Who, What, When, Where, and Why? A Computational Approach to Understanding Historical Events Using State Department Cables

Allison J.B. Chaney, Hanna Wallach, David M. Blei

October 6, 2015

> *We can do nothing but scrutinize historical events themselves if we want to discover what they are.*
>
> – Dean W.R. Matthews, *What is an Historical Event?*

Abstract

We develop computational methods for analyzing historical documents to identify events of potential historical significance. Significant events are characterized by interactions between entities (e.g., countries, organizations, individuals) that deviate from typical interaction patterns. When studying historical events, historians and political scientists commonly read large quantities of text to construct an accurate picture of who, what, when, and where—a necessary precursor to answering the more nuanced question, "Why?" Our methods help historians identify possible events from the texts of historical communication. Specifically, we build on topic modeling to distinguish between topics that describe "business-as-usual" and topics that deviate from these patterns, where deviations are also indicated by particular entities interacting during particular periods of time. To demonstrate our methods, we analyze a corpus of 2 million State Department cables from 1973 to 1977. For example, we show that we are able to detect and characterize the Fall of Saigon.

1 Introduction

Communications between the U.S. State Department and its embassies have historically been called *diplomatic cables*, derived from the time when physical cables were used for such communications. We obtained around two million of these cables sent between 1973 and 1977 via the History Lab at Columbia,[1] which received them from the Central Foreign Policy Files at the National Archives. 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 message (often multiple entities), and the date the message was sent. Figure 1 shows three cables which were sent in April 1975 concerning orphans from the Vietnam War.

Historians and political scientists are interested in the process of sending cables—of particular interest is identifying historical events in a collection of these cables. These experts want

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

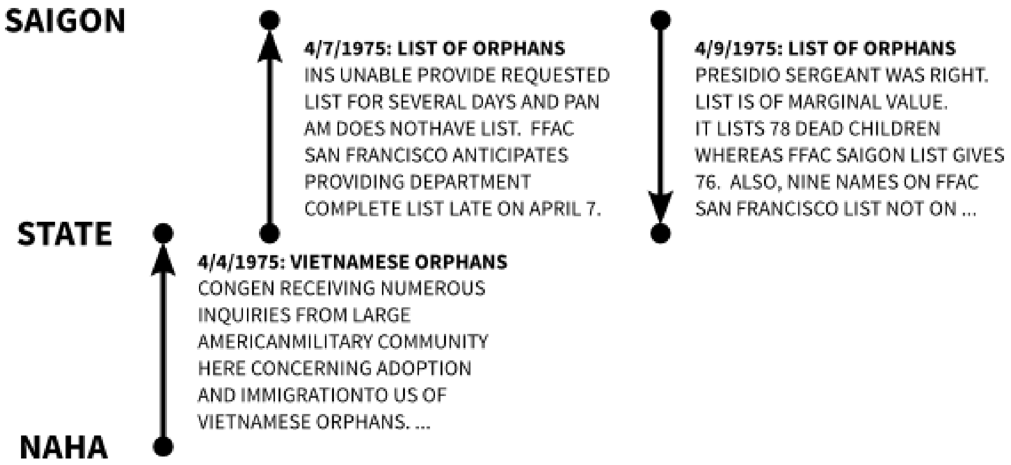
SAIGON
STATE
NAHA
4/7/1975: LIST OF ORPHANS
INS UNABLE PROVIDE REQUESTED LIST FOR SEVERAL DAYS AND PAN AM DOES NOTHAVE LIST. FFAC SAN FRANCISCO ANTICIPATES PROVIDING DEPARTMENT COMPLETE LIST LATE ON APRIL 7.
4/9/1975: LIST OF ORPHANS
PRESIDIO SERGEANT WAS RIGHT. LIST IS OF MARGINAL VALUE. IT LISTS 78 DEAD CHILDREN WHEREAS FFAC SAIGON LIST GIVES 76. ALSO, NINE NAMES ON FFAC SAN FRANCISCO LIST NOT ON ...
4/4/1975: VIETNAMESE ORPHANS
CONGEN RECEIVING NUMEROUS INQUIRIES FROM LARGE AMERICANMILITARY COMMUNITY HERE CONCERNING ADOPTION AND IMMIGRATIONTO US OF VIETNAMESE ORPHANS. ...

Figure 1: Example excerpts of cables.

to know *when* events happen, as well as *what* happens during the event, which includes the parties sending and receiving messages (*who*).

