Detecting and Characterizing Events

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

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Matthew Connelly
Columbia University
mjc96@columbia.edu

David M. Blei
Columbia University
david.blei@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 detect 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 (declassified) diplomatic history of the United States.[1] 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 scientists, 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 marked the end of the war and induced a mass exodus of refugees. Here is one cable 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

[1] The National Archives' corpus also includes messages sent by diplomatic pouch; however, for brevity, and at the risk of being imprecise, we also refer to these messages as "cables."

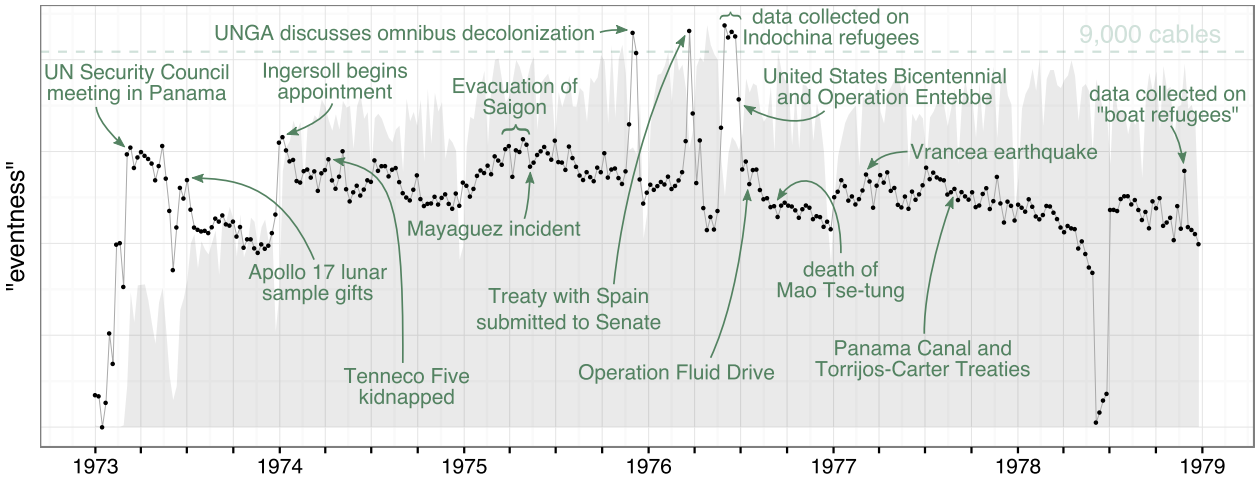

Figure 1: Capsule's analysis (described in detail in section 5) of two million cables from the National Archives' corpus. The y-axis represents a loose measure of "eventness" (equation (5)). The gray background depicts the number of cables sent over time.

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 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.)

This figure shows that Capsule detects many well-known events between 1973 and 1978, including the fall of Saigon (April 30, 1975) and the death of Mao Tse-tung (September 9, 1976). Capsule also uncovers obscure, but significant, events that have largely escaped the attention of scholars, such as when the U.S. defended its control of the Panama Canal before the United Nations Security Council (March 19, 1973). Capsule therefore provides a new way to detect and characterize historical moments that may be of interest to historians, political scientists, and journalists.

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 a similar 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 4, we validate Capsule using simulated data, and in section 5, 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, ana-

lysts 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 (Atefeh and Khreich, 2015; 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; each article is assumed to be associated with one event, which does not allow for distinctions 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 may 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.

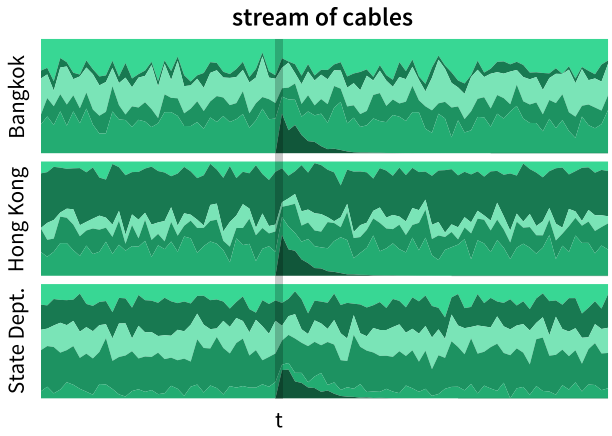

Figure 2: Cartoon intuition. The y-axis represents the stacked proportions of cables about various topics, while the x-axis represents time. The Bangkok embassy, Hong Kong embassy, and U.S. State Department all have typical diplomatic business, about which they usually send cables. When an event occurs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

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.

Finally, 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 strengths (described in section 3.1) reveals that the use of a vocabulary term by one embassy can "excite" the use of that term by another. This suggests a close relationship to Hawkes processes (Hawkes, 1971).

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[2]—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.

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

Table 1: The highest-probability vocabulary terms for examples of the three types of topics (general, entity, and event). These examples come from the analysis that we describe section 5.

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 Hong Kong),[3] 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.

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. The function $f(\cdot)$ ensures that the events influences decay over time. As we describe in appendix B, we

[2] Readers familiar with topic modeling may expect a multinomial model of term occurrences, but Poisson models of counts better capture messages with different lengths (Canny, 2004).

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

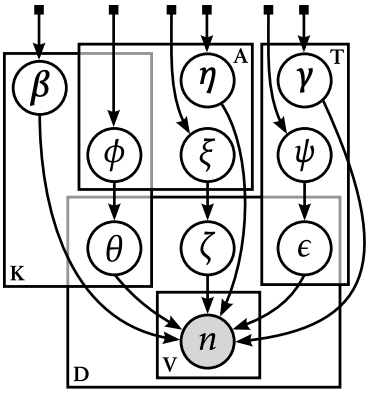

Figure 3: Graphical model for Capsule. Observed term counts depend on general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, entity topics $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_A$, and event topics $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T$, as well as message-specific strengths $\boldsymbol{\theta}_d$, ζ_d, and $\boldsymbol{\epsilon}_d$. Variables $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A$ and $\xi_1, \ldots, \xi_A$ represent entity-specific strengths, while $\psi_1, \ldots, \psi_T$ allow time intervals to be more or less "eventful." Black squares denote hyperparameters (unlabeled for visual simplicity).

compared several different decay functions (exponential, linear, and step) and found that the following exponential decay function works well in practice:

$$f(t_d, t) = \begin{cases} 0 & t \leq t_d < t + \tau \\ \exp\left(\frac{-(t_d - t)}{\tau / 5}\right) & \text{otherwise.} \end{cases} \quad (3)$$

Dividing τ by five means that we can interpret it as the number of time intervals after which an event will have little impact on the content of the messages.

