Detecting and Characterizing Events

Allison J.B. Chaney
Princeton University
achaney@cs.princeton.edu

Hanna Wallach
Microsoft Research
wallach@microsoft.com

David M. Blei
Columbia University
david.blei@columbia.edu

Matthew Connelly
Columbia University
mjc96@columbia.edu

Abstract

Significant events are characterized by interactions between entities, such as, countries, organizations, or individuals, that deviate from typical interaction patterns. Analysts, including historians, political scientists, and journalists, commonly read large quantities of text to construct an accurate picture of when and where an event happened, who was involved, and in what ways. In this paper, we present the *Capsule* model for analyzing documents to identify and characterize events of potential significance. Specifically, we develop a model based on topic modeling that distinguishes between topics that describe "business as usual" and topics that deviate from these patterns. To demonstrate this model, we analyze a corpus of over two million U.S. State Department cables from the 1970s. We provide an open-source implementation of an inference algorithm for the model and a pipeline for exploring its results.

1 Introduction

Foreign embassies of the United States government communicate with one another and with the U.S. State Department through diplomatic cables. The National Archive collects these cables in a corpus, which traces the (unclassified) diplomatic history of the United States. The corpus contains, for example, over two million cables sent between 1973 and 1978.

Most of these cables describe diplomatic "business as usual," such as arrangements for visiting officials, recovery of lost or stolen passports, or obtaining lists of names for meetings and conferences. For example, the embassies sent 8,635 cables during the week of April 21, 1975. Here is one, selected at random:

> Hoffman, UNESCO Secretariat, requested info from PermDel concerning an official invitation from the USG RE subject meeting scheduled 10–13 JUNE 1975, Madison, Wisconsin. Would appreciate info RE status of action to be taken in order to inform Secretariat. Hoffman communicating with Dr. John P. Klus RE list of persons to be invited.

But, hidden in the corpus are also cables about important diplomatic events—the cables and events that are most interesting to historians, political sceintists, and journalists. For example, during that same week, the U.S. was in the last moments of the Vietnam war and, on April 30, 1975, lost its hold on Saigon. This triggered the end of the war and a max exodus of refugees. Here is one of the cables about this event:

> GOA program to move Vietnamese Refugees to Australia is making little progress and probably will not cover more than 100-200 persons. Press comment on smallness of program has recognized difficulty of getting Vietnamese out of Saigon, but "Canberra Times" Apr 25 sharply critical of government's performance. [...] Labor government clearly hopes whole matter will somehow disappear.

Our goal in this paper is to develop a tool to help historians, political scientists, and journalists wade through corpora of documents to find potentially significant events and the primary sources around them. We present *Capsule*, a probabilistic model for detecting and characterizing important events, such as the

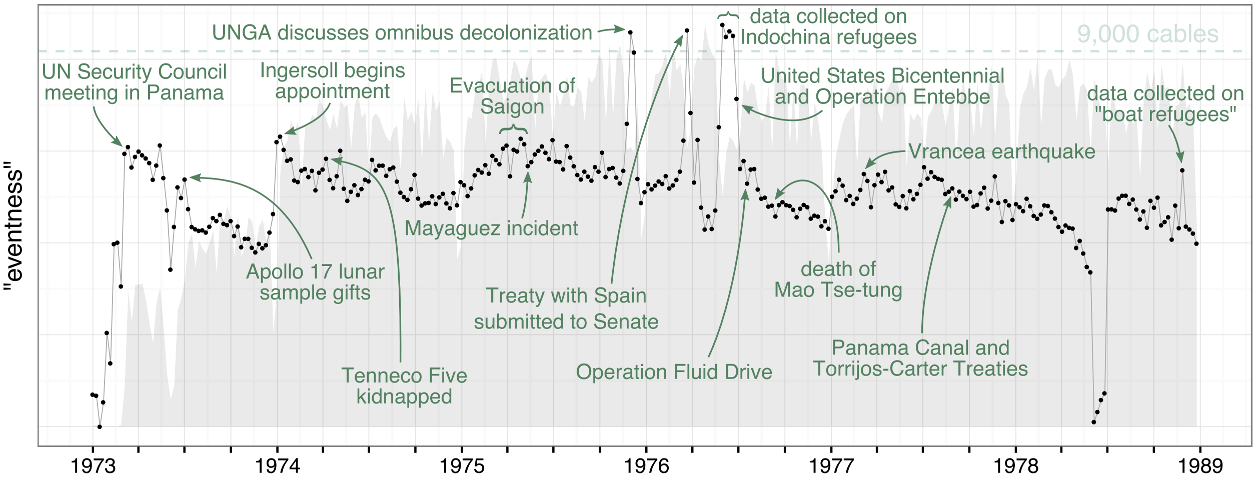


Figure 1: Measure of "eventness," or time interval impact on cable content (Eq. 3.1). Grey background indicates the number of cables sent over time. This comes from the model fit we discuss in Section 4. Capsule successful detects real-world events from National Archive diplomatic cables.

fall of Saigon, in large corpora of historical communication, such as diplomatic cables from the 1970s.

Figure 1 illustrates Capsule's analysis of two million cables from the National Archives' corpus. The y-axis represents "eventness," a loose measure of how strongly a week's cables deviate from typical diplomatic "business as usual" to discuss some matter that is common to many embassies. (We describe this measure of "eventness" in detail in section 3.)

The figure shows that Capsule detects many of the important moments during this five-year span, including the Air France hijacking and Israeli rescue operation "Operation Entebbe" (June 27–July 4, 1976), and the fall of Saigon (April 30, 1975). It also identifies other moments, such as the U.S. sharing lunar rocks with other countries (March 21, 1973) and the death of Mao Tse-tung (Sept. 9, 1976). Broadly speaking, Capsule gives a picture of the diplomatic history of these five years; it identifies and characterizes moments and source material that might be of interest to a historian.

The intuition behind Capsule is this: Embassies write cables throughout the year, usually describing typical diplomatic business, such as visits from government officials. Sometimes, however, important events occur, such as the fall of Saigon, that pull embassies away from their typical activities and lead them to write cables that discuss these events and their consequences. Capsule therefore operationalizes an "event" as a moment in history when multiple embassies deviate from their usual topics of discussion and each embassy deviates in the same way.

Capsule embeds this intuition into a Bayesian model that uses latent variables to encode what "business as usual" means for each embassy, to characterize the events of each week, and to identify the cables that discuss those events. Given a corpus of cables, the corresponding posterior distribution of the latent variables provides a filter for the cables that isolates important moments in diplomatic history. Figure 1 depicts the mean of this posterior distribution.

We present the Capsule model in section 3, providing both a formal model specification and guidance on how to use the model to detect and characterize real-world events. In section **??**, we validate Capsule using simulated data, and in section 4, we use it to analyze over two million U.S. State Department cables. Although we describe Capsule in the context of diplomatic cables, it is suitable for exploring any corpus with the same underlying structure: text (or other discrete multivariate data) generated over time by known entities. This includes email, consumer behavior, social media posts, and opinion articles.

2 Related Work

We first review previous work on automatic event detection and other related concepts, to contextualize our approach in general and Capsule in particular.

In both univariate and multivariate settings, analysts often want to predict whether or not rare events will occur (Weiss and Hirsh, 1998; Das et al., 2008). In contrast, Capsule is intended to help analysts explore and understand their data; our goal is human interpretability rather than prediction or forecasting.

Events can be construed as either anomalies—temporary deviations from usual behavior—or "changepoints" that mark persistent shifts in usual behavior (Guralnik and Srivastava, 1999; Adams and MacKay, 2007). We focus on events as anomalies.

Event detection in the context of news articles (Zhao et al., 2012; Zhao et al., 2007; Zhang et al., 2002; Li et al., 2005; Wang et al., 2007; Allan et al., 1998) and social media posts (VanDam, 2012; Lau et al., 2012; Jackoway et al., 2011; Sakaki et al., 2010; Reuter and Cimiano, 2012; Becker et al., 2010; Sayyadi et al., 2009) usually means identifying clusters of documents. For news, the goal is to create new clusters as novel stories appear, without distinguishing between typical content and rare events; for social media, the goal is to identify rare events, but the resultant methods are intended for short documents, and are not appropriate for longer documents that contain information about a variety of subjects.

Many existing methods for detecting events from text focus on individual vocabulary terms, often weighted by tf-idf values (Fung et al., 2005; Kumaran and Allan, 2004; Brants et al., 2003; Das Sarma et al., 2011; Zhao et al., 2007; Zhao et al., 2012). We characterize events by bursts in groups of terms.

Although groups of terms can be summarized directly (Peng et al., 2007; Chakrabarti and Punera, 2011; Gao et al., 2012), topic models (Blei, 2012) provide a way to automatically identify groups of related terms and reduce the dimensionality of text data. Researchers have previously used topic models to detect events mentioned in social media posts (Lau et al., 2012; Dou et al., 2012) and to find posts relevant to particular, monitored events (VanDam, 2012). Capsule uses topics to characterize both typical diplomatic content and potentially significant events.

In addition to modeling text over time, researchers have also used spatial information (Neill et al., 2005; Mathioudakis et al., 2010; Liu et al., 2011) and information about authors (Zhao et al., 2007) and news outlets (Wang et al., 2007) to enhance event detection. We rely on author information to characterize diplomatic "business as usual" for each embassy.

