Mining events with declassified diplomatic documents

Citation: Gao, Yuanjun et al. "Mining events with declassified diplomatic documents." Annals of Applied Statistics, 14, 4 (December 2020): 1699 - 1723 © 2020 The Author(s)

As Published: 10.1214/20-AOAS1344

Publisher: Institute of Mathematical Statistics

Persistent URL: https://hdl.handle.net/1721.1/130446

Version: Original manuscript: author's manuscript prior to formal peer review

Mining Events with Declassified Diplomatic Documents

Yuanjun Gao*[1], Jack Goetz†[2], Rahul Mazumder‡[3], and Matthew Connelly §[4]

[1]Department of Statistics, Columbia University, New York, NY.
[2]Department of Statistics, University of Michigan, Ann Arbor, MI.
[2]Massachusetts Institute of Technology, Cambridge, MA.
[3]Department of History, Columbia University, New York, NY.

December 21, 2017

Abstract

Since 1973 the State Department has been using electronic records systems to preserve classified communications. Recently, approximately 1.9 million of these records from 1973-77 have been made available by the U.S. National Archives. While some of these communication streams have periods witnessing an acceleration in the rate of transmission; others do not show any notable patterns in communication intensity. Given the sheer volume of these communications – far greater than what had been available until now – scholars need automated statistical techniques to identify the communications that warrant closer study. We develop a statistical framework that can semi-automatically identify from a large corpus of documents a handful that historians would consider more interesting electronic records. Our approach brings together related but distinct statistical concepts from nonparametric signal estimation and statistical hypothesis testing – which when put together help us identify and analyze various geometrical aspects of the communication streams. Dominant periods of heightened and sustained activities aka *bursts*, as identified through these methods, correspond well with historical events recognized by standard reference works on the 1970s.

1 Introduction

For more than forty years, social scientists have been developing datasets of events for the quantitative analysis of world politics. The last decade has witnessed a dramatic increase in activity in this area, much of it focused on automatic event detection for purposes of explaining and predicting political crises [2]. All of these efforts however, have used public information, such as newspaper or wire service reporting. Rather than directly measuring political activity, these systems can only count

*gaoyuanjun0430@gmail.com
†jrgoetz@umich.edu
‡rahulmaz@mit.edu
§mjc96@columbia.edu

what reporters write about, which can vary over time and geography depending on many extraneous factors [13]. Together with the intrinsic challenges in automatic extraction, this has produced datasets that purport to track the same kind of events, such as political protests, but that are completely uncorrelated [11]. Moreover, some of the most important political activity is not immediately reported, and may not become publicly known until decades later, when formerly secret records are declassified. Even then, the sheer volume of these records can make it difficult even for the diligent researcher to identify individual events and assess their relative importance.

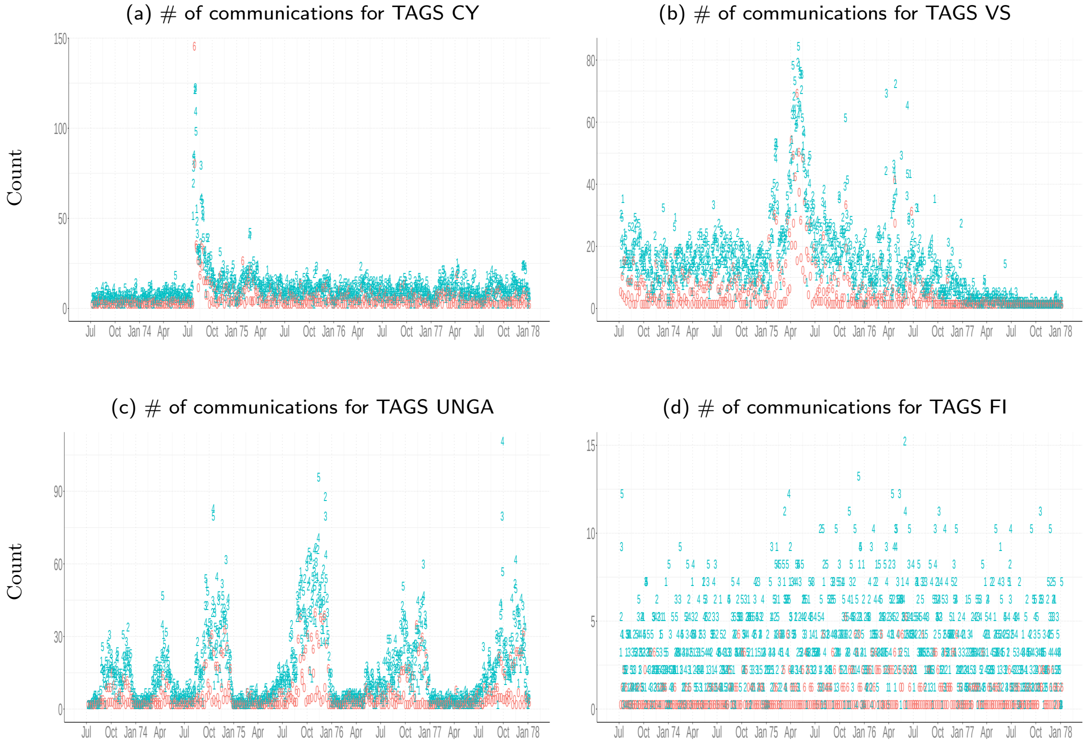
In this paper we study a new dataset of declassified documents and use statistical methods to identify and rank political events. Since 1973, the State Department has been using electronic records systems to preserve classified communications. The National Archives[1] now makes available over 1.4 million declassified cables from 1973-77 and in addition, the metadata of more than 0.4 million other communications originally delivered by diplomatic pouch. They are all machine-readable and rich with metadata, creating many opportunities for statistical modeling. Our goal is to explore modern statistical techniques that can automatically identify statistically interesting events in an important corpus of historical documents, which will continue to grow year-by-year as millions more communications are declassified. For these communication streams, we are interested in studying and identifying "interesting" statistical patterns – we contend that these patterns correspond to heightened diplomatic activity; and validate our findings with standard reference works on the 1970s. A statistically interesting pattern can mean several things. Loosely speaking, this can correspond to sudden localized changes or abrupt "jumps" in communication traffic, regardless of the overall series-specific baseline activity (a communication stream may be very active or have very low traffic intensity overall). There can also be continuous periods in a communication stream, where the data lies consistently above a series-specific baseline that corresponds to a representative global activity-level of that stream – these are "bursts" of activity in the temporal structure of the document streams that probably correspond with heightened diplomatic activity, such as the start or end of a war. An interesting event can also correspond to a heightened traffic intensity that plays out over longer periods, such as an increase over time, whether or not there are jumps.

