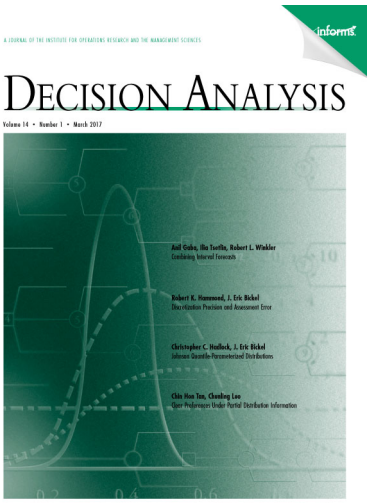

Decision Analysis

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Timely Decision Analysis Enabled by Efficient Social Media Modeling

Theodore T. Allen, Zhenhuan Sui, Nathan L. Parker

To cite this article:
Theodore T. Allen, Zhenhuan Sui, Nathan L. Parker (2017) Timely Decision Analysis Enabled by Efficient Social Media Modeling. Decision Analysis 14(4):250-260. https://doi.org/10.1287/deca.2017.0360

http://pubsonline.informs.org/journal/deca/

DECISION ANALYSIS
Vol. 14, No. 4, December 2017, pp. 250–260
ISSN 1545-8490 (print), ISSN 1545-8504 (online)

Timely Decision Analysis Enabled by Efficient Social Media Modeling

Theodore T. Allen,[a] Zhenhuan Sui,[a] Nathan L. Parker[b]

[a] Integrated Systems Engineering, Ohio State University, Columbus, Ohio 43210; [b] TRADOC Analysis Center, Monterey Naval Postgraduate School, Monterey, California 93943

Contact: allen.515@osu.edu, http://orcid.org/0000-0002-9522-3252 (TTA); sui.19@osu.edu (ZS); nparker@nps.edu (NLP)

Received: October 24, 2016
Revised: May 7, 2017; July 17, 2017
Accepted: August 11, 2017
Published Online in Articles in Advance: November 20, 2017

https://doi.org/10.1287/deca.2017.0360

Abstract. Many decision problems are set in changing environments. For example, determining the optimal investment in cyber maintenance depends on whether there is evidence of an unusual vulnerability, such as "Heartbleed," that is causing an especially high rate of incidents. This gives rise to the need for timely information to update decision models so that optimal policies can be generated for each decision period. Social media provide a streaming source of relevant information, but that information needs to be efficiently transformed into numbers to enable the needed updates. This article explores the use of social media as an observation source for timely decision making. To efficiently generate the observations for Bayesian updates, we propose a novel computational method to fit an existing clustering model. The proposed method is called *k*-means latent Dirichlet allocation (KLDA). We illustrate the method using a cybersecurity problem. Many organizations ignore "medium" vulnerabilities identified during periodic scans. Decision makers must choose whether staff should be required to address these vulnerabilities during periods of elevated risk. Also, we study four text corpora with 100 replications and show that KLDA is associated with significantly reduced computational times and more consistent model accuracy.

Funding: The authors thank the U.S. Army's TRADOC Analysis Center (TRAC) for supporting part of this research [Grant W9124N-15-T-0033]. The National Science Foundation [Grant 1409214] also supported part of this research.

Keywords: Bayes' theorem • applications: engineering • statistics • applications: security • applications

1. Introduction

Consider solving a similar problem for several decision periods. In each period, an updated model is used. Decisions in different periods are assumed here to have no interactive effect, and time enters as a factor only through changes to the environment for different periods. As a result, sequential decision policies, such as decision trees (e.g., Cao 2014) or continuous control policies (e.g., Borrero et al. 2015), are not instructive here. The motivating example is a cybersecurity investment decision (Paté-Cornell 2012, Gao et al. 2013, Parnell et al. 2015, Miller et al. 2016). This problem relates to monthly basic maintenance for periods of usual or, alternatively, elevated risk. Over 90% of breaches involve the exploitation of known vulnerabilities that have not been patched or remediated (Cockburn 2009). Organizations often exert intense efforts to address critical and high-level vulnerabilities but ignore medium vulnerabilities. Equilibrium game behavior is not relevant because, if all higher-level vulnerabilities are eliminated, the attacker has no recourse except to attempt to exploit lower vulnerabilities that remain. These vulnerabilities are more difficult to exploit and offer less access if exploited Mell et al. (2007).

Social media offer a useful source of timely information to update assumptions for many decision problems, for example, about the rate in which medium vulnerabilities are presently being exploited. The primary objective of this article is to provide computationally efficient and repeatable methods for updating period-specific decision models using Twitter or other streaming text data. Streaming social media data contain timely and valuable warnings about threats. For

example, experts on Twitter mentioned the medium vulnerability Heartbleed before more than 85% of the associated warnings/incidents appear in our university data set. Therefore, instead of using social media as an application area for decision analyses (Charalabidis and Loukis 2012, Bok et al. 2012), we seek to use it as a source of timely data for decision analysis problems. In a few periods in our motivating example, vulnerabilities constituted a large fraction of cybersecurity expert tweets. These were precisely the periods in which the most warnings and incidents were eventually observed. By monitoring tweets, it was possible for system administrators to anticipate and mitigate the attacks that followed. We expect that social media-based monitoring could aid updating for many types of decision analysis problems.

The method must be computationally efficient because text processing can be prohibitively slow (e.g., see Blei et al. 2003, Packiam and Prakash 2015). For our purposes, any method that transforms streaming text into numbers that strongly correlate with the system state could be used. For example, one might use sentiment analysis, which scores positive and negative words or even simpler counts of word mentions. Here, we use clustering methods primarily because the Twitter experts wrote about many subjects unrelated to medium vulnerabilities. Through clustering, all their topics can be mapped, including those that relate to the decision problem. We seek to generate useful data by recalling tweets from the key clusters.