Event detection is closely related to detecting and characterizing relationships between entities (Schein et al., 2015; Linderman and Adams, 2014; Das Sarma et al., 2011). Capsule can trivially use sender–receiver pairs instead of authors, and the model specification can be tailored to reflect network structure.

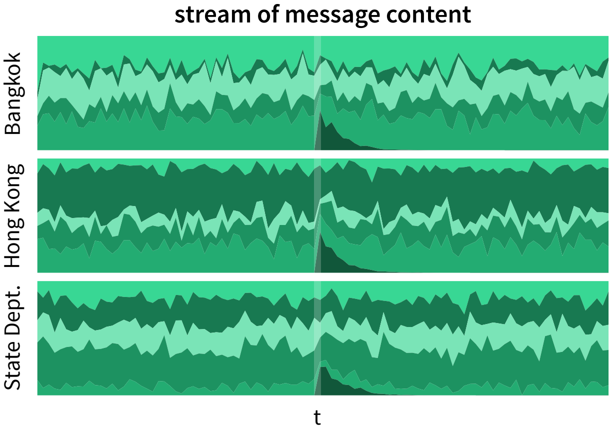


Figure 2: Cartoon intuition. The y-axis represents the stacked proportions of cables about various topics, while the x-axis represents time. The Bangkok embassy, Honk Kong embassy, and U.S. State Department all have typical diplomatic business, about which they usually send cables. When an event occurs during time interval t, the cables alter to cover the event before returning to "business as usual." Capsule discovers the embassies' typical concerns, as well as the timing and content of events.

3 The Capsule Model

In this section, we present the Capsule model for detecting and characterizing significant diplomatic events. We first provide the intuition behind Capsule, and then formally specify the model. We also explain how to use Capsule to explore a corpus and how to learn the posterior distribution of the latent variables.

Consider an entity like the Bangkok embassy, as illustrated in figure 2. We can imagine that this entity sends a stream of diplomatic cables over time—some to the U.S. State Department, others to other American embassies, such as the one in Hong Kong. Embassies usually write cables that describe typical diplomatic business. For example, the Bangkok embassy might write about topics regarding southeast Asia more generally. We can think of a topic as being a probability distribution over vocabulary terms.

Now imagine that an event, such as the capture of Saigon during the Vietnam war, occurs during a particular time interval t. We cannot directly observe the occurrence of this event, but we can observe the stream of cables and the event's impact on it. When the event occurs, multiple entities deviate from their usual topics of discussion simultaneously, before returning to their usual behavior, as depicted in figure 2. For example, the day after the capture of Saigon, the majority of the diplomatic cables written by the Bangkok embassy and several other entities were about Vietnam war refugees. If we think of the event as another probability distribution over vocabulary terms, then each entity's stream of cables reflects its typical concerns, as well as any significant events.

3.1 Model Specification

We now define the Capsule model. Our data come from *entities* (e.g., embassies) who send *messages* (e.g., diplomatic cables) over *time*; specifically, we observe the number of times n_{dv} that each vocabulary term v occurs in each message d. Each message is associated with an author entity a_d and a time interval t_d within which that message was sent.

We model each message with a bank of Poisson distributions—one for each vocabulary term:

$$n_{dv} \sim \text{Poisson}\left(\lambda_{dv}\right). \tag{1}$$

The rate λ_{dv} blends the different influences on message content. Specifically, it blends three types of *topics*, intended to capture "business-as-usual" discussion and content related to significant events.

We operationalize each topic as a specialized probability distribution over vocabulary terms (the set of unique words in the corpus of messages), as is common in topic models (Blei et al., 2003; Canny, 2004; Gopalan et al., 2014)—i.e., each term is associated with each topic, but with a different probability.

Each message blends 1) general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ about diplomacy (e.g., terms about diplomats, terms about communication), 2) an entity topic $\boldsymbol{\eta}_{a_d}$ specific to the author of that message (e.g., terms about Asia),[1] and 3) event topics $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T$ that are specific to the events in recent time intervals (e.g., terms about a coup, terms about the death of a dignitary).

Examples of these three types of topics are in table 1. The general topic relates to planning travel, the entity topic captures words related to the U.S.S.R., and the event topic captures words related to the evacuation of Saigon toward the end of the Vietnam War.

Topic Type	Top Terms
General	visit, hotel, schedule, arrival
Entity	soviet, moscow, ussr, agreement
Event	saigon, evacuation, vietnam, help

Table 1: The four highest-probability vocabulary terms for examples of each of the three types of topics. These example topics come from the analysis that we describe in section 4.

The messages share the three types of topics in different ways: all messages share the general topics, messages written by a single entity share an entity topic, and messages in the same time interval use the event topics in similar ways. Each message blends its corresponding topics with a set of message-specific strengths. As a result, each message captures a different mix of general diplomacy discussion, entity-specific terms, and recent events. Specifically, the Poisson rate for vocabulary term v in message d is

$$\lambda_{dv} = \sum_{k=1}^{K} \theta_{dk} \beta_{kv} + \zeta_d \eta_{a_d v} + \sum_{t=1}^{T} f(t_d, t)\, \epsilon_{dt} \gamma_{tv}, \tag{2}$$

where θ_{dk} is message d's strength for general topic k, ζ_d is message d's strength for a_d's entity topic, and ϵ_{dt} is message d's strength for event topic t. Function $f(\cdot)$ ensures that the events influences decay over time. We find that an exponential decay function, as in equation 6, works well in practice.

We place gamma priors on the topic strengths and Dirichlet priors on the topics. The distributions of general and entity topic strengths are defined hierarchically by entity, capturing the different topics that each entity tends to discuss. The prior on the entity strength is also defined hierarchically; different weeks are more or less "eventful." The graphical model is shown in Figure 3 and the generative process is in Figure 4.

Given a collection of messages, posterior inference uncovers the different types of topics and how each message exhibits them. We will see below, how inferences about the event strengths enable us to filter the corpus to find important messages.

[1] The entity-specific topics play a similar role to the background topics first introduced by Paul and Dredze (2012).

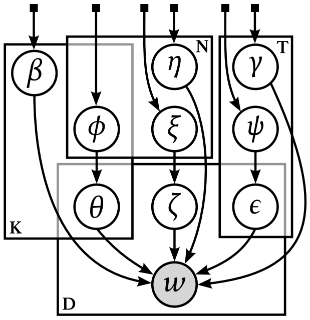


Figure 3: The graphical model for Capsule. Observed words w depend on general topics β, entity-specific topics η, and event topics γ, as well as document representations θ, ξ, and ϵ. Variables ϕ and ζ represent entity concerns (with general topics and entity-specific topics, respectively) and ψ represents the event strength of a given time interval. Hyper-parameters are indicated by black squares, but not labeled for visual simplicity.

- for each time step $t = 1{:}T$,
  - draw interval description over vocabulary (event topic) $\gamma_t \sim \text{Dirichlet}_V(\alpha)$
  - draw interval strength $\psi_t \sim \text{Gamma}(s_\psi, r_\psi)$
- for each entity $n = 1{:}N$,
  - draw entity-specific topics over vocabulary $\eta_n \sim \text{Dirichlet}_V(\alpha)$
  - draw entity-specific topic strength $\xi_n \sim \text{Gamma}(s_\xi, r_\xi)$
- for each topic $k = 1{:}K$,
  - draw general topic distribution over vocabulary $\beta_k \sim \text{Dirichlet}_V(\alpha)$
  - for each entity $n = 1{:}N$,
    - draw general entity concern $\phi_{n,k} \sim \text{Gamma}(s_\phi, r_\phi)$
- for each document $d = 1{:}D$ sent at time i_d by author a_d,
  - draw local entity concern $\zeta_d \sim \text{Gamma}(s_\zeta, \xi_{a_d})$
  - for each topic $k = 1{:}K$,
    - draw local entity concern $\theta_{d,k} \sim \text{Gamma}(s_\theta, \phi_{a_d,k})$
  - for each time $t = 1{:}T$,
    - draw local interval relevancy $\epsilon_{d,t} \sim \text{Gamma}(s_\epsilon, \psi_t)$
  - for each vocabulary term $v = 1{:}V$,
    - set $\lambda_{d,v} = \theta_d^\top \beta_v + \zeta_d \eta_{a_d} + \sum_{t=1}^{T} f(i_d, t)\epsilon_{d,t}\gamma_{t,v}$
    - draw word counts $w_{d,v} \sim \text{Poisson}\left(\lambda_{d,v}\right)$

Figure 4: The generative process for Capsule.

There are connections between Capsule and recent work on Poisson processes. In particular, we can interpret Capsule as a collection of related discrete time Poisson processes with random intensity measures. Further, marginalizing out the event strength prior reveals that word use from one entity can "excite" word use in another, which suggests a close relationship to Hawkes processes (Hawkes, 1971).

Detecting and characterizing events. Once we estimate the posterior distribution of the Capsule parameters, described in the following section, we can use the expectations of the latent parameters to explore the original data. To detect events, we consider the proportion of the document about event j, and take a weighted average of these proportions:

$$m_j = \frac{1}{\sum_d f(i_d, j)} \sum_d \frac{\varepsilon_{d,j}}{\zeta_d + \sum_t \varepsilon_{d,t} + \sum_k \mathbb{E}[\theta_{d,k}]},$$

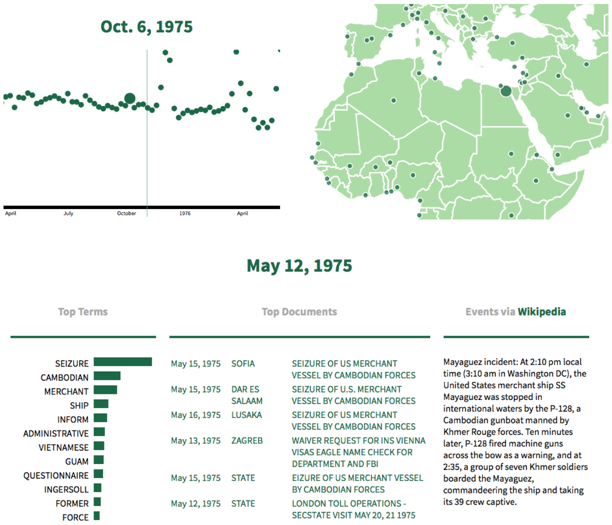
where $\varepsilon_{d,t} = f(i_d, t)\mathbb{E}[\epsilon_{d,t}]$. This measure of "eventness" provides an estimate of the proportion of words that are related to a real-world event in that interval. Figure 1 shows events detected with this metric.

Given an identified event, we can characterize it in terms of its top terms under γ, but we can also use weighted event relevancy parameters $\varepsilon_{d,t}$ to sort documents; Section 4 explores relevant documents for events found in the National Archive diplomatic cables data. In addition to detecting and characterizing events, Capsule can be used to explore entity concerns and the general themes in a given collection.

To make Capsule more accessible, we developed an open source tool for visualizing its results.[2] Our tool creates a navigator of the documents and latent parameters, allowing users to explore events, entities, topics, and the original documents. Figure 5 shows several screenshots of this browsing interface.

Learning the hidden variables. In order to use the Capsule model to explore the observed documents, we must compute the posterior distribution.

[2] Source code: https://github.com/ajbc/capsule-viz; demo: http://www.princeton.edu/~achaney/capsule/.



May 12, 1975

Top Terms

- SEIZURE
- CAMBODIAN
- MERCHANT
- SHIP
- INFORM
- ADMINISTRATIVE
- VIETNAMESE
- GUAM
- QUESTIONNAIRE
- INGERSOLL
- FORMER
- FORCE

Top Documents

May 15, 1975	SOFIA	SEIZURE OF US MERCHANT VESSEL BY CAMBODIAN FORCES
May 15, 1975	DAR ES SALAAM	SEIZURE OF U.S. MERCHANT VESSEL BY CAMBODIAN FORCES
May 16, 1975	LUSAKA	SEIZURE OF US MERCHANT VESSEL BY CAMBODIAN FORCES
May 13, 1975	ZAGREB	WAIVER REQUEST FOR INS VIENNA VISAS EAGLE NAME CHECK FOR DEPARTMENT AND FBI
May 15, 1975	STATE	EIZURE OF US MERCHANT VESSEL BY CAMBODIAN FORCES
May 12, 1975	STATE	LONDON TOLL OPERATIONS - SECSTATE VISIT MAY 20, 21 1975

Events via Wikipedia

Mayaguez incident: At 2:10 pm local time (3:10 am in Washington DC), the United States merchant ship SS Mayaguez was stopped in international waters by the P-128, a Cambodian gunboat manned by Khmer Rouge forces. Ten minutes later, P-128 fired machine guns across the bow as a warning, and at 2:35, a group of seven Khmer soldiers boarded the Mayaguez, commandeering the ship and taking its 39 crew captive.

Figure 5: Screenshots of Capsule visualization of US State Department cables. Top-left: events over time (similar to Figure 1). Right-top: entities shown on a map. Bottom: time interval summary, including top terms, relevant documents, and text scraped from Wikipedia.

Conditional on the observed word counts w, our goal is to compute the posterior values of the hidden parameters—general topics β, entity topics η, event topics γ, entity concerns ϕ (for general topics) and ξ (for their own topic), overall event strengths ψ, and document-specific strengths for general topics θ, entity topics ζ, and event topics ϵ.

As for many Bayesian models, the exact posterior for Capsule is not tractable to compute; approximating it is our central statistical and computational problem. We develop an approximate inference algorithm for Capsule based on variational methods (Jordan et al., 1999),[3] which is detailed in **??**.[4] This algorithm produces a fitted variational distribution which can then be used as a proxy for the true posterior, allowing us to explore a collection of documents with Capsule.

4 Evaluation

In this section we explore the performance of Capsule on simulated data and a collection of over 2 million U.S. State Department diplomatic cables from the 1970s.

