the different contexts of word usage in early America (e.g., the word *beauty* was used mostly in literary topics, but also frequently in topics relating to the American Revolution). This method can also find topics at arbitrary levels of detail, providing an ad hoc hierarchy of topics that can be tied to a historian's particular interests. We also presented a valuable extension of computing topics: determining the time-dynamics of these topics. These trends over time add another interesting analytical layer, allowing us to find relationships between different topics.

The ability to compute these topics and topic trends in a totally automated and unsupervised fashion is of exceptional value to historians. Because there is no a priori designation of topics—in fact there are very few "knobs to turn" in the method—historians do not need to rely on fallible human indexing or their own preconceived identification of topics. But the most important advantage of this method is its ability to analyze orders-of-magnitude more documents than a person can reasonably view. Thus, instead of resorting to

sampling to analyze a large volume of documents, the computer can analyze the entire corpus. This is especially useful in studies of print culture that strive to understand how historical actors read and understood entire publications. In addition, this methodology provides a useful tool for indexing text documents with analytic topics.

Extensions of this research might include further analysis of topic trends over time and at different resolutions. As more early American newspapers become available online, we can begin to compare the content of several newspapers in the period, which could reveal regional and editorial differences in particular publications. Other serials, such as almanacs and magazines, might provide a fertile ground for cultural analyses of shifting topics in short fiction, poetry, or bawdy humor. Scholars of material culture would be particularly interested in the shifting focus of advertisements for consumer goods. Beyond such print culture sources, analyzing large collections of personal writings, such as diary entries or letters, would reveal how different individuals understood and recorded the world around them. In conclusion, we believe that this approach provides a crucial, and to date under-used, methodology for analyzing the ever increasing numbers of full-text historical sources.

Acknowledgment

References

Accessible Archives Inc. (2004). Retrieved September 15, 2004, from http://www.accessible.com

Aldridge, A.O. (1962). Benjamin Franklin and the *Pennsylvania Gazette*. Proceedings of the American Philosophical Society, 106, 77–81.

Amory, H., & Hall, D.D. (2000). The colonial book in the Atlantic world. American Antiquarian Society. New York: Cambridge University Press.

Berry, M.W., Drmac, Z., & Jessup, E.R. (1999). Matrices, vector spaces and information retrieval. SIAM Review, 41(2), 335–362.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Block, S. (2002). Rape without women: Print culture and the politicization of rape, 1765–1815. Journal of American History, 89(3). Retrieved September 15, 2004, from http://www.historycooperative.org/journals/jah/89.3/block.html

Chakrabarti, S. (2003). Mining the Web: Discovering knowledge from hypertext data. San Francisco: Morgan Kaufman.

Clark, C.E., & Wetherell C. (1989). The measure of maturity: *The Pennsylvania Gazette*, 1728–1765. William and Mary Quarterly, 46(2), 279–303.

Coccaro, N., & Jurafsky, D. (1998). Towards better integration of semantic predictors in statistical language modeling. Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP).

Cohen, D.A. (1993). Pillars of salt, monuments of grace: New England crime literature and the origins of American popular culture, 1674–1860. New York: Oxford University Press.

Deerwester, S., Dumais, G.W., Furnas, S.T., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41, 391–407.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B39, 1–38.

Desrochers, R.E., Jr. (2002). Slave-for-sale advertisements and slavery in Massachusetts, 1704–1781. William and Mary Quarterly, 59, 623–664.

Dhillon, I.S., & Modha, D.S. (2001). Concept decompositions for large sparse text data using clustering. Machine Learning, 42, 143–175.

Duda, R.O., & Hart, P.E. (1973) Pattern classification and scene analysis. New York: Wiley.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl. 1, pp. 5228–5235).

Grubb, F. (1999). Lilliputians and Brobdingnagians, stature in British Colonial America: Evidence from servants, convicts, and apprentices. Research in Economic History, 19, 139–203.

Hofmann, T. (1999) Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42, 177–196.

Lee, D.D., & Sung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791.

Rosen-Zvi, M., Griffiths T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for authors and documents. 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada.

Salinger, S. (1987). "To serve well and faithfully": Labour and indentured servants in Pennsylvania, 1682–1800. New York: Cambridge University Press.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Sibson, R. 1973. SLINK: An optimally efficient algorithm for the single link cluster method. Computer Journal, 16, 30–34.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA.