Toward a Framework for the Large Scale Textual and Contextual Analysis of Government Information Declassification Patterns

Rachel Shorey
Computer Science

Hanna Wallach
Computer Science

Bruce Demarais
Political Science

University of Massachusetts, Amherst

1 Introduction

The US government protects a massive amount of secret data as part of its Security Classification System. This information is expensive to protect and maintain. In order to keep citizens informed, as well as to keep costs down, the government is constantly releasing newly declassified documents to the public. According to OpenTheGovernment.org's annual Secrecy Report Card, human readers manually declassified almost 29 million pages of information in 2009 alone (McDermott and Bennett, 2010).

Scholars interested in learning about government transparency history and policy face a daunting task in examining even a small portion of these documents. In order to make the process of learning about the content of these documents easier, we investigate the documents available through Gale's Declassified Documents Reference System, an electronic repository of once-classified documents created throughout the 20th century, along two dimensions. First, we perform survival analysis to consider questions relating to time, such as when documents were created and how long they tend to remain classified. Then, we examine the contents of the documents via the use of statistical topic models. Finally, we combine temporal and content information, both by using the output of a topic model to inform a proportional hazards survival model and by considering time information during inference in a topic model.

In this paper, we present a range of results that arise from a combined analysis of the temporal features and textual content of declassified documents. Since the results of statistical topic models tend to be easily-interpretable by humans, these new ways of looking at data relating to declassified government documents will likely be useful to experts in fields relating to government policy on secrecy.

2 Secrecy in the United States

Every time the U.S. Congress reviews the security classification system of the executive branch, it concludes that over-classification—the process by which too much information is classified for too long—places an unwarranted burden on government resources, keeps citizens uninformed, and poses a risk to national security (Relyea, 1999). However, faced with the daunting task of determining what, among millions of pages of information, can be released to the public, understandably risk-averse government information officials default to secrecy (Pozen, 2005). The post-9/11 era has witnessed two new developments in the over-classification problem—the proliferation of pseudo-secrecy (i.e., 'sensitive but unclassified' markings) (Relyea, 2003) and the executive's de-prioritization of transparency in the face of the threat of terrorism (Jaeger, 2007). It is therefore more critical than ever that information officials be certain of security risks related to the declassification of information. We seek to produce tools that could help policy makers better understand and handle the oft-juxtaposed objectives of precision and comprehensiveness in reviewing information for possible public release.

The available models for automatic analysis of the text of declassified documents focus directly on determining whether a new document can be safely declassified. Automatic declassification is often approached from the perspective of modeling all secret information and removing or preventing release of previously unreleased information. Previously-proposed approaches include an ontological semantic approaches (Attallah et al, 2001) and the use of information mined from documents to create boolean formulas representing data that should or should not be released (Sánchez et al, 2002). In this paper, we hope to further the understanding of the ways in which document content influences the decision to declassify documents without explicitly modeling specific pieces of information contained in the document. We hope to inform declassification policy using historical patterns as a guide in addition to assisting in the declassification decisions regarding specific documents. While we do not propose a fully automated document declassification scheme, the statistical models we present in this paper could certainly be a component of such a system by suggesting documents ripe for release or helping to funnel documents to appropriate departments or experts for review on the basis of document contents.

3 Survival Analysis

Survival analysis, also known as 'duration modeling' or 'event history analysis' is a statistical method for analyzing time spans (e.g., the time from birth until death, the incubation time of HIV, or the length of a war). The conventional objective in survival analysis is to estimate the effects of some covariates on the location of the distribution of time spans. As such, it is very similar to regression modeling. Tools for survival analysis have been developed to accommodate common features of time span data, and the generative processes underlying that data. The most basic of these characteristics include the facts that the distribution of time spans must have strictly positive support, and observations are often right- or interval-censored (i.e., it is only known that the time span is between two observation points or that the time span is greater than some censoring point, respectively) (Klein and Moeschberger, 2003).

The first major choice in survival modeling is the selection of the quantity in the distribution of the time span (denoted by Y) which is conditioned by the covariates. Two common choices are (1) the expectation of the log of Y, and (2) the log of the instantaneous rate of failure—known as the log hazard rate (Box-Steffensmeier and Jones, 1987). In accelerated failure time (AFT) models

$$\mathrm{E}\left[\ln(Y_i)\right] = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where $\boldsymbol{x}$ is a p-vector of covariates and $\boldsymbol{\beta}$ a p-vector of regression coefficients. Alternatively, in a proportional hazards (PH) model, the assumption is that

$$\frac{f(Y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta})}{1 - F(Y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta})} = \Lambda_0(Y)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

where $f(\cdot)$ is the density of Y, and $\Lambda_0(Y)$ is the hazard rate when $\boldsymbol{x} = \mathbf{0}$—the baseline hazard. Either parameterization can be considered, but it is important to note that in the AFT form, the effect of x_j on the location of Y is in the same direction as β_j, but in the PH parameterization, the effect is opposite the direction of β_j. Here, we use a PH model.

The second major decision is whether to specify a fully- or semi-parametric model. The Cox proportional hazards model (Cox, 1972) is a very popular survival model because, given the assumption that the hazard rate is proportional to $\exp(\boldsymbol{x}'\boldsymbol{\beta})$, unbiased estimation of $\boldsymbol{\beta}$ does not require the specification of the baseline hazard rate. This means that the analyst need not make any (possibly arbitrary) assumptions about how the risk of failure varies over the duration of the event under study, outside of the dependence of that risk on the covariates. The major drawback of the Cox framework is that it does not permit the simulation/prediction of event times, but only hazard ratios. For our initial analyses we use the Cox model, but we turn to a parametric assumption when we incorporate time into a topic model.

4 Topic Models

Statistical topic models, such as latent Dirichlet allocation (Blei et al, 2003) discover clusters of words, known as topics, based on their co-occurrence patterns within documents. These models are capable of producing clearly-interpretable topics (see table 2 for example topics extracted from approximately 80,000 declassified documents) without any labeled data or human intervention. A key strength of statistical topic models over other document clustering methods (such as k-means clustering or naïve Bayes) is that statistical topic models are based on the assumption that each document is represented by a mixture (i.e., a probabilistic combination) of topics, where each topic is represented by a probability distribution over words in some vocabulary. Thus, for instance, a document that discusses both the Vietnam War and the USSR will be explicitly represented as such. Additionally, the use of a fully-probabilistic framework has the significant benefit of enabling the incorporation of any additional relevant evidence and structure beyond text (such as timestamps or author identities) directly into the model.

Statistical topic models have been used for a variety of analyses relating to political science and policy. One particularly rich source for text analysis that is of interest to political scientists is parliamentary proceedings from various legislative bodies. The bills and votes that have come before the US Senate are the subject of a wide variety of studies based on topic modeling. In their "Group-Topic" model, McCallum et al. jointly discover topics based on the text of each bill and groups of Senators based on individuals' voting patterns (McCallum et al, 2007). Topic models have been used to inform ideal point estimation (Gerrish and Blei, 2010), demonstrating that including topic information in an ideal point spatial model produces better accuracy in predicting roll call votes. In addition to Senate bills, other legislative text has been analyzed using topic models, including press releases (Grimmer, 2010) and floor speeches (Fader et al, 2007) by politicians.

4.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) (Blei et al, 2003) is a generative, probabilistic model of documents. That is, LDA is characterized by a set of probabilistic rules that describe the process of generating word tokens in hypothetical documents, using unobserved (latent) random variables—in this case, topics (word clusters). The generative process for LDA is relatively simple: to generate a document, a document-specific mixture of topics is first selected. Then, to generate each word token in that document, a topic is selected at random from the document-specific mixture and a word is drawn from the distribution over word types corresponding to the chosen topic.

Clearly, real documents are not actually created using this process. When fitting a generative model to real-world data, such as a collection of documents, the goal is to find the set of latent variables that best characterize the observed data, assuming the model was in fact responsible for generating the data. Standard statistical techniques can be used to invert the generative process and infer the latent topics (probability distributions over words) that best characterize a particular document collection, as well as the document-specific topic mixtures.

Mathematically, LDA is a model for documents $\mathcal{W} = \{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(D)}\}$. A "topic" t is a discrete distribution over words with probability vector $\boldsymbol{\phi}_t$. A symmetric Dirichlet prior with concentration parameter β is placed over $\Phi = \{\boldsymbol{\phi}_1, \ldots \boldsymbol{\phi}_T\}$:

$$\begin{aligned} P(\Phi) &= \textstyle\prod_t \mathrm{Dir}\left(\boldsymbol{\phi}_t; \beta\right) \\ &= \textstyle\prod_t \frac{\Gamma(\beta)}{\prod_w \Gamma(\frac{\beta}{W})} \prod_w \phi_{w|t}^{\frac{\beta}{W}-1} \delta\left(\sum_w \phi_{w|t} - 1\right). \end{aligned}$$

Each document, indexed by d, has a document-specific distribution over topics $\boldsymbol{\theta}_d$. The prior over $\Theta = \{\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_D\}$ is also assumed to be a symmetric Dirichlet, this time with concentration parameter α.

The tokens in every document $\boldsymbol{w}^{(d)} = \{w_n^{(d)}\}_{n=1}^{N_d}$ are associated with corresponding topic assignments $\boldsymbol{z}^{(d)} = \{z_n^{(d)}\}_{n=1}^{N_d}$, drawn from the document-specific distribution over topics, while the tokens are drawn from the topics' distributions over words:

$$\begin{aligned} P(\boldsymbol{z}^{(d)} \mid \boldsymbol{\theta}_d) &= \textstyle\prod_n \theta_{z_n^{(d)}|d} \\ P(\boldsymbol{w}^{(d)} \mid \boldsymbol{z}^{(d)}, \Phi) &= \textstyle\prod_n \phi_{w_n^{(d)}|z_n^{(d)}}. \end{aligned}$$

Dirichlet–multinomial conjugacy allows parameters Θ and Φ to be marginalized (integrated) out.

For real-world documents, the word tokens $\mathcal{W}$ are observed, while the corresponding topic assignments $\mathcal{Z}$ are unobserved (latent). Variational methods (Blei et al, 2003; Grimmer, 2011) and MCMC methods (Griffiths and Steyvers, 2004) are both effective at inferring $\mathcal{Z}$. Here, we use MCMC methods—specifically Gibbs sampling (Geman and Geman, 1984), which involves sequentially resampling each topic assignment $z_n^{(d)}$ from its conditional posterior given $\mathcal{W}$, α, β and $\mathcal{Z}_{\backslash d,n}$ (the current topic assignments for all tokens other than the token at position n in document d):

$$P(z_n^{(d)} \,|\, \mathcal{W}, \mathcal{Z}_{\backslash d,n}, \alpha, \beta) \propto \frac{N^{\backslash d,n}_{w_n^{(d)}|z_n^{(d)}} + \frac{\beta}{W}}{N^{\backslash d,n}_{z_n^{(d)}} + \beta} \, \frac{N^{\backslash d,n}_{z_n^{(d)}|d} + \frac{\alpha}{T}}{N_d - 1 + \alpha},$$

where sub- or super-script "$\backslash d, n$" denotes a quantity excluding data from position n in document d.

4.2 Including time information

Statistical topic models such as LDA produce human-interpretable topics (i.e., specialized probability distributions over some shared vocabulary) that characterize the textual content of document collections. As we discuss in section 6.1, we can combine these topics in a post-hoc manner with temporal information to produce easily-understood information about which concepts are discussed frequently at what times. In addition, we can make use of the inferred topics to inform a survival analysis of the documents, as we discuss in section 6.2.

In neither of these cases, however, does the temporal information have any influence in the *creation* of topics. In contrast, allowing time to play a model-internal role provides some evidence for (1) separating documents that contain similar words but were written at very different times and (2) placing words from documents written at the same time together.

A number of statistical topic models incorporate information from different modalities into topic inference. Among them, the most closely-related to this work is the Topics-Over-Time model (Wang and McCallum, 2006), in which the date that a document was written is used to inform topic inference.

Here, we expand on Wang and McCallum's model in several ways. As in LDA, we model each document as a discrete distribution over topics and each topic as a discrete distribution over words. We give each of these distributions a symmetric Dirichlet prior. We model the creation time of each word using topic-specific normal distribution with unknown mean and variance. We place a normal-gamma prior on the parameters of this distribution over times. Finally, we model the duration of classification for each document using topic-specific normal distribution in log-space, again with a normal-gamma prior over the unknown mean and variance parameters. Our notation is described in table 1. The conditional independence assumptions embodied by the joint distribution over all observed and unobserved variables are represented by the graphical model in figure 1 and described below. Note that instead of the typical parameterization in terms of variance, we parameterize all normal distributions using precision (inverse variance) in order to simplify marginalization. Note also that the s variables are log durations.

$$\begin{aligned}
z_n^{(d)} &\sim \text{Discrete}\,(\boldsymbol{\theta}_d) \\
\boldsymbol{\theta}_d &\sim \text{Dir}\,(\boldsymbol{\theta}_d; \alpha) \\
w_n^{(d)} &\sim \text{Discrete}\,(\boldsymbol{\phi}_{z_n^{(d)}}) \\
\boldsymbol{\phi}_t &\sim \text{Dir}\,(\boldsymbol{\phi}_t; \beta) \\
c_n^{(d)} &\sim \text{Normal}\,(\mu_{z_n^{(d)}}, \rho_{z_n^{(d)}}) \\
\mu_t, \rho_t &\sim \text{Gamma-Normal}\,(\mu', \rho', \zeta, \gamma) \\
s_n^{(d)} &\sim \text{Normal}\,(\nu_{z_n^{(d)}}, \tau_{z_n^{(d)}}) \\
\nu_t, \tau_t &\sim \text{Gamma-Normal}\,(\nu', \tau', \iota, \kappa)
\end{aligned}$$

Topic assignments are inferred using Gibbs sampling: each $z_n^{(d)}$ is resampled from its conditional posterior given the words, creation dates, classification durations, and other topic assignments:

$$\begin{aligned}
&P(z_n^{(d)} \,|\, \mathcal{W}, \mathcal{C}, \mathcal{S}, \mathcal{Z}_{\backslash d,n}, H) \propto \\
&\quad \frac{N^{\backslash d,n}_{w_n^{(d)}|z_n^{(d)}} + \frac{\beta}{W}}{N^{\backslash d,n}_{z_n^{(d)}} + \beta} \, \frac{N^{\backslash d,n}_{z_n^{(d)}|d} + \frac{\alpha}{T}}{N_d - 1 + \alpha} \\
&\quad \times \frac{\left(\zeta + \frac{C^{\backslash d,n}}{2} + \frac{(N_{z_n} - 1)\rho'\left(\mu' - \hat{c}^{\backslash d,n}\right)}{2(N_{z_n} - 1 + \rho')}\right)^{\gamma + \frac{N_{z_n} - 1}{2}}}{\left(\zeta + \frac{C}{2} + \frac{N_{z_n}\rho'(\mu' - \hat{c})^2}{2(N_{z_n} + \rho')}\right)^{\gamma + \frac{N_{z_n}}{2}}} \\
&\quad \times \frac{\left(\sqrt{\rho' + N_{z_n} - 1}\right)\Gamma\left(\gamma + \frac{N_{z_n}}{2}\right)}{\left(\sqrt{\rho' + N_{z_n}}\right)\Gamma\left(\gamma + \frac{N_{z_n} - 1}{2}\right)} \\
&\quad \times \frac{\left(\iota + \frac{S^{\backslash d,n}}{2} + \frac{(N_{z_n} - 1)\tau'\left(\nu' - \hat{s}^{\backslash d,n}\right)}{2(N_{z_n} - 1 + \tau')}\right)^{\kappa + \frac{N_{z_n} - 1}{2}}}{\left(\iota + \frac{S}{2} + \frac{N_{z_n}\tau'(\nu' - \hat{s})^2}{2(N_{z_n} + \tau')}\right)^{\kappa + \frac{N_{z_n}}{2}}} \\
&\quad \times \frac{\left(\sqrt{\tau' + N_{z_n} - 1}\right)\Gamma\left(\kappa + \frac{N_{z_n}}{2}\right) e^{s_n^{(d)}}}{\left(\sqrt{(\tau' + N_{z_n}}\right)\Gamma\left(\kappa + \frac{N_{z_n} - 1}{2}\right)}
\end{aligned}$$

Symbol	Description
$w_n^{(d)}$	the type of the n^{th} word token in document d
$z_n^{(d)}$	the topic assignment for the n^{th} word token in document d
$c_n^{(d)}$	the creation date of n^{th} word token in document d
$s_n^{(d)}$	the log of the duration between the creation time and declassification time of the n^{th} word token in document d
N_d	the total number of word tokens in document d
$N_{z_n^{(d)}}$	the total number of word tokens assigned to topic $z_n^{(d)}$
D	number of documents in corpus
T	the number of topics
W	the size of the vocabulary

Table 1: Notation used throughout this paper.

where H represents the set of model hyperparameters, $\hat{c}$ and $\hat{s}$ the mean date and duration, respectively, $C = \sum(c_n - \hat{c})^2$, and $S = \sum(s_n - \hat{s})$.

We use normal distributions for two reasons. First, it is quite reasonable to assume that the dates of document creation and the log classification durations are normal. Second, our goal is to produce tools for researchers in the field of transparency research who might not be intimately familiar with a wide variety of probability distributions. One of the main advantages of statistical topic models is their easily-interpretable topics. We hope to create a model that remains very easy to interpret even when additional information is modeled. We therefore believe that using the familiar normal distribution furthers that end. We operate in log space for the duration component to prevent negative durations.

It seems somewhat odd to consider the creation date and classification duration of every word, since each document (and hence all words in that document) has a single creation date and classification duration. Following (Wang and McCallum, 2006), we choose to consider the dates in this way because we are most interested in determining which topics are frequently used at what times. Ensuring dependence between topics and temporal properties can be accomplished in one of two ways: we could constrain each document to a single topic, or we could consider each word to be written and declassified independently at times deterministically inherited from the document to which it belongs. We choose the latter because we are unsatisfied with the constraint that a document can express only a single topic. By representing documents as mixtures of topics, we obtain a more nuanced picture of both the individual documents and the corpus as a whole.

5 Declassified Documents Corpus

We draw the documents modeled in this paper from the Gale Declassified Documents Reference System. This database contains approximately 90,000 documents created since the 1920s, of which we were able to mine 88,045. We discarded approximately 10% of these documents because either the creation or declassification date was not available.

6 Results and discussion

In this section, we present results several analyses of the declassified documents in our corpus based on time and content data available for the documents. Our goal is to provide tools to assist researchers interested in transparency and secrecy in interpreting large collections of declassified documents. As such, we present several analyses of the content of the corpus produced by combining textual and temporal information in various different ways.

6.1 Topic Model

Using the MALLET topic modeling package (McCallum, 2002), we ran LDA with 1000 topics for 1000 iterations of Gibbs sampling on all eligible documents. Many of the resulting topics correspond well with distinct events or periods in US History or to common subjects of diplomatic communication

Philippines	**Congress**	**Oil**	**Vietnam Bombings**
PHILIPPINES	BILL	OIL	TARGETS
MARCOS	LEGISLATION	PETROLEUM	AIR
MANILA	CONGRESS	CRUDE	ATTACKS
FILIPINOS	HOUSE	GAS	BOMBING
PRESIDENT	ACT	PRODUCTION	NORTH
BASES	ADMINISTRATION	PIPELINE	STRIKES
MAGSAYSAY	SENATE	SUPPLY	MILITARY
MACAPAGAL	AUTHORITY	IMPORTS	HANOI
HUK	PASSED	FUEL	HAIPHONG
PHILCAG	ACTION	BARRELS	DAMAGE

Table 2: The ten most frequent words for each of four example topics. Topics headings were selected by hand.

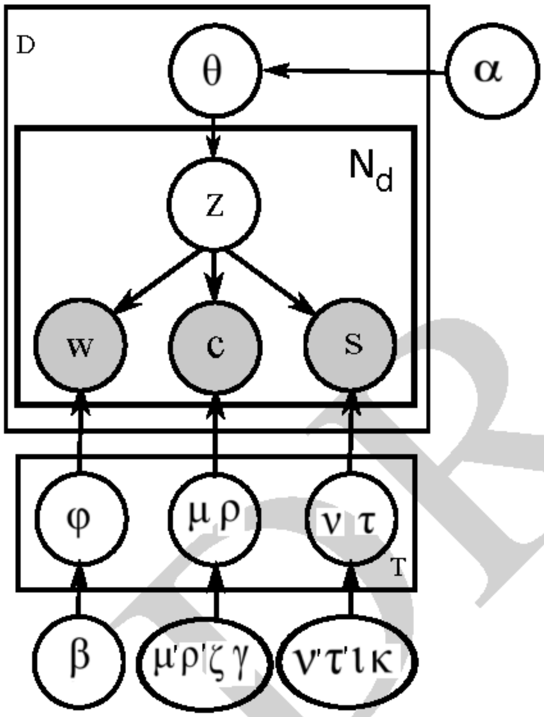

Figure 1: Topic-Time-Duration model

or government investigation. We present several example topics in table 2. Semantically-coherent topics such as these allow humans to quickly get a sense of the content of document collections. A scholar investigating communication regarding military activities during the Vietnam War could, for instance, choose to investigate those documents containing a high percentage of the fourth topic in table 2.

While examining the topics produced from a large document collection can help in navigation, a more nuanced picture of US government secrecy and declassification policy emerges when topics are examined in conjunction with timestamp information. Figures 2 and 3 each show the average percentage of documents from a given year created (in black) and declassified (in red) that are drawn from the topic in question. In Figure 2, we see the classification and declassification patterns for documents relating to military organization, and in Figure 3 for documents relating to Martin Luther King, Jr. The declassification pattern for the military organization topic is not particularly surprising, as the location and organization of military units is not very relevant after those troops have been moved. The comparatively long wait, followed by a spike of declassification relating to King may be of particular interest to researchers studying the results and consequences of the now well-known surveillance programs of the 1960s.

6.2 Survival Analysis

In addition to examining the proportion of documents created and declassified at various times that belong to a given topic, we are interested in whether

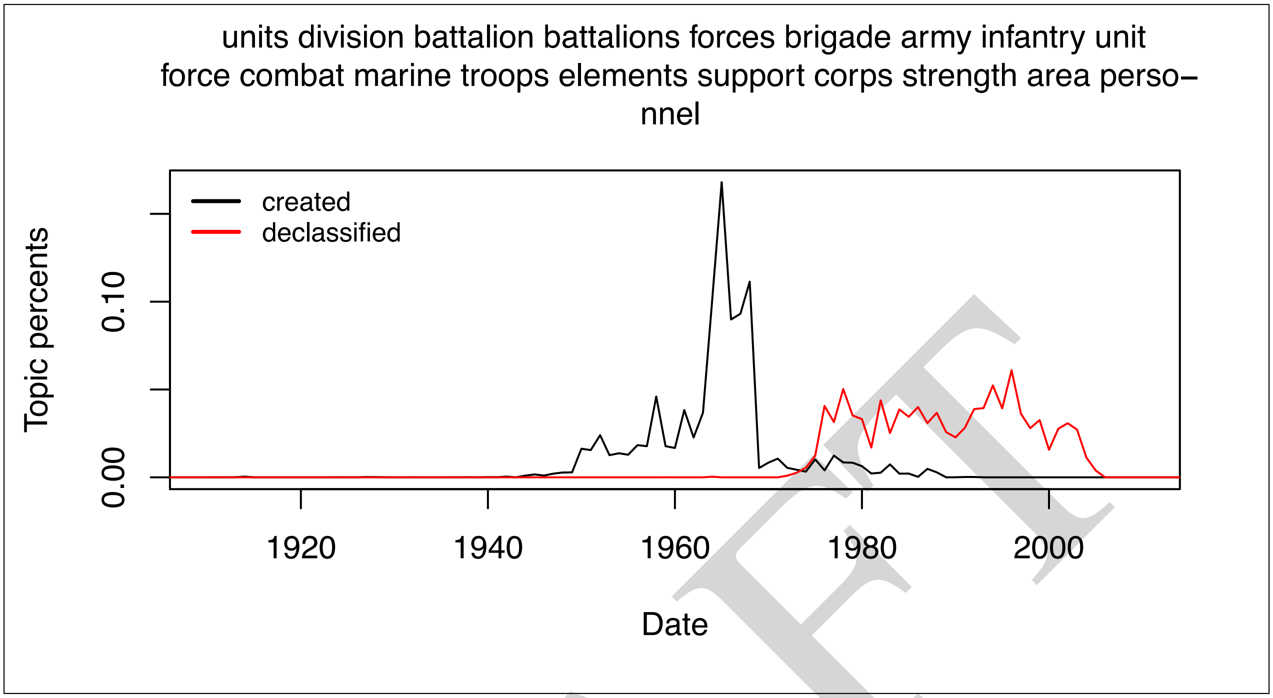

Figure 2: Classification and declassification patterns for documents relating to military structure

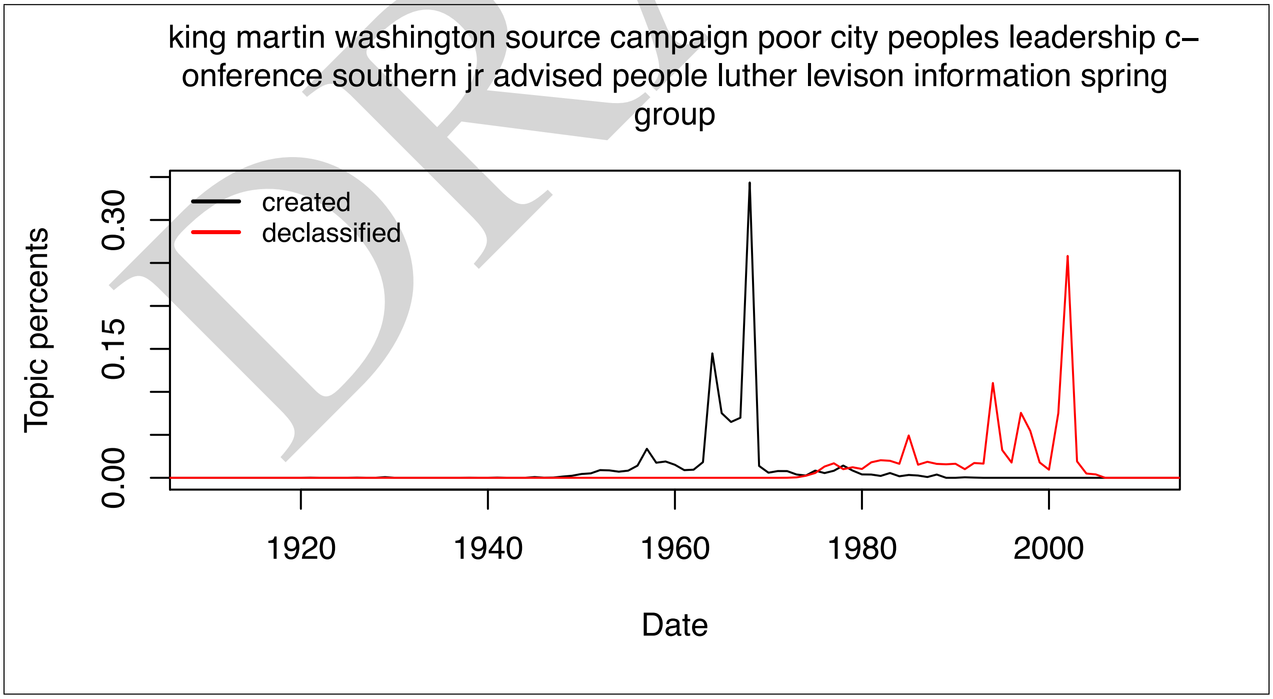

Figure 3: Classification and declassification patterns for documents relating to Martin Luther King, Jr.