Probably the most widely studied methods for clustering text data are variants of latent Dirichlet allocation (LDA) or "topic models" (Blei et al. 2003, Packiam and Prakash 2015). There are several ways to fit topic models to data, including collapsed Gibbs sampling, a form of Markov chain Monte Carlo simulation (Teh et al. 2006; Allen 2011, p. 14), and "mean field variational inference" (Blei et al. 2003), an approximate maximum likelihood fit of the clustering (distribution) model.

Yet, both collapsed Gibbs sampling and variational inference can be prohibitively expensive computationally for corpora involving tens of thousands of documents. Collapsed Gibbs is known for its lack of repeatability. Here, we seek computationally efficient methods to fit approximate topic models with improved repeatability. Specifically, we propose to explore the concept of transforming k-means clustering results to estimate topic model parameters. Lee (2012) had used fuzzy c clustering to generate "fuzzy LDA," which permits documents that cover multiple topics like LDA and unlike k-means clustering. Yet Ghosh and Dubey (2013) show that k-means scaled more efficiently than fuzzy c clustering.

The remainder of this article is organized as follows. First, we describe the time decision analysis formulation. Then, we describe the proposed methods for efficient clustering needed to generate the decision formulation inputs. Next, we compare the proposed estimation methods with alternatives. Finally, we illustrate the methods on a cyber investment problem and conclude with a summary of the results and future work possibilities.

2. Timely Decision Modeling

Consider a two-phase approach for estimating the probabilities in our decision problem. The first phase is a startup phase in which the model is estimated and matrices are estimated to facilitate Bayesian updates. The second phase is steady state in which new text data are analyzed and Bayesian updates potentially change the results for subsequent decision problems. In each period, the decision maker observes the system state from social media, then chooses an action. The Bayesian updates require the collection of observation data and the estimation of "observation matrices" (Smallwood and Sondik 1973), both of which steps we describe.

2.1. Two-Phase Approach

In the first phase, data are gathered carefully such that the true state of the system can be assumed to be known. We denote the system state as Y with possible values $y = 1, \ldots, s$ and the chosen action in period i is $a_i = 1, \ldots, a$. The state is independent of the action and observed before the action selection. The reward depends on the action and state and is $r[(y, a_i)]$, and the utility function is $u[r(y, a_i)]$. The current probability distribution for the state in period i is $p_i(y, a_i)$, and the initial probability distribution is $p_0(y, a_0)$. In each time period, the decision maker selects the option that maximizes the expected utility given by

$$\max_{a_i} E[u(a_i)] = \sum_{y=1}^{s} p_i(y, a_i) u[r(y, a_i)], \tag{1}$$

which is essentially the Von Neumann and Morgenstern (2007) problem. Note that the utility could be equivalently applied to each reward, state, and action set, resulting in a simplified exposition.

In the cybersecurity investment context, the model in Equation (1) is analogous to the model by Parnell et al. (2015). An exception is that the probability distribution may depend on the time period, i, as for time-dependent formulations (DeGroot 2005). This time-dependence persists throughout the observation, O, which is here assumed to be one of m levels, that is, $O \in \{1,\ldots,m\}$. The key idea here is that the social media text is converted to a series of observations, $o_1, o_2 \ldots$, one for each period, with relevance to the decision problem. Then the probabilities are updated using Bayes' theorem:

$$p_i(y, a_i \mid O = o) = \frac{p_0(y, a_i)p(O = o \mid y, a_i)}{\sum_{y=1}^{s} p_0(y, a_i)p(O = o \mid y, a_i)}, \quad (2)$$

where p_0 is the initial or prior probability, and the so-called "observation" matrix is $p(O = o \mid y, a_i)$ for indices $o = 1,\ldots,m$, $y = 1,\ldots,s$, and each possible action a_i. Establishing the prior during the "burn in" Phase 1 is part of preparing for continuing fluctuations in Phase 2. The formulation in Equations (1) and (2) is relevant for problems in which the system resets between periods, a phenomenon that applies only approximately to our cybersecurity case study.

The objective of the startup phase is to estimate the observation matrix, $p(O = o \mid y, a_i)$, using training data. Then, in steady state (Phase 2), the analysis method is used to provide observations, O, from the social media. The true state is not known, only the observation values. Updates are performed using Equation (2), and the result is used to solve Equation (1) to generate the optimal action for the relevant time period i. In each period, action follows the observation.

2.2. Observations and Observation Matrices

The following sections describe a computationally efficient method to derive random observations $O_1,\ldots,O_n$ over n periods for which the system states $y_1,\ldots,y_n$ are assumed known for known actions $a_1,\ldots,a_n$. Counts for the number of times an observation was observed in each state are $C_{o,a,y}$ for $o = 1,\ldots,m$, $a_i = 1,\ldots,v$, and $y = 1,\ldots,s$. Then, the observation matrix, $p(O \mid y, a_i)$ is estimated using

$$p(O = o \mid y, a_i) = \frac{C_{o,a,y}}{\sum_{o'=1}^{m} C_{o',a,y}} \quad \text{for } o = 1,\ldots,m, \; a_i = 1,\ldots,a, \text{ and } y = 1,\ldots,s, \quad (3)$$

which derives the standard frequentist probability estimates. Observation matrices are displayed for each action, a_i with rows corresponding to states, y, and columns corresponding to observation levels (Smallwood and Sondik 1973). Observations are informative about the system state if the probabilities have dissimilar values along the columns of the observation matrices. Then, if the relevant observation level occurs, the Bayesian update in Equation (2) generates a high probability that the system is in a specific state.