We place hierarchical gamma priors over the message-specific strengths, introducing entity-specific strengths $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A$ and $\xi_1, \ldots, \xi_A$ that allow different entities to focus on different topics and event strengths $\psi_1, \ldots, \psi_T$ that allow different time intervals to be more or less "eventful." We place Dirichlet priors over the topics. The graphical model is in figure 3 and the generative process is in figure 4.

Given a corpus of messages, learning the posterior distribution of the latent variables uncovers the three types of topics, the message- and entity-specific strengths, and the event strengths. In section 3.3, we explain how an analyst can use the event strengths as a filter that isolates potentially significant messages.

3.2 Learning the Posterior Distribution

In order to use Capsule to to explore a corpus of messages, we must first learn the posterior distribution of the latent variables—the general topics, the entity topics, the event topics, the message- and entity-specific strengths, and the event strengths—conditioned on the observed term counts. As for many Bayesian models, this posterior distribution is not tractable to compute; approximating it is therefore our central statistical and computational problem. We introduce an approximate inference algorithm for Capsule, based on variational methods (Jordan et al., 1999),[4], which

- for $k = 1, \ldots, K$,
  - draw general topic $\boldsymbol{\beta}_k \sim \text{Dirichlet}_V(\alpha, \ldots, \alpha)$
  - for each entity $a = 1, \ldots, A$,
    - ► draw entity-specific strength $\phi_{ak} \sim \text{Gamma}(s, r)$
- for each entity $a = 1, \ldots, A$,
  - draw entity topic $\boldsymbol{\eta}_a \sim \text{Dirichlet}_V(\alpha, \ldots, \alpha)$
  - draw entity-specific strength $\xi_a \sim \text{Gamma}(s, r)$
- for each time interval $t = 1, \ldots, T$,
  - draw event topic $\boldsymbol{\gamma}_t \sim \text{Dirichlet}_V(\alpha, \ldots, \alpha)$
  - draw event strength $\psi_t \sim \text{Gamma}(s, r)$
- for each message $d = 1, \ldots, D$, sent during time interval t_d by author entity a_d,
  - for each general topic $k = 1, \ldots, K$,
    - ► draw message-specific strength $\theta_{dk} \sim \text{Gamma}(s, \phi_{a_d k})$
  - draw message-specific strength $\zeta_d \sim \text{Gamma}(s, \xi_{a_d})$
  - for each time interval $t = 1, \ldots, T$,
    - ► draw message-specific strength $\epsilon_{dt} \sim \text{Gamma}(s, \psi_t)$
  - for each vocabulary term $v = 1, \ldots, V$,
    - ► set $\lambda_{dv} = \sum_{k=1}^{K} \theta_{dk}\beta_{kv} + \zeta_d \eta_{a_d v} + \sum_{t=1}^{T} f(t_d, t)\, \epsilon_{dt}\gamma_{tv}$
    - ► draw term counts $n_{d,v} \sim \text{Poisson}(\lambda_{dv})$

Figure 4: Generative process for Capsule. We use s and r to denote top-level (i.e., fixed) shape and rate hyperparameters; they can be set to different values for different variables.

[4] Source code: https://github.com/ajbc/capsule.

we outline in appendix A.[5] This algorithm produces a fitted variational distribution which be can then be used as a proxy for the true posterior distribution.

3.3 Detecting and Characterizing Events

We can use the mean of the fitted variational distribution to explore the data. Specifically, we can explore "business-as-usual" content using the posterior expected values of the general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ and the entity topics $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_A$, and we can detect and characterize events using the posterior expected values of the event strengths and the event topics.

To detect events, we define an measure that quantifies the "eventness" of time interval t. Specifically, we first compute how relevant each message d is to that time interval: $m_{dt} = f(t_d, t)\,\mathbb{E}[\epsilon_{dt}]$. Using these relevancy values, we then compute the proportion of each message's term counts that are associated with the event topic specific to time interval t:

$$p_{dt} = \frac{m_{dt}}{\sum_k \mathbb{E}[\theta_{dk}] + \mathbb{E}[\zeta_d] + \sum_{t'} m_{dt'}}. \quad (4)$$

Finally, we aggregate these values over messages:

$$\frac{1}{\sum_d f(t_d, t)} \sum_{d=1}^{D} p_{dt}, \quad (5)$$

where the multiplicative fraction ensures that messages that were sent during time intervals that are further from t contribute less than than messages that were sent during time intervals that are closer to t.

We can characterize an event t by selecting the highest-probability vocabulary terms from $\mathbb{E}[\boldsymbol{\gamma}_t]$. By ordering the messages according to $m_{dt} = f(t_d, t)\,\mathbb{E}[\epsilon_{dt}]$, we can also identify the messages that are most strongly associated with event t.

In section 5, we explore the cables associated with significant events in the National Archives' corpus of diplomatic cables. To make Capsule more accessible for historians, political scientists, and journalists, we have released an open-source tool for visualizing its results.[6] This tool allows analysts to browse a corpus of messages and the mean of the corresponding posterior distribution, including general topics, entity topics, and event topics. Figure 5 contains several screenshots of the tool's browsing interface.

[5] Appendices are in the supplemental material.

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

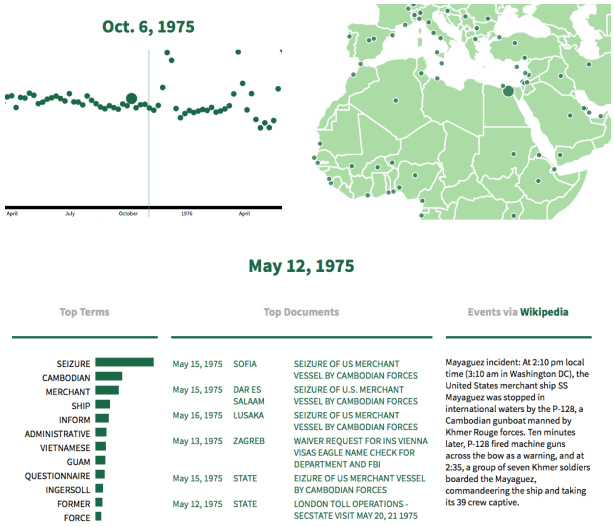

Figure 5: Screenshots of the Capsule visualization tool used to explore U.S. State Department cables. Top left: events over time (similar to figure 1). Top right: entities located on a map. Bottom: summary of the week of May 12, 1975, including top vocabulary terms, relevant cables, and text from Wikipedia.