For this work, we characterize an event in two ways: *when* it occurs and *what* occurs. We assume that only one event can begin on any given day, which allows us to describe each day in terms of the probability of an event starting that day. We can also assume that an event will influence cables for a set time period after it starts—this way, we only need to discover the starting day for each event. To understand *what* occurs, we can summarize the cable message content with a topic model such as LDA (Blei et al., 2003) and model event content in that same space.

2 A Generative Model of Events

Our model is a generative process—we make assumptions about how the data came to be and describe these assumptions in terms of probability distributions. Given our model and observed data, the task is then to reverse the generative process to find the hidden quantities that (retrospectively) generated the data.

Consider an entity like the Bangkok American embassy. We can imagine that there is a stream of cables being sent by this embassy—some might be sent to the US State Department, others to another American embassy like Hong Kong, and perhaps a few are sent to individuals by name. Each of these cables has an associated send date, and we can represent the content of the cable with a topic model; we call these cable descriptions θ, which is a matrix of D cables (or documents) by K topics.[2]

An entity will usually talk about certain topics and with certain frequency. The Bangkok embassy, for instance, sent and average of 22 cables per day in the 1970s, and was concerned with topics regarding southeast Asia more generally. We can describe the general interests of entities in the same topic space we use to describe individual cables and we will call these per-entity interests ϕ and can assign them a formal distribution.

- Draw the entity's base topics: $\phi_{0k} \sim \text{Gamma}(\alpha, \beta)$

Eventually we will want to model the interactions between entities, but for now we can consider a single entity at a time.

[2] This allows us to represent the cable in terms of about 100 topics rather than in terms of hundreds of thousands of vocabulary words. We can discover these topics with LDA and treat them as fixed observations going forward.

Now imagine that at a particular time, an event occurs, such as the capture of Saigon during the Vietnam war. We do not directly observe that events occurs, but each event can again be described in the same topic space used to describe individual cables. Whether or not an event occurs at a particular time step is represented by ϵ_t and the content of the event (or its topical representation) is called π_t.

- For each day i with date a_i:
  - Generate event occurrence/strength $\epsilon \sim \text{Poisson}(\eta_\epsilon)$,[3] where η_ϵ is a fixed, non-negative hyperparameter for the mean event strength.
  - Generate the day/event's description in terms of each topic k: $\pi_{ik} \sim \text{Gamma}(\alpha_0, \beta_0)$, where α_0 and β_0 are fixed hyperparameters.

When an event occurs, both the frequency of cables being sent and the cable content changes. The Bangkok embassy sent 31 cables the day following the capture of Saigon (a 36% increase over the average), and the majority of these cables are about Vietnam war refugees. Thus we imagine that an entity's stream of cables is controlled by what it usually talks about (and how often) as well as the higher level stream of unobserved events. The influence of an event does not last indefinitely, however, so we model the decay of its magnitude with some function f. While many decay functions can be used, we define the event decay function to be a simple linear decrease:

$$f(a,c) = \begin{cases} 1 - \frac{c-a}{d}, & \text{if } a \leq c < a + d \\ 0, & \text{otherwise,} \end{cases}$$

where d is the time distance (in days) after event a at which point the event is no longer relevant.

When we analyze the cables with this model setup, we disentangle cables that represent "business as usual" from those that reflect the higher-order event stream. Consequently, we infer what that stream is, i.e., when something happened and what happened.

Recall that the key hidden values are event descriptions π ("what"), event occurrences ϵ ("when"), and entity interests ϕ. (Since entities are tied to individuals and places, we can use them to describe "who" is involved and "where" and event occurs after fitting our model.) These are all tied together when we model the observed data itself, as we assume that each cable is generated from a combination of these parameters.

- For each cable j on date c_j:
  - Set cable topic parameter: $\phi_{jk} = \phi_{0k} + \sum_i f(a_i, c_j)\pi_{ik}\epsilon_i$.
  - Draw cable topic: $\theta_{jk} \sim \text{Gamma}(\beta_c \phi_{jk}, \beta_c)$.

2.1 Inference

Posterior inference is the central computational problem. We want to learn the hidden values describe above (event descriptions π, event occurrences ϵ, and entity descriptions ϕ) from our observed data. We construct a black box variational inference algorithm following Ranganath et al. (2014) to determine the values of these latent parameters. Full details on the derivation of this inference algorithm can be found in Appendix A.

[3] Event occurrence can alternatively be modeled with a Bernoulli distribution. The advantage to the Bernoulli is that it is more interpretable. Poisson-distributed event occurrences lay the groundwork for more complicated models based in Poisson processes, and allow for a nuanced interpretation of multiple real-world events occurring in a single day, but convolving into a single event in model space.

3 Discussion

In this section, we discuss how to validate our model and explore some preliminary results.

Validation We have collected examples of known historical events and have manually identified multiple cables that are associated with each event. We plan to compare the events discovered by our model to these known events. We will assess how many of the known events are recovered with our model, as well as how the average topic distribution of the known cables compares to the discovered event distribution.

Another approach to validating our model is to present the discovered events—including date, topic distribution, and entities involved—to an expert historian or political scientist who can assess the value of the discovered events.

A technical approach to validation is to verify that the ELBO is increasing appropriately during inference. The ELBO is discussed in greater detail in Appendix A, and Figure 2 (also in the appendix) show a plot of the ELBO on a small sample of data.

Exploration

Conclusions Traditional topic models can describe documents, but they cannot identify when events occur—only a model like ours that explicitly models event occurrences and event content can attribute document observations to discrete events. Further, by modeling entities, we can distinguish between "business-as-usual" document content and content that is attached to particular events—we are also unable to capture this phenomenon with traditional topic models.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *JMLR*, 3:993–1022.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *AISTATS*.

A Inference Details

For now, we assume that we know the LDA topics β and only observe the documents in terms of their topics θ; breaking this assumption makes inference a little more complicated as the updates for θ would have new dependencies.

As usual, inference is the central computational problem. Variational inference approaches this problem by minimizing the KL divergence from an approximating distribution q to the true posterior p. This is equivalent to maximizing the ELBO:

$$\mathscr{L}(q) = \mathbb{E}_{q(\epsilon,\pi,\phi)}[\log p(\theta,\epsilon,\pi,\phi) - \log q(\epsilon,\pi,\phi)].$$

We define the approximating distribution q using the mean field assumption:

$$q(\epsilon,\pi,\phi) = \prod_i q(\epsilon_i) \prod_k \left[q(\phi_{0k}) \prod_i q(\pi_{ik}) \right].$$

The variational distributions $q(\pi)$ and $q(\phi)$ are both gamma-distributed with free variational parameters λ^π and λ^ϕ, respectively. Because these parameters are free, we use the softmax function $\mathscr{M}(x) = \log(1 + \exp(x))$ to constrain them so that they do not violate the requirements of the gamma distribution. The variational distribution $q(\epsilon)$ is Poisson-distributed with variational parameter λ^ϵ; this free parameter is also constrained by the softmax function.

The expectations under q, which are needed to maximize the ELBO, do not have a simple analytic form, so we use "black box" VI techniques. F or each variable, we can write the log probability of all terms containing that variable, giving us

$$\log p_i^\epsilon(\theta,\epsilon,\pi,\phi) \triangleq \log p(\epsilon_i \mid \eta_\epsilon) + \sum_{j:f(a_i,c_j)\neq 0} \sum_k \log p(\theta_{jk} \mid \phi_{0k}, c_j, a_i, d, \beta_c, \pi_{ik}, \epsilon_i),$$

$$\log p_{ik}^\pi(\theta,\epsilon,\pi,\phi) \triangleq \log p(\pi_{ik} \mid \alpha_0, \beta_0) + \mathbf{1}[\epsilon_i \neq 0] \sum_{j:f(a_i,c_j)\neq 0} \log p(\theta_{jk} \mid \phi_{0k}, c_j, a_i, d, \beta_c, \pi_{ik}, \epsilon_i),$$

and

$$p_k^\phi(\theta,\epsilon,\pi,\phi) \triangleq \log p(\phi_{0k} \mid \alpha, \beta) + \sum_j \log p(\theta_{jk} \mid \phi_{0k}, c_j, a_i d, \beta_c, \pi_{ik}, \epsilon_i).$$