3. Efficient Methods for Obtaining Observations from Social Media

In this section, we review the LDA model, which is a probability distribution from Blei et al. (2003). Then, we review the associated estimation methods from Blei et al. (2003), Teh et al. (2006), and Griffiths and Steyvers (2004). In the next section, we propose a new estimation method based on transforming a k-means clustering model into an LDA model.

Note that virtually all text-modeling methods begin with a natural language-processing step in which text is transformed into numbers with irrelevant words removed and words "stemmed" (e.g., "jumping" and "jumps" are both shortened to "jump," see Feldman and Sanger 2007, Porter 1980).

3.1. Latent Dirichlet Allocation

Our notation follows Blei et al. (2003) and Carpenter (2010) so that $w_{d,j}$ is the jth word in the dth document with $d = 1,\ldots,D$ and $j = 1,\ldots,N_d$. Therefore, "D" is the number of documents or tweets, and "N_d" is the number of words in the dth document. We transform words into numbers using the method of Porter (1980). Therefore, $w_{d,j} \in \{1,\ldots,W\}$, where W is the number of distinct words in all documents.

The clusters or "topics" are defined by the estimated probabilities, $\hat{\phi}_{t,c}$, that a randomly selected word in cluster $t = 1,\ldots,T$ (on that topic) is the word $c = 1,\ldots,W$. The value $\hat{\theta}_{d,t}$ represents the estimated probability a randomly selected word in document d is

assigned to cluster t of the T possible. Estimating the $\hat{\phi}_{t,c}$ and $\hat{\theta}_{d,t}$ for $t = 1,\ldots,T$, $d = 1,\ldots,D$ and $c = 1,\ldots,W$ permits estimation of the observations needed for our timely decision analysis problem. This follows because we are interested in clusters or topics related to our problem by the probabilities, $\phi_{t,c}$ and periods in which the document probabilities, $\theta_{d,t}$, on these topics are high. The model variables $z_{d,j}$ are the cluster assignments for each word in each document, $d = 1,\ldots,D$ and $j = 1,\ldots,N_d$.

Generally, low values or diffuse prior parameters α and β are applied (Griffiths and Steyvers 2004). Note that these priors are relevant to Bayesian estimation of LDA only. The joint probability of the data, $w_{d,j}$, and the parameters to be estimated, $(z_{d,j}, \theta_{d,t}, \phi_{t,c})$, are provided in many references, including Carpenter (2010). The key quantities to be estimated in the distribution are the counts of words on topic t in document d, $n_t^{(d)}$, given by

$$n_t^{(d)} = \sum_{j=1}^{N_d} \sum_{c'=1}^{W} I(z_{d,j} = t \text{ and } c = c'), \tag{4}$$

and the number of instances of word c with topic t, $n_t^{(c)}$, given by

$$n_t^{(c)} = \sum_{d=1}^{D} \sum_{j=1}^{N_d} I(z_{d,j} = t \text{ and } w_{d,j} = c), \tag{5}$$

where $I(\ldots)$ is an indicator function giving 1 if the equalities hold and 0 otherwise.

Note Equation (4) is a simple representation of human speech in which words, $w_{d,j}$, are multinomial draws associated with given topics, $z_{d,j}$, which are also multinomial draws. The probabilities, $\phi_{t,c}$, that define the topics are also random; that is, it is a hierarchical distribution. Technically, the estimates that are often used for these probabilities are Monte Carlo estimates for the posterior means of the Dirichlet distributed probabilities, $\hat{\phi}_{t,c}$.

Once the parameters $\hat{\phi}_{t,c}$ and $\hat{\theta}_{d,t}$ have been estimated, the derivation of the observations is relatively easy. Studying the estimated posterior mean probabilities of $\phi_{t,c}$, the clusters or topics (t) relevant to the decision problem are identified. Then, retrieving the documents on these topics with values of $\hat{\theta}_{d,t}$ that exceed a threshold in each time period gives the needed observation counts, $O_1,\ldots,O_n$. For example, if there are many tweets on cyber vulnerabilities, the period is likely associated with elevated threats necessitating additional investment.

3.2. Collapsed Gibbs Sampling

Perhaps the most popular way to estimate the parameters in the LDA model in Equation (4) is called "collapsed Gibbs" sampling (Teh et al. 2006, Griffiths and Steyvers 2004). To implement collapsed Gibbs, the values of the topic assignments for each word, $z_{d,j}$, are sampled uniformly. Then, iteratively, multinomial samples are drawn for each topic assignment $z_{d,j}$ iterating through each document, d, and word, j, using the last iterations of all other assignments, $z_{-(d,j)}$. The multinomial draw probabilities are given in Teh et al. (2006). In the collapsed Gibbs sampling method, each word is randomly assigned to a cluster with probabilities proportional to the counts for that word being assigned multiplied by the counts for that document being assigned. After M iterations, the last set of topic assignments generates the counts and the estimated posterior means:

$$\hat{\phi}_{t,c} = \frac{n_t^{(c)} + \beta}{n_t^{(\cdot)} + W\beta} \tag{6}$$

and the posterior mean topic definitions using

$$\hat{\theta}_{d,t} = \frac{n_t^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}. \tag{7}$$

Therefore, if words are assigned commonly to certain topics by the Gibbs sampling chain, their frequency increases the posterior probability estimates both in the topic definitions, $\hat{\phi}_{t,c}$, and the document probabilities $\hat{\theta}_{d,t}$. From $\hat{\theta}_{d,t}$, we can see periods when certain topics dominate.