4 Model Validation with Simulated Data

Before using Capsule to explore a corpus of real messages (described in section 5), we provide a quantitative validation of the model using simulated data.

We used the generative process in figure 4 to create ten data sets, each with 100 time intervals, ten general topics, ten entities, and roughly 20,000 messages. We then used these data sets to compare Capsule's event detection performance to that of four baseline methods. We also compared the methods' abilities to identify the most relevant messages for each event.

4.1 Detecting Events

For each data set, we ordered the time intervals from most to least eventful, using the "eventness" measure described in section 3.3 and the simulated values of the latent variables. We then treated these ranked lists of time intervals as "ground truth" and assessed how well each method was able to recover them.

For Capsule itself, we used our approximate inference algorithm to obtain a fitted variational distribution for each simulated data set. We then ordered the time intervals using our "eventness" measure and the posterior expected values of the latent variables.

For our first baseline, we constructed an "event-only" version of Capsule by dropping the first and

second terms in equation (2). We used this baseline to test whether modeling "business as usual" discussion makes it easier to detect significant events. We obtained a fitted variational distribution for this model using a variant of our approximate inference algorithm, and then ordered the time intervals using our "eventness" measure, modified appropriately, and the posterior expected values of the latent variables.

For our second baseline, we drew inspiration from previous work on event detection in the context of news articles, and focused on each time interval's deviation in term counts from the average. Specifically, we ordered the time intervals $1, \ldots, T$ for each simulated data set according to this measure:

$$\sum_{v=1}^{V} \sum_{\substack{d=1 \\ t_d=t}}^{D} \left| n_{dv} - \frac{1}{D} \sum_{d=1}^{D} n_{dv} \right|. \quad (6)$$

We added tf-idf term weights for our third baseline:

$$\sum_{v=1}^{V} \text{tf-idf}\,(v) \sum_{\substack{d=1 \\ t_d=t}}^{D} \left| n_{dv} - \frac{1}{D} \sum_{d=1}^{D} n_{dv} \right|. \quad (7)$$

Finally, we randomly ordered the time intervals for each data set to serve as a straw-man baseline.

We also experimented with baselines that involved term-count deviations on the entity level and topic-usage deviations on the message level (Dou et al., 2012), but found that they were not competitive.

For each data set, we compared each method's ranked list of time intervals to the corresponding "ground-truth" list of time intervals, by dividing the sum of the lists' actual set overlap at each rank by the sum of their maximum set overlap at each rank:

$$\frac{\sum_{r=1}^{T} |S_r^{\text{truth}} \cap S_r^{\text{method}}|}{\sum_{r=1}^{T} r}, \quad (8)$$

where S_r^{truth} is a set of the top r time intervals according to the "ground-truth" list and S_r^{method} is a set of the top r time intervals according to the method.

Figure 6 shows that Capsule outperforms all four baseline methods. These results serve as a sanity check for both the model and its implementation.

4.2 Identifying Relevant Messages

For each data set, we created a list of the most relevant messages for each time interval t by computing $f(t_d, t)\,\epsilon_{dt}$ for each message d (using the simulated values of ϵ_{dt}) and ordering the messages accordingly. We then treated these ranked lists of messages as "ground truth" and assessed how well Capsule and the baseline methods were able to recover them.

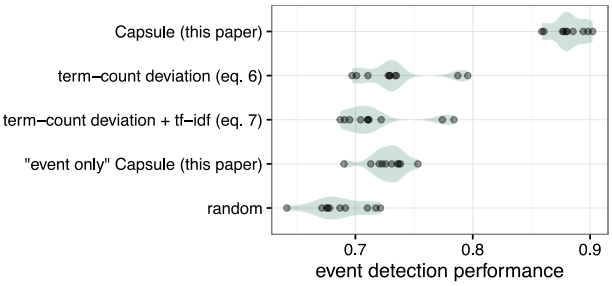

Figure 6: Event detection performance using ten simulated data sets. Each dot represents the performance (equation (8); higher is better) of a single method on a single data set; each shaded green area summarizes the distribution of performance for a single method. Capsule outperforms all four baseline methods.

For Capsule, we used our approximate inference algorithm to obtain a fitted variational distribution for each data set, and then, for each time interval, ordered the messages according to $m_{dt} = f(t_d, t)\,\mathbb{E}[\epsilon_{dt}]$. For our second and third baselines, we ordered the messages sent during each time interval according message-specific versions of equations (6) and (7).

For each data set, we compared each method's ranked list of messages for each time interval to the corresponding "ground-truth" list, by computing precision at ten messages. The average precision for Capsule was was 0.44, while the average precision for the "event-only" version of the model was 0.09. The other baselines recovered zero relevant messages.

5 Exploratory Analysis

Capsule is intended to help analysts explore and understand their data. In this section, we demonstrate its capabilities by analyzing a corpus of over two million U.S. State Department cables from the 1970s.

5.1 Data

The National Archive collects diplomatic cables sent between the U.S. State Department and its foreign embassies. We obtained a subset of this corpus from the Central Foreign Policy Files at the National Archives, via the History Lab at Columbia Univer-

sity;[7] the subset contains cables sent between 1973 and 1978. In addition to the text of the cables, each message is labeled with its author (e.g., the U.S. State Department, a particular embassy, or an individual), the date the cable was sent, and other metadata. We used a vocabulary of 6,293 terms and omitted cables with fewer than three terms, resulting in 2,021,852 cables sent by 22,961 entities. We used weekly time intervals, as few cables were sent on weekends.

5.2 Model Settings

We ran our approximate inference algorithm for Capsule to obtain a fitted variational distribution. We used $K = 100$ general topics, the exponential decay function in equation (3) with $\tau = 4$, and top-level hyperparameters $s = r = 0.3$. With these settings, a single iteration of the algorithm took about an hour.[8]

5.3 Detecting Well-Known Events

To evaluate Capsule's ability to detect well-known events, we used a list, provided to us by the History Lab, of thirty-nine well-known events that took place between 1973 and 1978. Each event is present in at least one of six reputable collections of historic events, such as the Office of the Historian's Milestones in the History of U.S. Foreign Relations.[9] We treated this list of events as "ground truth" and assessed how well Capsule and each of the baselines described in section 4.1 were able to recover them—or, in other words, how well the methods identify these eventful weeks, compared to more typical weeks.