4.1 Results on Simulated Data

Prior to exploring Capsule results on data of historical interest, we provide a quantitative assessment of the model on simulated data.

We generated ten data sets, each with 100 time steps, 10 general topics, and 100 entities. Each simulation contained about 20,000 documents and followed the generative process assumed by Capsule, as shown in Figure 4.

To evaluate event detection, we created a ranked list of all time intervals and computed the overlap between a method and the simulated ground at every threshold; this generates an curve under which we can compute the area and normalized based on ideal performance—we refer to this metric as event detection AUC:

$$\frac{\sum_{i=1}^{T} |\mathrm{Truth}_i \cap \mathrm{Model}_i|}{\sum_{i=1}^{T} i}, \tag{3}$$

where Model_i is a set of the top i most eventful intervals according to the model, and Truth_i is the known set of the top i most eventful intervals. As the data is simulated, we can order all intervals by their known "eventness"—this metric captures how well the model recovers the true ordering.

The most successful of the baseline methods for event detection were based on absolute error in word count relative to the mean. This can be computed for all words:

$$\sum_{v=1}^{V} \left[\sum_{d=1}^{D} \mathrm{abs} \left(w_{d,v} - \frac{1}{|D|} \sum_{d=1}^{D} w_{d,v} \right) \right], \tag{4}$$

and can also be weighted by tf-idf,

$$\sum_{v=1}^{V} \text{tf-idf}(v) \left[\sum_{d=1}^{D} \mathrm{abs} \left(w_{d,v} - \frac{1}{|D|} \sum_{d=1}^{D} w_{d,v} \right) \right]. \tag{5}$$

We also considered metrics that computed deviations on the entity and document level, but the simplest overall metrics performed best.

Figure 6 shows that Capsule[5] outperforms these approaches for event detection. We also consider

[3] Source code is available at `https://github.com/ajbc/capsule`.

[4] Appendices are located in the supplemental materials document.

[5] The model was set with the same number of topics $K = 10$ and exponential decay f used to simulate the data. More details on the decay function surround its formal definition in Equation (6).

an "event only" model—this is a model that only uses the interval-related subset of Capsule's parameters; comparing to this shows that is it important to model "business as usual" for improved event detection. LDA based approaches like average deviation from mean in topic space ((Dou et al., 2012)) do not perform well for event detection as deviations in topic space are too coarse to provide a meaningful signal.

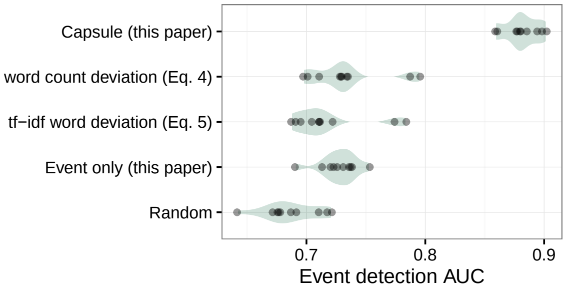


Figure 6: Event detection performance on ten simulated datasets; each dot is the performance on a single dataset, and the shaded green describes the distribution of the performances. Capsule detects events better than comparison methods.

Once events have been identified, our next task is to identify relevant documents; to evaluate this, we calculate precision of recovering the top N documents. Both Capsule and its event-only partial model outperform all comparison methods in terms of document recovery. For Capsule, average precision at 10 documents was 0.44; the event-only model had average precision of 0.09. LDA performed slightly worse than the event-only model, and the other comparison methods (similar to Equations 4 and 5) recovered zero relevant documents–equivalent to random.

As described in **??**, we assessed the sensitivity of our model to different settings of event duration τ three different decay functions f: exponential, linear, and step functions. We found that fitting Capsule with an exponential decay function, or

$$f(i_d, t) = \begin{cases} 0, & \text{if } t \leq i_d < t + \tau \\ \exp\left\{\frac{-(i_d - t)}{\tau/5}\right\}, & \text{otherwise,} \end{cases} \tag{6}$$

provided the best performance and the most interpretable results.

4.2 Results on U.S. State Department Diplomatic Cables

As Capsule is intended to be used to explore large collections of documents, we must demonstrate its use in that context. This sections describes and explores the application of Capsule to a real-world collection of diplomatic messages.

Data. The National Archive collects communications between the U.S. Sate Department and its embassies. We obtained a collection of these diplomatic messages from the History Lab at Columbia,[6] which received them from the Central Foreign Policy Files at the National Archives. The communications in this data set were sent between 1973 and 1978.

In addition to the text of the cables themselves, each document is supplemented with information about who sent the cable (e.g., the State Department, the U.S. Embassy in Saigon, or an individual by name), who received the cable (often multiple entities), and the date the cable was sent. We used a vocabulary of size 6,293 and omitted cables with fewer than three terms, resulting in a collection of 2,021,852 messages sent between 22,961 entities. We selected a weekly duration for the time intervals, as few cables were sent on the weekends.