4. k-Means-Based Latent Dirichlet Allocation (KLDA)

Gibbs sampling is noisy and inefficient since only a single iteration of topic assignments is used for the posterior estimates, and even approximate convergences can require thousands or millions of iterations. The proposed estimation method clusters documents. This is different from LDA, which permits documents to have specific words on multiple topics. Yet, for short documents such as tweets, the difference may be considered unimportant, and robustness is explored in Section 6.

Denote the word counts for each document, d, and word, c, as $X_{d,c}$. The standard k-means clustering in our notation is Lloyd (1982):

1. Select T documents, $d_1, \ldots, d_T$, uniformly from $\{1, \ldots, D\}$. Initialize the cluster centroids using $q_{t,c} = X_{d_t,c}$ for $c = 1, \ldots, W$ and $t = 1, \ldots, T$.

2. Compute the distances for each document to each centroid using

$$v_{d,t} = \sqrt{\sum_{c=1}^{W} (q_{t,c} - X_{d,c})^2} \quad \text{for } t = 1, \ldots, T,\ d = 1, \ldots, D. \tag{8}$$

3. Assign each document to a cluster, $\tilde{z}_d$, using

$$\tilde{z}_d = \arg\min_t v_{d,t} \quad \text{for } d = 1, \ldots, D. \tag{9}$$

The set S_t contains documents with $\tilde{z}_d = t$ for $t = 1, \ldots, T$.

4. Update the centroids using the average locations for documents in the cluster:

$$q_{t,c} = \frac{\sum_{d \in S_t} X_{d,c}}{|S_t|}. \tag{10}$$

5. Repeat steps 2 through 4 until the cluster assignments do not change.

A last step is added to permit fractional membership in clusters by documents and facilitate the interpretation as a topic model. The "membership" function, similar to fuzzy-c clustering (as distinct from fuzzy decision making), is

$$u_{d,t} = \quad 1/v_{d,t} \quad \text{for } t = 1, \ldots, T,\ d = 1, \ldots, D. \tag{11}$$

This permits estimation of the document topic probabilities using

$$\hat{\theta}_{d,t} = \frac{u_{d,t}}{\sum_{d'=1}^{D} u_{d',t}} \quad \text{for } t = 1, \ldots, T,\ d = 1, \ldots, D. \tag{12}$$

Also, the estimated topic definitions are generated using

$$\hat{\phi}_{t,c} = \frac{q_{t,c}}{\sum_{c'=1}^{W} q_{t,c'}} \quad \text{for } t = 1, \ldots, T \text{ for } c = 1, \ldots, W \tag{13}$$

as the topic proportions, which show the distribution of topics in all the document lists. Clearly, if the documents are long and cover many substantially different topics, the approximation will be poor. We explore the robustness computationally in Section 6. Intuitively, the memberships in Equation (11) are, for short documents at least, approximately proportional to the counts in Equations (4) and (5). Therefore, the ratios in Equations (12) and (13) are like the Bayesian estimates in Equations (6) and (7).

5. Numerical Studies

In this section, a computational comparison of Gibbs sampling and KLDA is provided. Four test corpora drawn from Allen et al. (2016) include two having multiple topics per document, permitting the sensitivity of KLDA performance to be studied. The purpose of this step is to clarify the computational and accuracy advantages of the alternative estimation methods.

5.1. Test Problems

In this section, four similar cases are studied to compare different estimation methods. To preview, Table 1 summarizes the results of the computational run times. Table A.1 (in the appendix) shows the four similar cases in which 40 documents are studied so that $D = 40$ for each case. Table A.2 (in the appendix) shows the true model topic proportion and topic definition, where topic number $T = 5$ for cases 1 and 2 and $T = 6$ for cases 3 and 4. In general, "true topics" are possible because they can be used to generate the documents. In this case, they are simply assumed. The dictionary size for all the cases is $W = 25$. This is a "robustness" study because four cases span a variety of cases in terms of topic diversity and overlap.

5.2. Evaluation Metrics

Because the estimated distribution topics have no natural ordering, it is hard to compare the result against the assumed ground truth. Therefore, Steyvers and Griffiths (2007) proposed that the permutations of cluster labels should be considered and the closest "distance" permutation should be selected. Define the function $t'(\mathbf{r}, t)$ as the selection of topic t in permutation $\mathbf{r}$. Use $\phi^{\text{true}}_{t,c}$ to denote the ground truth topic definitions for $t = 1, \ldots, T$ and for $c = 1, \ldots, W$. In the appendix, the ground truth is provided for one of the four cases. For all cases, see Allen et al. (2016). Further, denote $\mathbf{r}^*$ as the argmax permutation for Equation (13). The accuracy measure used here is the average root mean squared (RMS):

$$\text{RMS}(\phi) = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{c=1}^{W} (\phi^{\text{true}}_{t,c} - \phi_{t'(\mathbf{r}^*,t),c})^2}. \tag{14}$$

Intuitively, the RMS value indicates the typical size of errors in the topic definition estimation.

5.3. Comparison Results

Table 1 contains the comparison results of *k*-means LDA, Gibbs Sampling LDA with 10, 100, and 1,000 runs. Each value in the table is the average RMS for 100 replications, that is, starting from distinct random seeds. Using RMS metrics, *k*-means LDA could achieve a similar level of distance or even a smaller distance to the true model compared with other models. This holds even if there are multiple topics in each document (case 3 and case 4). For Gibbs sampling, Monte Carlo simulation introduces uncertainties. A higher number of iterations gives slightly better RMS than lower numbers, but the quality is highly influenced by the random seed. Table 1 gives the timing for estimation methods. Clearly, KLDA is significantly more efficient with comparable quality. It permits our VBA software to analyze 10,000 tweets in less than 20 minutes on an i5 processor.