Specifically, we used each method to construct a ranked list of time intervals. Then, for each method, we computed the discounted cumulative gain (DCG), which, in this context, is equivalent to computing

$$\sum_{e=1}^{39} \frac{1}{\log\left(\text{rank}\left(e, L_T^{\text{method}}\right)\right)}, \tag{9}$$

where L_T^{method} is the method's ranked list of time intervals and $\text{rank}\left(e, L_T^{\text{method}}\right)$ is the rank of the e^{th} well-known event in L_T^{method}. Finally, we divided the DCG by the ideal DCG—i.e., $\sum_{e=1}^{39} \frac{1}{\log(e)}$—to obtain the normalized DCG (nDCG). Table 2 shows that Capsule outperforms all four baseline methods.

Method	nDCG
Capsule (this paper)	0.693
term-count deviation + tf-idf (equation (7))	0.652
term-count deviation (equation (6))	0.642
random	0.557
"event-only" Capsule (this paper)	0.426

Table 2: Event detection performance (nDCG; higher is better) using thirty-nine well-known events that took place between 1973 and 1978. Capsule outperforms all four baseline methods.

5.4 Exploration

We now turn to our primary goal—using Capsule to explore and understand a corpus of messages. Figure 1 shows our "eventness" measure (equation (5)) over time. One of the tallest peaks occurs during the week of December 1, 1975, when the United Nations General Assembly discussed omnibus decolonization. As described in section 3.3, we can characterize this event by computing $m_{dt} = f(t_d, t)\,\mathbb{E}[\epsilon_{dt}]$ for each message d and then ordering the messages accordingly. Table 3 lists the highest-ranked messages.

Another notable event was the seizure of the S.S. Mayaguez, an American merchant vessel, during May, 1975, at the end of the Vietnam War. Table 4 lists the highest-ranked messages for this event. We can examine these messages to confirm their relevancy and learn more about the event. For example, here is the content of the most relevant message:

> 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, 1976. On July 4, the U.S. celebrated its Bicentennial, but on the same day, Israeli forces completed a hostage rescue mission because an Air France flight from Tel Aviv had been hijacked and taken to Entebbe, Uganda.[10] This event was mostly discussed the week

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

[8] Each iteration of our algorithm considers all messages. Modifying it to stochastically sample the data would reduce the time required to obtain an equivalent fitted variational distribution.

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

[10] Capsule assumes that only one event occurs during each

$f(t_d, t)\,\mathbb{E}[\epsilon_{dt}]$	Date	Author 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 3: Highest-ranked messages for the week of December 1, 1975, when the United Nations General Assembly discussed decolonization. Capsule accurately recovers messages related to this real-world event. Typos are intentionally copied from the data.

$f(t_d, t)\,\mathbb{E}[\epsilon_{dt}]$	Date	Author 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 4: Highest-ranked messages for the week of May 12, 1975, when the S.S. Mayaguez, an American merchant vessel, was captured. Capsule accurately recovers messages related to this real-world event. Typos are intentionally copied from the data.

after the event took place; the most relevant messages are listed in appendix B (table 5). The cable from Stockholm describing the "Ugandan role in Air France hijacking" begins with the following content, which reveals further information about this 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. ...

In addition to detecting and characterizing well-known events, such the S.S. Mayaguez incident and Operation Entebbe, Capsule can detect and characterize obscure, but significant, events, such as when Eritrean rebels kidnapped Tenneco oil employees (April 8, 1974) and when the U.S. Navy evacuated citizens from Lebanon ("Operation Fluid Drive," June 20, 1976). Both events appear in figure 1. Capsule uncovers events where analysts might not otherwise look.

Capsule also provides a way to explore "business-as-usual" discussion using the posterior expected values of the general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ and the entity topics $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_A$. Examples of each of these types of topics are in appendix B (tables 6 and 7, respectively); these examples illustrate that, as desired, the entity topics absorb location-specific terms, preventing them from overwhelming the general topics.

6 Conclusion

We presented Capsule, a Bayesian model for detecting and characterizing potentially significant events. We evaluated Capsule's ability to detect events and identify relevant messages; it outperformed four baseline methods. We used Capsule to analyze a large corpus of U.S. State Department cables from the 1970s, demonstrating that it can discover both well-known and obscure (but significant) events, as well as relevant documents. We anticipate that Capsule, and our visualization tool, will be useful for historians, political scientists, and journalists who wish to explore and understand large corpora of documents. This is increasingly important—the U.S. State Department alone produces around two billion e-mails annually.

Acknowledgments

This work was supported by NSF IIS-1247664; ONR N00014-11-1-0651; DARPA FA8750-14-2-0009 and N66001-15-C-4032; Adobe; the Alfred P. Sloan Foundation; the Columbia Global Policy Initiative.

time interval. This example is a clear violation of this assumption, but also serves to demonstrate that Capsule can successfully detect and characterize multiple events, even when they overlap.

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.

Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

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.

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 Conference 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.

Detecting and Characterizing Events: Appendices

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

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Matthew Connelly
Columbia University
mjc96@columbia.edu

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

A Inference

In this appendix, we describe the details of the approximate inference algorithm for Capsule.

Conditioned on the observed term counts—n_{dv} for vocabulary term v in message d; collectively $\mathbf{N}$—our goal is to learn the posterior distribution of the latent variables. Each message is associated with an author entity a_d and a time interval t_d within which that messages was sent. The latent variables are the general topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, the entity topics $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_A$, and the event topics $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_T$, as well as the message-specific strengths $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$, $\zeta_1, \ldots, \zeta_D$, and $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_D$, the entity-specific strengths $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A$ and $\xi_1, \ldots, \xi_A$, and the event strengths $\psi_1, \ldots, \psi_T$. See figures 3 and 4 for the graphical model and generative process.