Model Settings. We fit Capsule with $K = 100$ general topics and using the exponential decay f, shown in Equation (6), with event duration $\tau = 4$. With these settings on the cables data, fitting the model takes about one hour per iteration.[7]

Quantitative Results. The History Lab at Columbia provided a list of 39 real-world events in the time period covered by the cables data; they validated that these events were present in at least one of six reputable collections of events, such as the Office of the Historian list of milestones.[8]

We ran Capsule and baseline comparison methods to recover these events, and used the nDCG metric to evaluate the methods. The nDCG metric is dis-

[6] http://history-lab.org

[7] Our algorithm is batch–we consider each data point for every iteration. Modifying the algorithm to stochastically sample the data would reduce the time required to achieve an equivalent model fit.

[8] https://history.state.gov/milestones/1969-1976

counted cumulative gain,

$$\mathrm{DCG} = \sum_{j=1}^{T} \frac{\mathbf{1}[\text{interval at rank } j \text{ in known events}]}{\log j}, \tag{7}$$

divided by the ideal DCG value, or

$$\mathrm{nDCG} = \frac{\mathrm{DCG}}{\text{ideal DCG}}. \tag{8}$$

As shown in Table 2, Capsule outperforms the baselines.

Additionally, we can compute held-out validation data likelihood on the model and each of its component parts; Table 3 shows that the full Capsule model captures the data better than any of its component parts individually.

Model Exploration. The evaluations to this point are useful in validating that Capsule captures its intended constructs, but the objective of the model is not prediction; rather, it is to be used as a scaffold to explore large collections of documents. We now turn to exploring the cables data using Capsule.

We begin our exploration by detecting events using Capsule. With Section 3.1 as our metric of "eventness," we consider this metric over time, which is shown in Figure 1. Here, high values—often peaks—correspond to real-worlds events, several of which are labeled.

One of the tallest peak occurs the week of December 1, 1975, during which the United Nations General Assembly (UNGA) discussed omnibus decolonization. As discussed in Section 3, we sort documents by their weighted event relevancy parameters $f(i_d, t)\epsilon_{d,t}$ to find cables that reflect an event. Table 4 shows the top cables for this discussion. Capsule accurately identifies this real-world event and recovers relevant cables.

Another notable event was the seizure of the S.S. Mayaguez, an American merchant vessel, in May of 1975—at the end of the Vietnam War. The top documents for this week are shown in Table 5. We can inspect individual documents to confirm their relevancy and learn more about the events. For instance, the content of the most relevant document, according to Capsule, is as follows.

> In absence of MFA Chief of Eighth Department Avramov, I informed American desk officer Yankov of circumstances surrounding seizure and recovery of merchant ship Mayaguez and its crew. Yankov promised to inform the Foreign Minister of US statement today (May 15). Batjer

A third week of interest occurs in early July of 1976. On July 4th, the US celebrated its Bicentennial, but on the same day, Israeli forces completed a hostage rescue mission—an Air France flight from Tel Aviv had been hijacked and taken to Entebbe, Uganda. This event, like many events, is mostly discussed the week following the real-world event; relevant cables are shown in **??**, **??**. The cable from Stockholm describing the "Ugandan role in Air France hijacking" begins with the following content, which reveals further information about the event.

> 1. We provided MFA Director of Political Affairs Leifland with Evidence of Ugandan assistance to hijackers contained in Ref A. After reading material, Leifland described it a "quite good", and said it would be helpful for meeting MFA has scheduled for early this morning to determine position GOS will take at July 8 UNSC consideration of Israeli Rescue Operation. ...

Capsule assumes that only one event occurs in each time interval—this example is a clear violation of this assumption, but it also demonstrates that the model successfully captures both events, even when they overlap.

In addition to events, Capsule can be used to explore the general themes of a corpus and entities' typical concerns. Examples of general topics of conversation are shown in **??**, **??** and entity-exclusive topics are shown in **??**, **??**; these show us how entity topics absorb location-specific words, preventing these terms from overwhelming the general topics.

These exploratory results show that our model is successfully capturing when multiple entities are discussing the same subjects and that our model can be used to explore the underlying data by providing a structured scaffold from which to view the data.

Method	**nDCG**
Capsule	0.693
Average tf-idf weighted word count deviation	0.652
Average unweighted word count deviation	0.642
Single term maximum tf-idf weighted deviation	0.561
Random (10k ave)	0.557
Single term maximum unweighted deviation	0.555

Table 2: Evaluation of Capsule and comparison baselines on a collection of 39 real-world events. Capsule performs best.