6. Cybersecurity Twitter-Enabled Study

In this section, we use a routine decision problem faced by many organizations to illustrate the application of the formulation, modeling of social media data, observations, and results (Afful-Dadzie and Allen 2014). With only two actions and two states, this problem is simple and illustrative. The same approach could be applied to problems with more states, actions, and multiple sets of noninteracting time periods. Yet the authors are aware of an organization that suffered losses perhaps exceeding $1 M because of failure to solve this problem optimally. Often, organizations do not attempt to patch medium-level cyber vulnerabilities. Patching requires staff time and can cause disruptions because some software may not work after patching actions.

Yet, during times of elevated risks resulting from exceptionally problematic medium-level vulnerabilities, adjustments are potentially relevant. Also, in these cases, the actions of administrators do not affect the threat level, but only the rewards (or losses). This simplifies our formulation in Equation (1) since the probabilities do not depend on the actions. Experts tweet on Twitter continually on many subjects relevant to decision problems. The experts cover many topics, and there are hundreds of potentially relevant medium-level vulnerabilities. Continued discussion of a medium vulnerability by experts is likely an indicator of an elevated risk state.

Here, we study $D = 16{,}047$ tweets starting in January 2014 for 12 months from 16 selected Twitter accounts on multiple top 10 lists relating to cybersecurity: Mathewjschwartz, Neilweinberg, Scotfinnie, Secureauth, Lennyzeltser, Dangoodin001, Dstrom, Securitywatch, Cyberwar, Jason_Healey, FireEye, Lancope, Varonis, DarkReading, RSAsecurity, and Mcafee_Labs. The decision problem includes $s = 2$ states (normal and elevated risk), $a = 2$ actions (1–do not patch medium-level vulnerabilities, 2–patch medium-level vulnerabilities).

Table 1. Computational Accuracy (RMS) and Timing Results for the Case Studies

Case	Test model	Iterations	Average RMS	Std RMS	100 replicates time (sec)
1	*k*-means LDA	2	0.0453	0.0000	5
1	Gibbs sampling LDA	10	0.0507	0.0098	4
1	Gibbs sampling LDA	100	0.0451	0.0089	44
1	Gibbs sampling LDA	1,000	0.0436	0.0064	323
2	*k*-means LDA	2	0.0500	0.0000	5
2	Gibbs sampling LDA	10	0.0531	0.0076	6
2	Gibbs sampling LDA	100	0.0492	0.0063	43
2	Gibbs sampling LDA	1,000	0.0492	0.0049	301
3	*k*-means LDA	2	0.0401	0.0000	6
3	Gibbs sampling LDA	10	0.0482	0.0093	6
3	Gibbs sampling LDA	100	0.0416	0.0063	56
3	Gibbs sampling LDA	1,000	0.0409	0.0046	489
4	*k*-means LDA	2	0.0450	0.0000	6
4	Gibbs sampling LDA	10	0.0519	0.0080	7
4	Gibbs sampling LDA	100	0.0456	0.0075	59
4	Gibbs sampling LDA	1000	0.0459	0.0053	485

Table 2. Posterior Mean Topic Definitions, $\hat{\phi}_{t,c}$, Estimates from KLDA with 17 on Heartbleed

T1	0.0567	T2	0.0540	...	T17	0.0500	...
Word	Prob	Word	Prob	...	Word	Prob	...
(frequency) low	0.1393	(text) rt	0.1095	...	(name) mathewjschwartz	0.1040	...
(name) cyberwar	0.0291	(frequency) low	0.0927	...	(frequency) low	0.1018	...
(name) dangoodin001	0.0264	(name) dangoodin001	0.0183	...	(text) infosec	0.1009	...
(name) darkread	0.0182	(name) cyberwar	0.0169	...	(text) atinformationweek	0.0420	...
(month) 3	0.0164	(month) 2	0.0166	...	(frequency) medium	0.0192	...
(month) 2	0.0162	(name) securitywatch	0.0165	...	(text) breach	0.0152	...
(month) 4	0.0154	(frequency) high	0.0132	...	(text) new	0.0152	...
(month) 1	0.0153	(name) jasonhealei	0.0124	...	(text) risk	0.0147	...
(name) securitywatch	0.0135	(name) mcafeelab	0.0123	...	(text) malwar	0.0129	...
(month) 5	0.0131	(month) 3	0.0123	...	(month) 4	0.0125	...
(month) 8	0.0127	(text) secur	0.0112	...	(month) 5	0.0121	...
(month) 7	0.0120	(month) 1	0.0112	...	(text) attack	0.0121	...
(name) jasonhealei	0.0116	(text) atdavemarcu	0.0104	...	(text) hack	0.0116	...
(month) 6	0.0113	(month) 4	0.0100	...	(month) 6	0.0098	...
(month) 12	0.0099	(month) 8	0.0099	...	(text) secur	0.0098	...
(name) mcafeelab	0.0091	(month) 6	0.0094	...	**(text) heartble**	0.0098	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note. The words associated with medium-level cyber vulnerabilities are in bold.

We assume that the system was in state 1 except for four months starting in April as indicated in Table 3(a) because of the announcement of the well-known Heartbleed vulnerability. The database has $W = 894$ nonrare and not-stopping distinct words, that is, common words excluding articles, prepositions, and other relatively uninformative words.

Applying k-means-based LDA, one topic (T17) is identified as related to cyber vulnerabilities in general and Heartbleed. It is the only topic for which one of the top 20 defining words is a medium vulnerability. The stemmed results for the top words generated using Equation (13) are shown in Table 2. Note how obscure our decision problem is with so much discussion being largely irrelevant and the need for filtering.