Then we can write the gradients with respect to the variational parameters as:

$$\nabla_{\lambda_i^\epsilon} \mathscr{L} = \mathbb{E}_q \left[\nabla_{\lambda_i^\epsilon} \log q(\epsilon_i \mid \lambda_i^\epsilon) \left(\log p_i^\epsilon(\theta,\epsilon,\pi,\phi) - \log q(\epsilon_i \mid \lambda_i^\epsilon) \right) \right],$$

$$\nabla_{\lambda_{ik}^\pi} \mathscr{L} = \mathbb{E}_q \left[\nabla_{\lambda_{ik}^\pi} \log q(\pi_{ik} \mid \lambda_{ik}^\pi) \left(\log p_{ik}^\pi(\theta,\epsilon,\pi,\phi) - \log q(\pi_{ik} \mid \lambda_{ik}^\pi) \right) \right],$$

and

$$\nabla_{\lambda_k^\phi} \mathscr{L} = \mathbb{E}_q \left[\nabla_{\lambda_k^\phi} \log q(\phi_{0k} \mid \lambda_k^\phi) \left(\log p_k^\phi(\theta,\epsilon,\pi,\phi) - \log q(\phi_{0k} \mid \lambda_k^\phi) \right) \right].$$

Using this framework, we construct our black box algorithm below in Algorithm 1.

For Reference The gamma distribution and derivatives:

$$\log \mathrm{Gamma}(x \mid a, b) = ab \log b - \log \Gamma(ab) + (ab-1)\log x - bx \tag{1}$$

$$\nabla_a \log \mathrm{Gamma}(x \mid \mathscr{M}(a), \mathscr{M}(b)) = \mathscr{M}'(a)\mathscr{M}(b)[\log \mathscr{M}(b) - \Psi(\mathscr{M}(a)\mathscr{M}(b)) + \log x] \tag{2}$$

$$\nabla_b \log \mathrm{Gamma}(x \mid \mathscr{M}(a), \mathscr{M}(b)) = \mathscr{M}'(b)[\mathscr{M}(a)((\log \mathscr{M}(b) + 1) - \Psi(\mathscr{M}(a)\mathscr{M}(b)) + \log x) - x] \tag{3}$$

The Poisson distribution and derivative:

$$\log \text{Poisson}(x \mid \lambda) = x \log \lambda - \log(x!) - \lambda \tag{4}$$

$$\nabla_\lambda \log \text{Poisson}(x \mid \mathscr{M}(\lambda)) = \mathscr{M}'(\lambda)\left[\frac{x}{\mathscr{M}(\lambda)} - 1\right]. \tag{5}$$

The softmax function and derivative:

$$\mathscr{M}(x) = \log(1 + e^x)$$

$$\mathscr{M}'(x) = \frac{e^x}{1 + e^x}$$

Validation of Implementation To show that our algorithm fits our model appropriately, Figure 2 shows that the ELBO is maximized. While the ELBO does decrease slowly after reaching a local maximum, this sort of over-fitting can be avoided with an appropriate convergence metric, like considering the change in data likelihood. Using this as a metric, the shown ELBO would have converged around iteration 10.