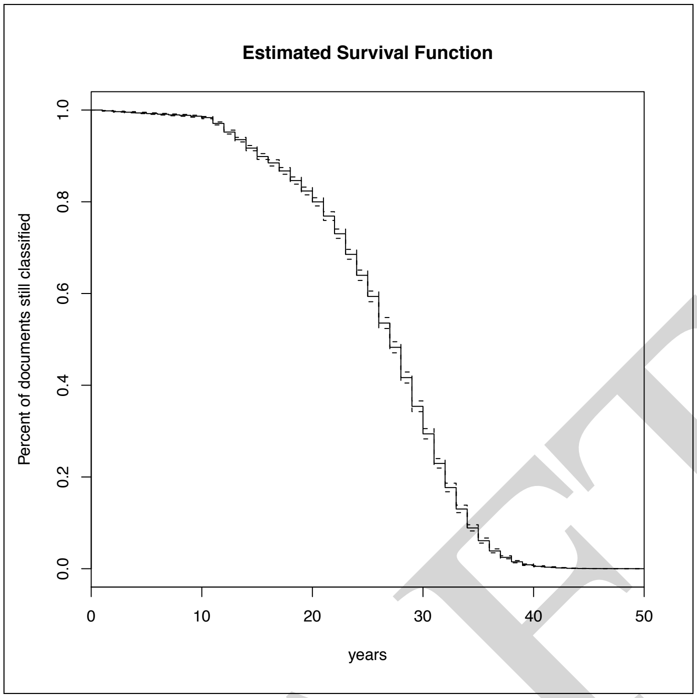

Figure 4: Estimated survival curve based on fitted Cox proportional hazard model

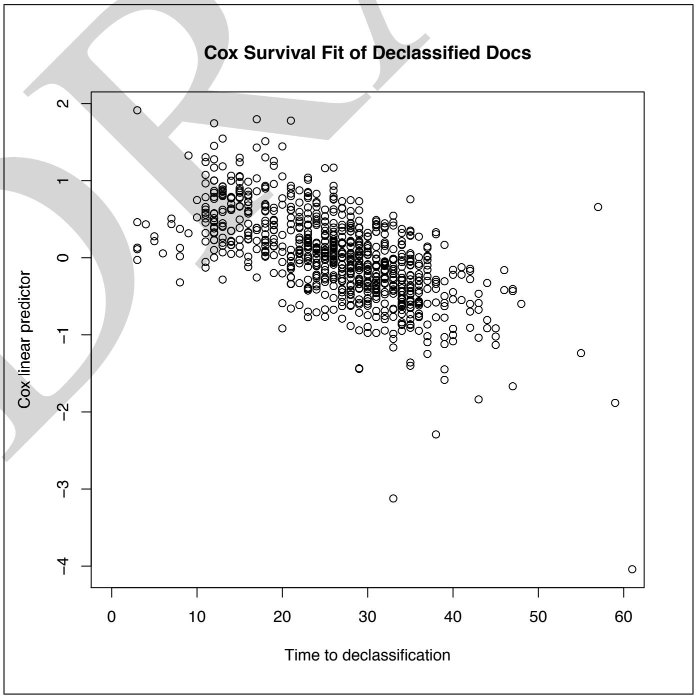

Figure 5: Cox model linear predictors

the textual content of a document can help us learn more about declassification patterns. In order to investigate this question, we ran a Cox proportional hazards survival model with topic membership of words in the document as covariates. We used the *survival* library from the R statistical package to perform this analysis. For this portion of our analysis, we restricted ourselves to approximately 8000 "training" documents (about 10% of our corpus) to train our model, and predicted linear predictors on a test set of approximately 1000 documents. We began by using MALLET to train LDA with 100 topics on the training documents, and using the trained topic model to infer topic distributions for the test documents. We then trained a Cox proportional hazards model using percentage of a document falling into each topic as regressors, and predict log hazard ratios on the test set using the topics inferred by MALLET. In figure 4 we show the estimated survival curve from the Cox model. This curve is estimated by considering the mean value of each covariate and it shows the general shape of the survival function for the corpus. Figure 5 plots the declassification date by the log hazard ratio produced by the Cox model. Survival analysis and the Cox model are not intended to provide exact predicted survival times for specific individuals. Rather, the model produces relative comparisons among documents, with large log hazard ratios indicating documents that "survive" for relatively short periods—that is, documents that are declassified relatively quickly as described in section 3. As is evident from figure 5, the Cox model with topic membership proportions as regressors indicates that relative durations of classification vary across topics in the corpus.

6.3 Topics and Times

In addition to combining the results from a topic model with an analysis of document declassification time patterns, we incorporated classification duration into topic inference. We used the model described in section 4.2 to infer latent topics based on word tokens, token creation time, and token declassification time. Using a subset of approximately 1000 documents, we inferred 100 topics using Gibbs sampling until convergence. We compare the declassification duration patterns of four different topics produced by the model below. The words most highly associated with each of the four example topics relating to budgets and funding, Southeast Asia, Vietnam and Berlin, respectively, are shown in table 3 and their classification durations are compared in figure 6. The mean of the fitted log normal distributions over creation times for these topics are 1977, 1964, 1966 and 1959, respectively. It is not particularly surprising that documents containing lots of words about budgeting and funding may be declassified earlier than those about military or diplomatic strategy in Southeast Asia or with the Soviets about Berlin. Most funding and budget related requests will eventually make it into the congressional record and thus the public sphere, after which point there is no need to keep budget requests secret. Also, budgetary matters follow the fiscal calendar, which might eliminate the need to keep classified budgetary information for more than a few years. It is the classification duration pattern of the topic we have titled 'Vietnam' in table 3 that shows the value of considering durations in a model-internal fashion. The mean time chronological for creation of words in this topic is similar to that of the 'Southeast Asia' topic (1966 vs. 1964), and the topic deals with a similar region of the world. The 'Vietnam' topic relates more directly to actions and diplomacy in Vietnam itself, where the US was openly engaged in combat in a war covered in great detail by the media of the day. In contrast, the 'Southeast Asia' topic, on the other hand, regards countries where the CIA is now known to have played a role despite the lack of openly-acknowledged warfare on the part of the US. This serves as a proof of concept—our joint Topic-Time-Duration model is capable of differentiating between topics along the covert/overt distinction—an important substantive difference for government information officials and researchers alike.

Contrasts like that described above, between different types of documents such as budget requests and diplomatic or military correspondence, could play an important role in a review of government secrecy and declassification policy. Funding-related material probably does not need to be classified for a long period of time, if at all. The information gleaned from these topics could also help government declassification workers screen candidate documents for ones that are ready for release. Discovering contrasts between latent topics like 'Vietnam'

Funding/Budgeting	Southeast Asia	Vietnam	Soviet Berlin
REQUEST	ACTION	VIETNAM	BERLIN
MILLION	VIENTAINE	SORTIES	MEETING
WHITE	LAOS	COPY	SOVIET
HOUSE	BANGKOK	MEMORANDUM	POSITION
DEVELOPMENT	PHOUMI	VIET	KHRUSHCHEV
BUDGET	IMMEDIATE	CONG	WESTERN
LEGISLATION	CINCPAC	AUTHORITY	SITUATION
MEMORANDUM	TELEGRAM	CONFIDENTIAL	EAST
ADMINISTRATION	AIR	POLITICAL	WEST
ECONOMIC	SAIGON	NORTH	SUMMIT
ADDITIONAL	AMEMBASSY	ATTACK	DATE
SUPPLEMENTAL	DEPT	SOUTH	GERMANY
CONFERENCE	LAO	GOVERNMENT	SOVIETS

Table 3: Most frequently used words for four topics from the Time-Duration-Topic model. The distribution of classification durations corresponding to these topics is plotted in figure 6