Then, KLDA identifies the top 20 documents by posterior mean estimate, $\hat{\theta}_{d,t}$, for each of the 12 months (not shown). Inspecting these tweets manually and tabulating relevant mentions of Heartbleed (or any other medium vulnerability) resulted in the raw mentions in Table 3(a). In most periods, medium vulnerabilities received no mentions. Yet, when there is a mention

Table 3. (a) States $y_1, \ldots, y_{12}$, Raw Observations, and $O_1, \ldots, O_{12}$; (b) Counts $C_{O,a,y}$; (c) Observation Matrices, $p(o \mid y, a)$; and (d) Posterior Values, $p(y, a \mid O)$, for Different Observation Levels

(a)				(b)			(c)		(d)			
Months	System state	Raw mentions	Observation	State	1 (0)	2 (>0)	1 (0)	2 (>0)	State	p_0	1 (0)	2 (>0)
1	1	0	1	State 1	1	3	0.250	0.750	State 1	0.333	0.125	0.750
2	1	0	1	State 2	7	1	0.875	0.125	State 2	0.667	0.875	0.250
3	1	0	1									
4	2	7	2									
5	2	4	2									
6	2	1	2									
7	2	0	1									
8	1	0	1									
9	1	1	2									
10	1	0	1									
11	1	0	1									
12	2	0	1									

of a medium-level vulnerability, for example, Heartbleed, the count of tweets is increased. For simplicity, observations are divided into two levels, that is, level 1—zero mentions of Heartbleed or level 2—greater than zero mentions. This results in the observations $O, \ldots, O_{12}$, and cross-tabulating generates the counts in Table 3(b) $C_{O,a,y}$. There are eight total counts of observation 1 and four total of observation 2. The frequentist estimates for the observation matrices are given in Table 3(c). The prior values and posterior estimates from Equation (2) are provided for different observations in Table 3(d).

We assume that attempting to patch medium vulnerabilities will reduce the number of successful intrusions. We assume rewards of $r(y=1, a_i=1) = -\$300{,}000$, $r(y=2, a_i=1) = \$100{,}000$, $r(y=1, a_i=2) = -\$200{,}000$, and $r(y=2, a_i=2) = -\$50{,}000$ and the exponential utility function $u(r) = 1 - e^{r/r_0}$ from Kirkwood (1997, p. 22). To represent moderate risk aversion, we assume that the reference parameter r_0 is \$200,000, which is smaller than the highest rewards in absolute value. If we observe $O_i = 1$ (no mentions of medium vulnerabilities), the expected utilities are $E[(a_i=1)] = 0.125 \times (-3.48) + 0.875 \times 0.393 = -0.090$ and $E[u(a_i=2)] = 0.125 \times (-1.72) + 0.875 \times 0.284 = -0.4633$. With observation $O_i = 2$ (mentions of medium vulnerabilities), the expected utilities are $E[(a_i=1)] = 0.75 \times (-3.48) + 0.25 \times 0.393 = -2.513$ and $E[(a_i=2)] = 0.75 \times (-1.72) + 0.25 \times 0.284 = -1.360$. Therefore, if the experts tweet about medium vulnerabilities, the optimal action is patching. Otherwise, patching is not recommended. Smaller values of the reference value r_0 correspond to more risk-averse decision makers (Kirkwood 1997). The threshold level is \$115,067. For smaller values, patching medium vulnerabilities is always recommended.

This example illustrates how social media analytics can inform timely decision problems. Note that our default model assumes that the prior for each month resets. For cyber maintenance decision making, this is justified by the fact that the cycle of exploitation and patching has a finite duration. Vulnerabilities such as Heartbleed become known and, after a period, almost all systems are patched with little period-to-period dependence. Admittedly, the time scale of the reset might be longer than a single month and carryover effects of patching could be important. Fully sequential methods using, for example, partially observable Markov decision processes (POMDP), are proposed as a topic for future research.

7. Conclusions and Future Research

In this article, we proposed a method to link social media analytics with routine decision analyses. We also proposed an innovative topic estimation technique based on k-means clustering called KLDA. This permits the rapid estimation of LDA models. The latter incorporate human high-level domain knowledge so that users can direct or perturb the model and results. Applying the techniques to test problems, we demonstrated that KLDA can achieve improved repeatability and comparable subjective accuracy. Specifically, we used four cases to test our new model against the true models. The improved efficiency is important for enabling spreadsheet applications, allowing users to benefit from text processing and information retrieval for private text corpora.

Yet a number of topics remain for future study. Problems in which the current state selection may depend on previous states can potentially be investigated by simply using the current probabilities for the update in the next period using partially observable Markov decision process (POMDP) formulations. Incorporating risk aversion in multi-period decision making is an active area of research (Homem-de-Mello and Pagnoncelli 2016); however, other techniques besides k-means-based estimation, such as fuzzy c clustering, can be explored. Also, additional comparison metrics and test cases might better clarify the accuracy limitations of KLDA methods. New evaluation metrics could be more objective and interpretable than RMS. Currently, the computational experiments involve only small test corpora from Allen et al. (2016). Larger corpora from the literature can be explored. Methods that permit experts to edit topics offer the promise of more informative observations (Zhao et al. 2012, Sun 2014, Allen et al. 2016, Sui et al. 2015). Timely pricing enabled by social media analysis and local elicitation can also be investigated (Allen and Maybin 2004).

Acknowledgments

The authors thank the associate editor and the anonymous reviewers for their helpful ideas, which greatly improved the manuscript. Please note that the views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Army, the Department of Defense, or the U.S. Government.

Appendix. Numerical Study Details

This appendix contains data for the case studies including the true model, which originally appeared in Allen et al. (2016).