Model	**LL at 10 iterations**	**LL at convergence**
Full Capsule	-1.62e7	-1.52e7
Entity Topics Only	-1.64e7	–
General Topics Only	-1.71e7	-1.53e7
Event Only	-1.79e7	–

Table 3: Log likelihood (LL) computed on validation data at 10 iterations and at convergence—the event only and entity only models are small enough that they converge with very few iterations. The full Capsule model achieves the lowest log likelihood in both cases.

5 Conclusion

We have presented Capsule, a Bayesian model that identifies when events occur, characterizes these events, and discovers the typical concerns of author entities. We have shown that Capsule outperforms comparison methods and explored its results on a real-world datasets. We anticipate that Capsule can be used by historians, political scientists, and others who wish to investigate events in large text corpora.

References

Ryan Prescott Adams and David JC MacKay. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 291–300.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, January.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 330–337.

John Canny. 2004. Gap: a factor model for discrete data. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 11:66–73.

Kaustav Das, Jeff Schneider, and Daniel B Neill. 2008. Anomaly pattern detection in categorical datasets. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 169–176.

Anish Das Sarma, Alpa Jain, and Cong Yu. 2011. Dynamic relationship and event discovery. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 207–216.

Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X Zhou. 2012. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102. IEEE.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 181–192. VLDB Endowment.

$f * \epsilon$	Date	Entity	Subject
4.60	1975-12-05	Canberra	30th UNGA: Item 23, Guam, Obmibus Decolonization and ...
4.26	1975-12-05	Mexico	30th UNGA-Item 23: Guam, Omnibus Decolonization and ...
4.21	1975-12-06	State	30th UNGA-Item 23: Guam, Omnibus Decolonization and ...
4.11	1975-12-03	Dakar	30th UNGA: Resolutions on American Samoa, Guam and ...
4.08	1975-12-04	Monrovia	30th UNGA: Item 23: Resolutions on decolonization and A...

Table 4: Top documents for the time interval of week December 1, 1975, when the UN discussed decolonization resolutions; Capsule recovers relevant documents related to this real-world event. Typos intentionally copied from original data.

$f * \epsilon$	Date	Entity	Subject
5.06	1975-05-15	Sofia	Seizure of US merchant vessel by Cambodian forces
5.05	1975-05-15	Dar es Salaam	Seizure of U.S. merchant vessel by Cambodian forces
4.92	1975-05-16	Lusaka	Seizure of US merchant vessel by Cambodian forces
4.61	1975-05-13	Zagreb	Waiver request for INS Vienna visas Eagle name check...
4.59	1975-05-15	State	eizure of US merchant Vessel by Cambodian forces

Table 5: Top documents for the week during which the S.S. Mayaguez was captured. Capsule identifies documents relevant to the real-world event. Typos intentionally copied from original data.

Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 1173–1182.

Prem K Gopalan, Laurent Charlin, and David Blei. 2014. Content-based recommendations with Poisson factorization. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 3176–3184. Curran Associates, Inc.

Valery Guralnik and Jaideep Srivastava. 1999. Event detection from time series data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 33–42.

Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. 2011. Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32. ACM.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 297–304.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1519–1534.

Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. 2005. A probabilistic model for retrospective news event detection. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 106–113.

Scott W Linderman and Ryan P Adams. 2014. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*.

Xueliang Liu, Raphaël Troncy, and Benoit Huet. 2011. Using social media to identify events. In *Proceedings of the ACM SIGMM International Workshop on Social Media (WSM)*, pages 3–8.

Michael Mathioudakis, Nilesh Bansal, and Nick Koudas. 2010. Identifying, attributing and describing spatial bursts. *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, 3(1-2):1091–1102.

Daniel B Neill, Andrew W Moore, Maheshkumar Sabhnani, and Kenny Daniel. 2005. Detection of emerging space-time clusters. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 218–227.

Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from twitter. *Health*, 11:16–6.

Wei Peng, Charles Perng, Tao Li, and Haixun Wang. 2007. Event summarization for system management. In *Proceedings of the ACM SIGKDD International Confer-*

ence on Knowledge Discovery and Data Mining (KDD), pages 1028–1032.

Timo Reuter and Philipp Cimiano. 2012. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 22. ACM.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 851–860.

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. 2015. Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1045–1054.

Courtland VanDam. 2012. A probabilistic topic modeling approach for event detection in social media. Master's thesis, Michigan State University.

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 784–793. ACM.

Gary M Weiss and Haym Hirsh. 1998. Learning to predict rare events in event sequences. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 359–363.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88.

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 7, pages 1501–1506.

Wayne Xin Zhao, Rishan Chen, Kai Fan, Hongfei Yan, and Xiaoming Li. 2012. A novel burst-based text representation model for scalable event detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 43–47. Association for Computational Linguistics.