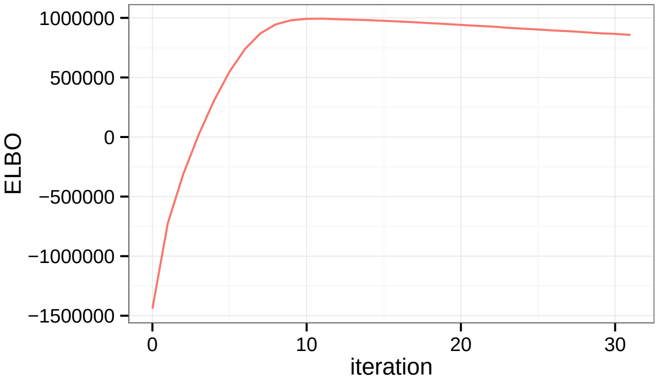

Figure 2: The ELBO on a small collection of documents.

```
Algorithm 1: Inference for Cables Model
Input: document topics $\theta$
Output: estimates of latent parameters entity topics $\phi$, event topics $\pi$, and event occurrences $\epsilon$
Initialize $\lambda^\pi$, $\lambda^\phi$, and $\lambda^\epsilon$ to respective priors
Initialize iteration count $t = 0$
while change in validation likelihood $< \delta$ do
    initialize $\sigma^\pi = 0$
    for each sample $s = 1, \ldots, S$ do
        for each component $k$ do
            draw sample entity topics $\phi_{0k}[s] \sim \text{Gamma}(\mathscr{M}(\lambda_k^\phi))$
            set $p_k^\phi[s] = \log p(\phi_{0k}[s] \mid \alpha, \beta)$                       // see Eqn. 1
            set $q_k^\phi[s] = \log q(\phi_{0k}[s] \mid \lambda_k^\phi)$          // Eqn. 1 with params $\mathscr{M}(\lambda_k^\phi)$
            set $g_k^\phi[s] = \nabla_{\lambda_k^\phi} \log q(\phi_{0k}[s] \mid \lambda_k^\phi)$                 // see Eqns. 2, 3
        end
        for each timestep $i$ do
            draw sample event occurance $\epsilon_i[s] \sim \text{Poisson}(\mathscr{M}(\lambda_i^\epsilon))$
            set $p_i^\epsilon[s] = \log p(\epsilon_i[s] \mid \eta)$                            // see Eqn. 4
            set $q_i^\epsilon[s] = \log q(\epsilon_i[s] \mid \lambda_i^\epsilon)$               // Eqn. 4 with param $\mathscr{M}(\lambda_i^\epsilon)$
            set $g_i^\pi[s] = \nabla_{\lambda_i^\epsilon} \log q(\epsilon_i[s] \mid \lambda_i^\epsilon)$                       // see Eqn. 5
            if $\epsilon_i[s] \neq 0$ then
                for each component $k$ do
                    draw sample event topics $\pi_{ik}[s] \sim \text{Gamma}(\mathscr{M}(\lambda_{ik}^\pi))$
                    set $p_{ik}^\pi[s] = \log p(\pi_{ik}[s] \mid \alpha_0, \beta_0)$             // see Eqn. 1
                    set $q_{ik}^\pi[s] = \log q(\pi_{ik}[s] \mid \lambda_{ik}^\pi)$    // Eqn. 1 with params $\mathscr{M}(\lambda_{ik}^\pi)$
                    set $g_{ik}^\pi[s] = \nabla_{\lambda_{ik}^\pi} \log q(\pi_{ik}[s] \mid \lambda_{ik}^\pi)$         // see Eqns. 2, 3
                end
            end
        end
    end
    for each document $j$ (sent on date $c_j$ and has topics $\theta_j$), sample $s$ and component $k$ do
        set $\phi_{jk}[s] = \phi_j[s] + \sum_i f(a_i, c_j)\epsilon_i[s]\pi_{ik}[s]$
        set $p_{jk}^\theta[s] = \log p(\theta_{jk} \mid \phi_{jk}[s], \beta_c)$                    // see Eqn. 1
        $p_k^\phi[s] \mathrel{+}= p_{jk}^\theta[s]$
        for each timestep $i$ where $a_i \leq c_j < a_i + d$ do
            $p_i^\epsilon[s] \mathrel{+}= \sum_k p_{jk}^\theta[s]$
            if $\epsilon_i[s] \neq 0$ then
                $p_{ik}^\pi[s] \mathrel{+}= p_{jk}^\theta[s]$
                update $\sigma_i^\pi \mathrel{+}= 1$
            end
        end
    end
    set $\hat{\nabla}_{\lambda^\phi}\mathscr{L} \triangleq \frac{1}{S}\sum_s g^\phi[s](p^\phi[s] - q^\phi[s])$
    set $\hat{\nabla}_{\lambda^\epsilon}\mathscr{L} \triangleq \frac{1}{S}\sum_s g^\epsilon[s](p^\epsilon[s] - q^\epsilon[s])$
    set $\hat{\nabla}_{\lambda^\pi}\mathscr{L} \triangleq \frac{1}{\sigma_\pi}\sum_s g^\pi[s](p^\pi[s] - q^\pi[s])$
    set $\rho = (t + \tau)^\kappa$
    set $\lambda^\pi \mathrel{+}= \rho\hat{\nabla}_{\lambda^\pi}\mathscr{L}$
    set $\lambda^\epsilon \mathrel{+}= \rho\hat{\nabla}_{\lambda^\epsilon}\mathscr{L}$
    set $\lambda^\phi \mathrel{+}= \rho\hat{\nabla}_{\lambda^\phi}\mathscr{L}$
end
set $\mathbb{E}[\pi] = \lambda^{\pi,a}$
set $\mathbb{E}[\phi] = \lambda^{\phi,a}$
set $\mathbb{E}[\epsilon] = \lambda^\epsilon$
return $\mathbb{E}[\pi]$, $\mathbb{E}[\phi]$, $\mathbb{E}[\epsilon]$
```