Table A.1. Synthetic Data for the Numerical Example

Doc	Document
1	The operator cut aluminum and dropped it at station1.
2	The inspector drilled plastic and overheated it at station2.
3	The manager milled steel and misaligned it at station3.
4	The engineer saw stone and over torqued on the truck.
5	The supplier welded and misdimensioned the titanium offsite.
6	The inspector drilled plastic and overheated it at station2.
7	It was drilled and overheated.
8	It was drilled and overheated.
9	The engineer and the manager at station3 and on the truck.
10	The welded titanium was misdimensioned.
11	The titanium was welded and misdimensioned offsite.
12	The steel was misdimensioned.
13	The operator cut the steel and plastic.
14	The manager welded it and misdimensioned it.
15	The operator cut and dropped the aluminum at station1.
16	The operator cut and dropped it at station1.
17	The engineer welded and misdimensioned the titanium.
18	It was drilled and overheated.
19	It was drilled and overheated.
20	The manager milled steel and misaligned it at station3.
21	The operator cut and dropped the steel at station1.
22	The engineer and the manager at station3 and offsite.
23	It was drilled and overheated.
24	The engineer saw stone and over torqued on the truck.
25	The stone was drilled and overheated.
26	It was drilled and overheated.
27	It was drilled and overheated.
28	It was drilled and overheated offsite.
29	The supplier welded titanium and misdimensioned it offsite.
30	The operator cut and dropped the titanium at station1.
31	The operator cut and dropped it at station1.
32	It was steel.
33	The steel was drilled and overheated.
34	It was drilled and overheated at station3.
35	The engineer and the manager at station1 and on the truck.
36	The welded titanium was misdimensioned.
37	It was drilled and overheated.
38	It was drilled and overheated.
39	The supplier welded titanium and misdimensioned it offsite.
40	It was drilled and overheated.

Table A.2. Assumed Ground Truth for the Numerical Examples

T1	0.4	T2	0.2	T3	0.15	T4	0.125	T5	0.125
Word	Prob	Word	Prob	Word	Prob	Word	Prob	Word	Prob
Oper	0	Oper	0	Oper	0.23	Oper	0	Oper	0
Cut	0	Cut	0	Cut	0.23	Cut	0	Cut	0
Aluminum	0	Aluminum	0	Aluminum	0.08	Aluminum	0	Aluminum	0
Drop	0	Drop	0	Drop	0.23	Drop	0	Drop	0
Station1	0	Station1	0	Station1	0.23	Station1	0	Station1	0

Table A.2. (Continued)

T1	0.4	T2	0.2	T3	0.15	T4	0.125	T5	0.125
Word	Prob	Word	Prob	Word	Prob	Word	Prob	Word	Prob
Inspector	0.1	Inspector	0	Inspector	0	Inspector	0	Inspector	0
Drill	0.35	Drill	0	Drill	0	Drill	0	Drill	0
Plastic	0.1	Plastic	0	Plastic	0	Plastic	0	Plastic	0.1
Overh	0.35	Overh	0	Overh	0	Overh	0	Overh	0
Station2	0.1	Station2	0	Station2	0	Station2	0	Station2	0
Manag	0	Manag	0	Manag	0	Manag	0.25	Manag	0.1
Mill	0	Mill	0	Mill	0	Mill	0	Mill	0.1
Steel	0	Steel	0	Steel	0	Steel	0	Steel	0.5
Misalign	0	Misalign	0	Misalign	0	Misalign	0	Misalign	0.1
Station3	0	Station3	0	Station3	0	Station3	0.25	Station3	0.1
Engin	0	Engin	0	Engin	0	Engin	0.25	Engin	0
Saw	0	Saw	0	Saw	0	Saw	0	Saw	0
Stone	0	Stone	0	Stone	0	Stone	0	Stone	0
Overtorqu	0	Overtorqu	0	Overtorqu	0	Overtorqu	0	Overtorqu	0
Truck	0	Truck	0	Truck	0	Truck	0.25	Truck	0
Supplier	0	Supplier	0.05	Supplier	0	Supplier	0	Supplier	0
Weld	0	Weld	0.3	Weld	0	Weld	0	Weld	0
Misdimens	0	Misdimens	0.3	Misdimens	0	Misdimens	0	Misdimens	0
Titanium	0	Titanium	0.3	Titanium	0	Titanium	0	Titanium	0
Offsit	0	Offsit	0.05	Offsit	0	Offsit	0	Offsit	0

Notes. These are the assumed probabilities that specific words will be generated if specific topics are selected and the chance that a random word is on each topic.

References

Afful-Dadzie A, Allen TT (2014) Data-driven cyber-vulnerability maintenance policies. *J. Quality Tech.* 46(3):234–250.

Allen TT (2011) *Introduction to Discrete Event Simulation and Agent-based Modeling: Voting Systems, Health Care, Military, and Manufacturing* (Springer, London).

Allen TT, Maybin KM (2004) Using focus group data to set new product prices. *J. Product Brand Management* 13(1):15–24.

Allen TT, Xiong H, Afful-Dadzie A (2016) A directed topic model applied to call center improvement. *Appl. Stochastic Models Bus. Indust.* 32(1):57–73.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3(January):993–1022.

Bok HS, Kankanhalli A, Raman KS, Sambamurthy V (2012) Revisiting media choice: A behavioral decision-making perspective. *Internat. J. e-Collaboration (IJeC)* 8(3):19–35.

Borrero JS, Prokopyev OA, Sauré D (2015) Sequential shortest path interdiction with incomplete information. *Decision Anal.* 13(1):68–98.

Cao Y (2014) Reducing interval-valued decision trees to conventional ones: Comments on decision trees with single and multiple interval-valued objectives. *Decision Anal.* 11(3):204–212.