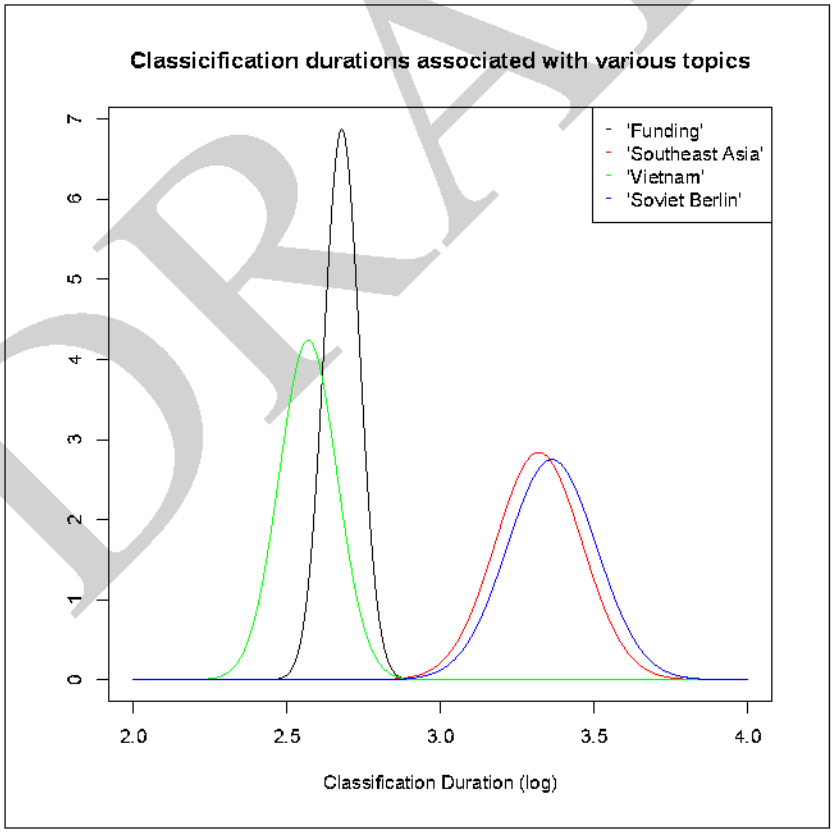

Figure 6: Fitted log-normal distributions over classification duration for four topics. The most common words for these topics are listed in table 3

versus ‘Southeast Asia’ provides new insight to historians, political scientists and watchdog groups interested in studying the openness and transparency of the US government in relation to historical events or investigating potential government coverups.

7 Conclusions

The work presented in this paper provides a brief discussion of methods for examining formerly-secret documents that have been declassified by the US government. We take advantage of the easy interpretability of statistical topic models to explore document content, while using methods taken from and inspired by survival analysis techniques to model temporal declassification patterns. In addition, we introduce a way of taking classification duration into account while inferring latent topics via statistical topic modeling. These techniques have the potential to contribute to a greater understanding for social scientists and policy-makers of classification policies and decisions. Although additional investigation is warranted regarding evaluation of these models and further ways of combining survival modeling and topic models, the initial results presented in this paper offer a new perspective on the vast quantities of information that are part of the US government’s Security Classification System.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, the University of Massachusetts Amherst College of Social and Behavioral Sciences and NSF grant #SBE-0965436. Any opinions, findings and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the sponsor.

References

Mikhail Atallah, Craig McDonough, Victor Raskin and Sergei Nirenburg. 2000. Natural language processing for information assurance and security: an overview and imple- mentations. *Proceedings of the 2000 workshop on new security paradigms*, pp. 51-65.

David Blei and Andrew Ng and Michael Jordan 2003. Latent Dirichlet Allocation *Journal of Machine Learning Research*, 3 pp. 993-1022.

Janet M. Box-Steffensmeier and Bradford S. Jones. 1997. Time is of the Essence: Event History Models in Political Science. *American Journal of Political Science*. 41(4), pp. 336-383.

1972. Regression Models and Life Tables. *Journal of the Royal Statistical Society: Series B* 34(2), pp. 187-220.

Anthony Fader, Dragomir Radev, Michael H. Crespin, Burt L. Monroe, Kevin M. Quinn and Michael Colaresi. 2010. MavenRank: Identifying Influential Members of the US Senate Using Lexical Centrality *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 658-666

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images *IEEE Transaction on Pattern Analysis and Machine Intelligence* 6 pp. 721-741.

Sean Gerrish and David Blei. 2010. The Ideal Point Topic Model: Predicting Legislative Roll Calls from Text. *Proceedings of the Computational Social Science and the Wisdom of Crowds Workshop*. Neural Information Processing Symposium.

Thomas Griffiths and Mark Steyvers 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1), pp. 5228-5235.

Justin Grimmer 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis* 18(1), pp. 1-35.

Justin Grimmer 2011. An Introduction to Bayesian Inference via Variational Approximations. *Political Analysis* 19(1), pp. 32-47.

Paul T. Jaeger. 2007. Information policy, information access, and democratic participation: The national and international implications of the Bush administration’s information politics. *Government Information Quarterly*. 24(4), pp. 840-859.

John P. Klein, Melvin L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer. New York, NY, USA.

Andrew McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Andrew McCallum, Xuerui Wang and Natasha Mohanty. 2007. Joint Group and Topic Discovery from Relations and Text. *Statistical Network Analysis: Models, Issues and New Directions, Lecture Notes in Computer Science* 4503, pp. 28-44.

Patrice McDermott and Amy Bennett. 2010. Secrecy Report Card 2010: Indicators of Secrecy in the Federal Government. http://www.openthegovernment.org/otg/SecrecyRC 2010.pdf.

Salvador Nieto Sánchez, Evangelos Triantaphyllou and Donald Kraft. 2002. A feature mining based approach for the classification of text documents into disjoint classes. *Information Processing & Management* 38(4) pp. 583-604.

David E. Pozen. 2005. The Mosaic Theory, National Security, and the Freedom of Information Act. *The Yale Law Journal*. 115(3), pp. 628-679.

Harold C. Relyea. 1999. Security classification reviews and the search for reform. *Government Information Quarterly*. 16(1), pp. 5-27.

Harold C. Relyea. 2003. Government secrecy: policy depths and dimensions. *Government Information Quarterly*. 20(4), pp. 395-418.

Xuerui Wang and Andrew McCallum. 2006. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. *Conference on Knowledge Discovery and Data Mining (KDD)*.