As for many Bayesian models, the posterior distribution is not tractable to compute; we must instead approximate it. We therefore introduce an approximate inference algorithm for Capsule, based on variational methods (Jordan et al., 1999; Wainwright and Jordan, 2008). Variational methods approximate the true posterior distribution p with a (simpler) variational distribution q. Inference then consists of minimizing the KL divergence from q to p. This is equivalent to maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(q) = \mathbb{E}_q \left[\log p(\mathbf{N}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) - \log q(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi})\right]. \tag{10}$$

We define q using the mean field assumption:

$$\begin{aligned} &q(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) = \\ &\prod_{d=1}^{D} \left(q(\zeta_d \mid \lambda_d) \prod_{k=1}^{K} q(\theta_{dk} \mid \lambda^{\theta}_{dk}) \prod_{t=1}^{T} q(\epsilon_{dt} \mid \lambda^{\epsilon}_{dt}) \right) \times \\ &\prod_{k=1}^{K} \left(q(\boldsymbol{\beta}_k \mid \lambda^{\beta}_k) \prod_{a=1}^{A} q(\phi_{ak} \mid \lambda^{\phi}_{ak}) \right) \prod_{a=1}^{A} \left(q(\boldsymbol{\eta}_a \mid \lambda^{\eta}_a)\, q(\xi_a \mid \lambda^{\xi}_a) \right) \prod_{t=1}^{T} \left(q(\boldsymbol{\gamma}_t \mid \lambda^{\gamma}_t)\, q(\psi_t \mid \lambda^{\gamma}_t) \right. \end{aligned} \tag{11}$$

The variational distributions for the topics $q(\boldsymbol{\beta}_k)$, $q(\boldsymbol{\eta}_a)$, and $q(\boldsymbol{\gamma}_t)$ are all Dirichlet distributions with free variational parameters λ^{β}_k, λ^{η}_a, and λ^{γ}_t, respectively. The variational distributions for the strengths $q(\theta_{dk})$, $q(\zeta_d)$, $q(\epsilon_{dt})$, $q(\phi_{ak})$, $q(\xi_a)$, and $q(\psi_t)$ are all gamma distributions with free variational parameters λ^{θ}_{dk}, λ^{ζ}_d, λ^{ϵ}_{dt}, λ^{ϕ}_{ak}, λ^{ξ}_a, and λ^{ψ}_t, respectively. Each of these parameters has two components: shape s and rate r.

The expectations under q, which we need to maximize the ELBO, have closed analytic forms. We therefore update each free variational parameter in turn, following a standard coordinate-ascent approach.

To obtain update equations for the free variational parameters, we introduce auxiliary latent variables:

$$z^{\mathcal{K}}_{dkv} \sim \text{Poisson}\left(\theta_{dk}\beta_{kv}\right) \tag{12}$$

$$z^{\mathcal{A}}_{dv} \sim \text{Poisson}\left(\zeta_d \eta_{a_d v}\right) \tag{13}$$

$$z^{\mathcal{T}}_{dtv} \sim \text{Poisson}\left(f(t_d, t)\,\epsilon_{dt}\gamma_{tv}\right), \tag{14}$$

where the superscripts $\mathcal{K}$, $\mathcal{A}$, and $\mathcal{T}$ indicate the general, entity, and event topics, respectively. When marginalized out, these variables—collectively $\mathbf{z}$—leave the model intact. Because the Poisson distribution has an additive property, the value of n_{dv} is completely determined by the values of these variables:

$$n_{dv} = \sum_{k=1}^{K} z^{\mathcal{K}}_{dkv} + z^{\mathcal{A}}_{dv} + \sum_{t=1}^{T} z^{\mathcal{T}}_{dtv}. \tag{15}$$

Coordinate-ascent variational inference depends on the conditional distribution of each latent variable given the values of the other latent variables and the data. We use $D(a)$ to denote the set of messages sent by entity a and $D(t)$ to denote the set of messages potentially affected by event t (e.g., all messages sent after time interval t, in the case of an exponential decay function). The conditional distributions are:

$$(\boldsymbol{\beta}_k \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Dirichlet}_V\left(\alpha + \sum_{d=1}^{D} z^{\mathcal{K}}_{dk1}, \ldots, \alpha + \sum_{d=1}^{D} z^{\mathcal{K}}_{dkV}\right) \tag{16}$$

$$(\boldsymbol{\eta}_a \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Dirichlet}_V\left(\alpha + \sum_{d \in D(a)} z^{\mathcal{A}}_{d1}, \ldots, \alpha + \sum_{d \in D(a)} z^{\mathcal{A}}_{dV}\right) \tag{17}$$

$$(\boldsymbol{\gamma}_t \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Dirichlet}_V\left(\alpha + \sum_{d \in D(t)} z^{\mathcal{T}}_{d1t}, \ldots, \alpha + \sum_{d \in D(t)} z^{\mathcal{T}}_{dVt}\right) \tag{18}$$

$$(\theta_{dk} \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Gamma}\left(s + \sum_{v=1}^{V} z^{\mathcal{K}}_{dkv},\ \phi_{a_d k} + \sum_{v=1}^{V} \beta_{kv}\right) \tag{19}$$

$$(\zeta_d \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Gamma}\left(s + \sum_{v=1}^{V} z^{\mathcal{A}}_{dv},\ \xi_{a_d} + \sum_{v=1}^{V} \eta_{a_d v}\right) \tag{20}$$

$$(\epsilon_{dt} \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Gamma}\left(s + \sum_{v=1}^{V} z^{\mathcal{T}}_{dtv},\ \psi_t + f(t_d, t) \sum_{v=1}^{V} \gamma_{tv}\right) \tag{21}$$

$$(\phi_{ak} \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Gamma}\left(s + |D(a)|\, s,\ r + \sum_{d \in D(a)} \theta_{dk}\right) \tag{22}$$

$$(\xi_a \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\psi}) \sim \text{Gamma}\left(s + |D(a)|\, s,\ r + \sum_{d \in D(a)} \zeta_d\right) \tag{23}$$

$$(\psi_t \mid \mathbf{N}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}) \sim \text{Gamma}\left(s + |D(t)|\, s,\ r + \sum_{d \in D(t)} \epsilon_{dt}\right). \tag{24}$$

The conditional distribution of the auxiliary latent variables is:

$$(\langle \mathbf{z}^{\mathcal{K}}_{dv}, \mathbf{z}^{\mathcal{A}}_{dv}, \mathbf{z}^{\mathcal{T}}_{dv} \rangle \mid \mathbf{N}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\epsilon}, \boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\psi}) \sim \text{Mult}\,(n_{dv}, \boldsymbol{\omega}_{dv})\,, \tag{25}$$

where

$$\boldsymbol{\omega}_{dv} \propto \langle \theta_{d1}\beta_{1v}, \ldots, \theta_{dK}\beta_{Kv}, \zeta_d \eta_{a_d v}, f(t_d, 1)\,\epsilon_{d1}\gamma_{1v}, \ldots, f(t_d, T)\,\epsilon_{dT}\gamma_{Tv} \rangle. \tag{26}$$

Given the conditional distributions, coordinate-ascent variational inference involves setting each free variational parameter to the expected value of the corresponding model parameter under the variational distribution. We provide pseudeocode in algorithm 1; we use $\boldsymbol{\lambda}$ to denote the entire set of free variational parameters and $V(d)$ to denote the set of vocabulary terms present in document d. Our approximate inference algorithm produces a fitted variational posterior distribution which can then be used as a proxy for the true posterior distribution. The source code is available online at `https://github.com/ajbc/capsule`.