Carpenter B (2010) Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling. Revision 1.4, LingPipe, Inc. Accessed November 1, 2017, http://lingpipe.files.wordpress.com/2010/07/lda1.pdf.

Charalabidis Y, Loukis E (2012) Participative public policy making through multiple social media platforms utilization. *Internat. J. Electronic Government Res.* 8(3):78–97.

Cockburn E (2009) Websites here, websites there, websites everywhere . . ., but are they secure? *Quaestor Quart.* 4(3):1–4.

DeGroot MH (2005) *Optimal Statistical Decisions*, Vol. 82 (John Wiley & Sons, New York). [Reprint 1970.]

Feldman R, Sanger J (2007) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (Cambridge University Press, New York).

Gao X, Zhong W, Mei S (2013) Information security investment when hackers disseminate knowledge. *Decision Anal.* 10(4):352–368.

Ghosh S, Dubey SK (2013) Comparative analysis of *k*-means and fuzzy *c*-means algorithms. *Internat. J. Adv. Comput. Sci. Appl.* 4(4):35–39.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci.* 101(S1):5228–5235.

Homem-de-Mello T, Pagnoncelli BK (2016) Risk aversion in multistage stochastic programming: A modeling and algorithmic perspective. *Eur. J. Oper. Res.* 249(1):188–199.

Kirkwood CW (1997) *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets* (Duxbury Press, Belmont, CA).

Lee SH (2012) Comparison and application of probabilistic clustering methods for system improvement prioritization. Unpublished doctoral dissertation, Ohio State University, Columbus.

Lloyd SP (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theory* 28(2):129–137.

Mell P, Scarfone K, Romanosky S (2007) A complete guide to the common vulnerability scoring system version 2.0. Accessed November 14, 2017, http://www.nazimkaradag.com/wp-content/uploads/2014/11/cvss-guide.pdf.

Miller S, Wagner C, Aickelin U, Garibaldi JM (2016) Modelling cybersecurity experts' decision making processes using aggregation operators. *Comput. Security* 62(September):229–245.

Packiam RM, Prakash VSJ (2015) An empirical study on text analytics in big data. *2015 IEEE Internat. Conf. Comput. Intelligence Comput. Res. (ICCIC)* (IEEE, Madurai, India), 1–4.

Parnell GS, Butler III RE, Wichmann SJ, Tedeschi M, Merritt D (2015) Air Force cyberspace investment analysis. *Decision Anal.* 12(2):81–95.

Paté-Cornell ME (2012) Games, risks, and analytics: Several illustrative cases involving national security and management situations. *Decision Anal.* 9(2):186–203.

Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3): 130–137.

Smallwood RD, Sondik EJ (1973) The optimal control of partially observable Markov processes over a finite horizon. *Oper. Res.* 21(5):1071–1088.

Steyvers M, Griffiths T (2007) Probabilistic topic models. *Handbook Latent Semantic Anal.* 427(7):424–440.

Sui Z, Milam D, Allen TT (2015) A visual monitoring technique based on importance score and twitter feeds. Working paper, Ohio State University, Columbus.

Sun X (2014) Textual document clustering using topic models. *Semantics, Knowledge Grids (SKG), 2014 10th Internat. Conf.* (IEEE, Beijing), 1–4.

Teh YW, Newman D, Welling M (2006) A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Schölkopf PB, Platt JC, Hoffman T, eds. *Adv. Neural Inform. Processing Systems 19, Proc. Twentieth Annual Conf. Neural Inform. Processing Systems, Vancouver, December 4–7* (MIT Press, Cambridge, MA), 1353–1360.

Von Neumann J, Morgenstern O (2007) *Theory of Games and Economic Behavior*, 2nd ed. (Princeton University Press, Princeton, NJ). [Reprint 1947.]

Zhao T, Li C, Li M, Wang S, Ding Q, Li L (2012) Predicting best responder in community question answering using topic model method. *Proc. 2012 IEEE/WIC/ACM Internat. Joint Conf. Web Intelligence Intelligent Agent Tech.*, Vol. 1 (IEEE, Macau, China), 457–461.

Theodore T. Allen is associate professor of integrated systems engineering at Ohio State University. He received his Ph.D. in industrial operations engineering from the University of Michigan (Ann Arbor). He is the author of over 50 peer-reviewed publications, including two textbooks. He is the president elect of the INFORMS social media analytics section and past treasurer of the public sector OR section. Also, he is a 2016 recipient of an INFORMS volunteer service prize and the director of the Security and Efficiency Analytics Laboratory at Ohio State University and the simulation area editor of *Computers and Industrial Engineering* journal (five-year impact factor 2.6). He is the winner of numerous teaching awards, including the College of Engineering Charles E. MacQuigg Student Award for Outstanding Teaching.

Zhenhuan Sui received his Ph.D. in integrated systems engineering from the Ohio State University. Zhenhuan was the runner-up for the 2015 Informs Social Media Analytics Best Student Paper Award. He was an intern at Dow Corning in Toledo and Goldman Sachs in Japan. His interests relate to hierarchical Bayesian models with applications in text modeling and financial derivative pricing.

Nathan L. Parker is an operations research analyst in the U.S. Army at the TRADOC Analysis Center. He received his M.S. from the Naval Postgraduate School and achieved the rank of major in the U.S. Army. His interests include the intersection of statistics and operations research with an emphasis on *R* programming. Major Parker has led projects including developing text modeling software called Subject Matter Expert Refined Topic (SMERT) models involving Dr. Allen, operations research methods to improve reserve recruiting operations, and refining the economic potential of biofuel production with CCS using spatially explicit modeling.