B Additional Results

In this appendix, we provide additional results for Capsule. Table 5 lists the highest-ranked messages for an event described in section 5. Tables 6 and 7 contain examples of general topics and entity topics, respectively.

B.1 Decay Functions

We assessed Capsule's sensitivity to several decay functions. We considered an exponential function,

$$f(t_d, t) = \begin{cases} 0 & t \leq t_d < t + \tau \\ \exp\left(\frac{-(t_d - t)}{\tau\,/\,5}\right) & \text{otherwise,}[1] \end{cases} \tag{27}$$

a linear function,

$$f(t_d, t) = \begin{cases} 1 - \frac{(t_d - t)}{\tau} & t \leq t_d < t + \tau \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

and a step function,

$$f(t_d, t) = \begin{cases} 1 & t \leq t_d < t + \tau \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

For each decay function, we used Capsule's generative process, with $\tau = 3$, to create ten simulated data sets. We used three versions of our approximate inference algorithm—one for each decay function—to obtain a fitted variational distribution for each data set. We experimented with $\tau = 1, \ldots, 5$. Figure 7 shows event detection results (using equation (8)) and message identification results (using precision at ten messages).

As expected, Capsule performs best when the decay function used in the inference algorithm matches the decay function used to generate the data. For both event detection and message identification, the exponential function is least sensitive to the value of τ used to generate the data and the value of τ used in the inference algorithm. We also found that the exponential function gave the most interpretable results for real messages.

[1] Unlike the linear and step functions, the exponential function could be evaluated for any time interval t after message d's appearance at t_d; however, we truncate the function for computational reasons. The mean lifetime of the exponential decay is τ divided by five, which ensures that 99.3% of the area under the curve has occurred before we truncate the function at τ.

Algorithm 1: Coordinate-ascent variational inference for Capsule.

Input: observed term counts $\mathbf{N}$
Output: approximate posterior distribution of the latent variables, in terms of free variational parameters $\boldsymbol{\lambda}$
Initialize $\mathbb{E}[\boldsymbol{\beta}_k]$ to slightly random around uniform for each general topic k
Initialize $\mathbb{E}[$ all other latent variables $]$ to uniform
for *iteration* $m = 1, \ldots, M$ **do**
  set $\lambda^{\theta,r}$, $\lambda^{\zeta,r}$, and $\lambda^{\epsilon,r}$ to 0 and set remaining $\boldsymbol{\lambda}$ using priors
  update $\lambda^{\theta,r}_{dk} \mathrel{+}= \sum_V \mathbb{E}[\boldsymbol{\beta}_{kv}]$ for each message d and general topic k
  for *message* $d = 1, \ldots, D$ **do**
    for *term* $v \in V(d)$ **do**
      set $\boldsymbol{\omega}_{dv}$ using expected values of the latent variables (equation (26))
      set $\mathbb{E}[\langle \mathbf{z}^{\mathcal{K}}_{dv}, \mathbf{z}^{\mathcal{A}}_{dv}, \mathbf{z}^{\mathcal{T}}_{dv} \rangle] = n_{dv}\boldsymbol{\omega}_{dv}$
      update $\lambda^{\beta}_{kv} \mathrel{+}= \mathbb{E}[z^{\mathcal{K}}_{dkv}]$ for each general topic k (equation (16))
      update $\lambda^{\eta}_{a_d v} \mathrel{+}= \mathbb{E}[z^{\mathcal{A}}_{dv}]$ (equation (17))
      update $\lambda^{\gamma}_{tv} \mathrel{+}= \mathbb{E}[z^{\mathcal{T}}_{dtv}]$ for each time interval t (equation (18))
      update $\lambda^{\theta,s}_{dk} \mathrel{+}= \mathbb{E}[z^{\mathcal{K}}_{dkv}]$ for each general topic k (equation (19))
      update $\lambda^{\zeta,s}_{d} \mathrel{+}= \mathbb{E}[z^{\mathcal{A}}_{dv}]$ (equation (20))
      update $\lambda^{\epsilon,s}_{dt} \mathrel{+}= \mathbb{E}[z^{\mathcal{K}}_{dtv}]$ for each time interval t (equation (21))
    end
    set $\lambda^{\theta,r}_{dk} = \mathbb{E}[\phi_{a_d k}] + \sum_v \mathbb{E}[\beta_{kv}]$ for each general topic k (equation (19))
    set $\lambda^{\zeta,r}_{d} = \mathbb{E}[\xi_{a_d}] + \sum_v \mathbb{E}[\eta_{a_d v}]$ (equation (20))
    set $\lambda^{\epsilon,r}_{dt} = \mathbb{E}[\psi_t] + f(t_d, t) \sum_v \mathbb{E}[\gamma_{tv}]$ for each time interval t (equation (21))
    set $\mathbb{E}[\theta_{dk}] = \lambda^{\theta,s}_{dk} \,/\, \lambda^{\theta,r}_{dk}$ for each general topic k
    set $\mathbb{E}[\zeta_d] = \lambda^{\zeta,s}_{d} \,/\, \lambda^{\zeta,r}_{d}$
    set $\mathbb{E}[\epsilon_{dt}] = \lambda^{\epsilon,s}_{dt} \,/\, \lambda^{\epsilon,r}_{dt}$ for each time interval t
    update $\lambda^{\phi,s}_{a_d k} \mathrel{+}= s$ for each general topic k (equation (22))
    update $\lambda^{\xi,s}_{a_d} \mathrel{+}= s$ (equation (23))
    update $\lambda^{\psi,s}_{t} \mathrel{+}= s$ for each time interval t where $f(t_d, t) \neq 0$ (equation (24))
    update $\lambda^{\phi,r}_{a_d k} \mathrel{+}= \theta_{dk}$ for each general topic k (equation (22))
    update $\lambda^{\xi,r}_{a_d} \mathrel{+}= \zeta_d$ (equation (23))
    update $\lambda^{\psi,r}_{t} \mathrel{+}= \epsilon_{dt}$ for each time interval t (equation (24))
  end
  set $\mathbb{E}[\boldsymbol{\beta}_k] = \boldsymbol{\lambda}^{\beta}_{k} \,/\, \sum_v \lambda^{\beta}_{kv}$ for each general topic k
  set $\mathbb{E}[\boldsymbol{\eta}_a] = \boldsymbol{\lambda}^{\eta}_{a} \,/\, \sum_v \lambda^{\eta}_{av}$ for each entity a
  set $\mathbb{E}[\boldsymbol{\gamma}_t] = \boldsymbol{\lambda}^{\gamma}_{t} \,/\, \sum_v \lambda^{\gamma}_{tv}$ for each time interval t
  set $\mathbb{E}[\phi_{ak}] = \lambda^{\phi,s}_{ak} \,/\, \lambda^{\phi,r}_{ak}$ for each entity a and general topic k
  set $\mathbb{E}[\xi_a] = \lambda^{\xi,s}_{a} \,/\, \lambda^{\xi,r}_{a}$ for each entity a
  set $\mathbb{E}[\psi_t] = \lambda^{\psi,s}_{t} \,/\, \lambda^{\psi,r}_{t}$ for each time interval t
end
return $\boldsymbol{\lambda}$

$f(t_d, t)\ \mathbb{E}[\epsilon_{dt}]$	**Date**	**Author Entity**	**Subject**
6.86	1976-07-07	Cairo	Possible SC meeting on Israeli rescue operation
6.18	1976-07-10	Kuwait	Media reaction to Bicentennial summary
6.15	1976-07-06	Damascus	Syria condemns Israeli operation to free Air France ...
5.91	1976-07-08	Tel Aviv	Passengers comment on Air France hijacking
5.89	1976-07-06	Stockholm	Possible SC meeting on Israeli rescue operation
5.38	1976-07-09	Nicosia	Bicentennial activities in Cyprus
5.09	1976-07-11	State	Security Council debate on Entebbe events CONFID...
4.77	1976-07-09	State	Travel of Peter M. Storm, House Budget Committee
4.76	1976-07-06	Jidda	Weekly Saudi Editorial Summary (June 30-July 6)
4.68	1976-07-08	Lusaka	SWAPO President seeks assessment of Kissinger-Vor...
4.56	1976-07-07	Stockholm	Ugandan role in Air France hijacking
4.45	1976-07-06	Karachi	Transitional quarter funding for RSS travel
4.43	1976-07-06	Athens	Bicentennial anniversary in Greece
4.37	1976-07-08	Damascus	Beirut travel
4.34	1976-07-10	State	Status of Mrs. Bloch
4.17	1976-07-07	Hong Kong	Hong Kong Communist press denounces Israeli resc...
4.12	1976-07-08	Dar es Salaam	President Nyerere's fourth of July messages
4.09	1976-07-10	Moscow	Pravda and Krasnaya Zvezda on Entebbe rescue oper...

Table 5: Highest-ranked messages for the week immediately following the U.S. Bicentennial Celebration and Operation Entebbe. Capsule accurately recovers messages related to both of these real-world events. Typos are intentionally copied from the data.

References

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.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, January.

Top Terms
church, vatican, catholic, bishop, pope, ford, cardinal, ban, religious, archbishop
program, university, grant, education, school, post, institute, research, center, american
security, council, terrorist, threat, sc, sabotage, protective, herein, unsc, honour
visit, hotel, schedule, arrival, arrive, depart, please, meet, day, room
labor, union, strike, ilo, employment, federation, afl cio, trade, worker, confederation
bank, credit, loan, investment, finance, payment, financial, eximbank, opic, central
law, case, court, legal, investigation, arrest, justice, sentence, trial, attorney
party, government, election, opposition, national, leader, campaign, vote, support, anti
tax, company, pay, lease, compensation, exemption, repatriation, income, taxation, fee
oil, petroleum, opec, crude, gulf, price, exploration, refinery, energy, company
israel, arab, israeli, middle, egypt, peace, plo, cairo, egyptian, lebanon
radio, television, broadcast, allotment, appropriation, obligation, zero, warc, transmitter, network
india, indian, pakistan, delhi, goi, ocean, bangladesh, transit, pakistani, afghan
turkish, turkey, cyprus, greek, greece, athens, ankara, morocco, cypriot, algeria
aid, relief, emergency, usaid, disaster, donor, wfp, sahel, ifad, unicef
aircraft, team, flight, clearance, transport, civair, aviation, traffic, charter, cargo
soviet, moscow, press, ussr, soviet union, american, one, war, communist, article
sea, zone, marine, maritime, fish, coastal, continental, territorial, mile, fishery

Table 6: The highest-probability vocabulary terms for a selection of general topics (one per row) according to $\mathbb{E}[\boldsymbol{\beta}_1], \ldots, \mathbb{E}[\boldsymbol{\beta}_K]$. These examples come from the analysis described in section 5. Capsule identifies diplomatic themes that are relevant to any entity.

Entity	Top Terms
Ankara	turkish, turkey, ankara, government, cyprus, greek, party, one, time
Athens	greek, athens, greece, gog, government, cyprus, turkish, press, minister
Auckland	new zealand, company, box, trade, contact, opportunity, united states
Baghdad	iraqi, iraq, goi, arab, state, regime, ministry, government, party
Berlin	berlin, frg, german, senat, time, bonn, trade, one, agreement
Bern	swiss, bern, federal, bank, snb, gold, end, interest, national
Brussels	belgian, belgium, brussels, government, firestone, european, ministry
Budapest	hungarian, hungary, trade, mudd, one, time, puja, well, policy
Buenos Aires	argentine, argentina, goa, us, hill, government, one, press, police
Cairo	egyptian, cairo, egypt, arab, israeli, israel, peace, agreement, president
Canberra	australian, australia, goa, government, minister, whitlam, end, dfa, time
Dakar	senegalese, president, african, summary, conference, end, support, one
Dar es Salaam	tangov, salaam, tanzanian, spain, president, government, african, one
Guayaquil	ecuador, ecuadorean, port, congen, one, tuna, local, time, boat
Islamabad	pakistan, gop, government, one, party, minister, general, opposition, ppp
Paris	paris, france, rush, french, one, government, amconsul, quai, european
Jerusalem	jerusalem, bank, israeli, us, israel, plo, one, arab, unifil
Jidda	saudi, jidda, saudi arabia, prince, us, fahd, one, time, government
Johannesburg	black, africa, african, trade, union, police, labor, one, committee
Kabul	afghan, government, goa, minister, one, pakistan, regime, time, ministry
Lima	peru, gop, lima, peruvian, dean, minister, general, marcona, government
Lisbon	portugal, portuguese, gop, lisbon, government, party, summary, minister
London	london, british, government, fco, labor, agreement, one, washdc, summary
Madrid	spanish, spain, madrid, one, govt, general, committee, government, time
Nairobi	kenya, nairobi, marshall, embassy, kenyan, unep, le, ref, state
Oslo	norwegian, norway, soviet, government, minister, ministry, policy
Ottawa	canadian, canada, goc, ottawa, us, extaff, government, minister, federal
Peking	chinese, peking, uslo, china, people, teng, one, trade, delegation, hong
Phnom penh	penh, phnom, khmer, rice, fank, enemy, cambodia, government, dean
Prague	czechoslovak, goc, czech, trade, embassy, one, mfa, time, cssr
Quito	ecuador, ecuadorean, gulf, government, minister, bloomfield, general, one
Sao Paulo	paulo, brazil, state, brazilian, president, government, congen, one, do
Seoul	korea, korean, rok, rokg, seoul, park, government, president, time
Singapore	singapore, asean, minister, government, one, prime, comment, vietnam
Sofia	bulgarian, trade, one, agreement, american, visit, committee, party
Sydney	australia, australian, one, general, american, state, government, post
Tokyo	japan, japanese, tokyo, fonoff, summary, miti, end, diet, time
Taipei	taiwan, groc, china, chinese, government, american, one, local, republic
The Hague	dutch, netherlands, hague, government, minister, party, stoel, mfa, one
USUN New York	committee, usun, priority, report, draft, resolution, sc, comite, rep, new york
Vancouver	canada, government, canadian, british, columbia, pipeline, federal, editorial
Zagreb	yugoslav, yugoslavia, croatian, fair, belgrade, american, one, ina, summary
Zurich	swiss, congen, consulate, general, american, bern, dollar, shipment

Table 7: The highest-probability vocabulary terms for a selection of entity topics (one per row) according to $\mathbb{E}[\eta_1], \ldots, \mathbb{E}[\eta_A]$. These examples come from the analysis described in section 5. Capsule identifies themes and interests that are specific to the entities.

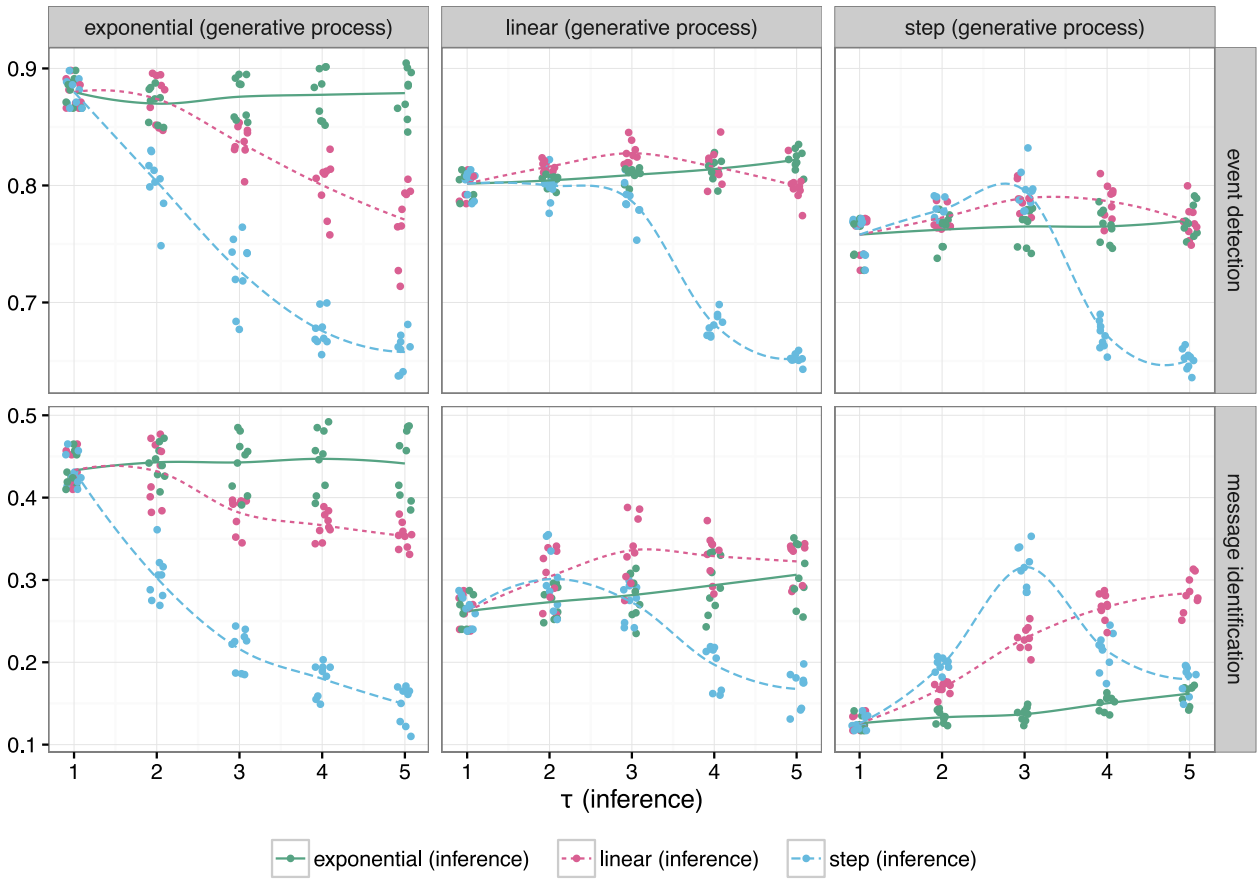

Figure 7: Capsule's sensitivity to several different decay functions (exponential, linear, and step) using simulated data. Capsule performs best when the decay function used in the inference algorithm matches the decay function used to generate the data. The exponential function is least sensitive to the value of τ used to generate the data and the value of τ used in the inference algorithm.