When these communications were first entered in the State Department system, they were assigned one or more TAGS (Traffic Analysis by Geography and Subject), which indicate what countries or subjects each cable is related to. For example, "VS" signifies South Vietnam, and "SHUM" concerns human rights. By using these content-based TAGS as the feature, we avoid the complication of language processing and focus on identifying statistically relevant activity patterns in the communication streams.

1.1 A brief exploratory description of the data

A glimpse of processed data in the form of communication streams is shown in Figure 1. The data shows that there is less traffic on weekends and holidays (including the end of the year). In addition, the number of communications sent in 1973 seems

[1] Website: https://aad.archives.gov/aad/series-list.jsp?cat=WR43


(a) # of communications for TAGS CY

(b) # of communications for TAGS VS

(c) # of communications for TAGS UNGA

(d) # of communications for TAGS FI

Figure 1: Figures showing counts of communications sent on each day, in the period 1973-1978. The numbers in the plot represent day-of-week (0-Sunday, 1-Monday, 2-Tuesday, ..., 6-Saturday), with weekdays colored in blue and weekends in red. Figures (a)–(d) show the communications restricted to different TAGS. The apparent heightened activities in the communication streams correspond to (a) Cyprus coup (b) Fall of Saigon (the most intense one) (c) the yearly United Nation General Assembly meetings. There does not seem to be any interesting activity for panel (d), showing cables related to Finland. A goal of this paper is to create statistical methods to automatically identify series with heightened diplomatic communications and further describe their structural patterns.

to be smaller compared to later years, due to fewer measurements. Since the overall (aggregated across all TAGS) number of communications sent across time did not have any systematic pattern indicative of events of historical importance, we decided to study the time series at a more granular level, by restricting to different types of TAGS. In Figure 1, panels (a)–(d) represent the communication streams, when restricted by TAGS type. Panels (a)–(c) show noticeable forms of increased activities in portions of the series – these are indicative of "interesting" historical events. For example, in panel (a) the increased activity in July 1974 corresponds to the Cyprus coup; in panel (b) the increase in number of diplomatic communications in April 1975 corresponds to the Fall of Saigon. Panel (c) shows multiple bursts in the number of cables, containing the particular TAGS UNGA (which stands for United Nations General Assembly) – interestingly, they all correspond to the annual United Nations General Assembly meetings. In addition to these visible bursts there seem to be some shorter periods of heightened activities, such as the smaller peaks for VS (South Vietnam) a year after the fall of Saigon corresponding to the ensuing refugee crisis.

In contrast to panels (a)–(c) in Figure 1, panel (d), for cables related to Finland (FI), does not seem to show any period of heightened activity during the time period under consideration. These prototypes are representative of the different TAGS-specific series: Exploratory analyses of the database of TAGS specific communication streams suggest that there are several series with some "interesting event" (as in Panels (a)–(c)), while others seem to be less active (as in panel (d)). A first goal of our work is to quantitatively define traits that separate communication streams like the figure in panel (d) from those in panels (a)–(c). We develop statistical methods that can *mine* these (TAGS specific) time-series and identify communication streams that exhibit statistically interesting activities in them. Once we identify these interesting communication streams, we develop algorithms that perform a deeper investigation of each series and identify time intervals where the signal undergoes abrupt localized changes in communication traffic. The informal ideas described above are made more precise in Section 2 of this paper.

Exploratory analysis suggests that changes in the proportion of a particular TAGS appearing in a communication stream are better representatives of identifying whether a period is active or not, as compared to tracking the corresponding counts. Due to the noticeable difference in the number of cables that were communicated over the weekdays and low-traffic days, the communication streams shown in Figure 1 seem to be a superposition of a high traffic series and a low-traffic (weekend and holiday) series. As a pre-processing step, we filtered out the days where the total number of cables being communicated were very small – they lead to more reliable estimates of proportions.

2 Statistical Methodology

For the reader's convenience, we first present an outline of the main statistical approaches that we discuss in this paper. Section 2.1 addresses our first question: how do we determine whether a communication stream, among several hundreds, is interesting or not? We address this as a statistical hypothesis testing problem, where we consider the problem of whether a communication stream is generated from a homogeneous binomial process – in other words: is the proportion of documents containing the particular TAGS uniform across time? We select a collection of TAGS for which there seems to be some form of a statistically interesting event, and explore their geometry in further detail. If p_t denotes the proportion of cables containing a TAGS at time t, we estimate the function $t \mapsto p_t$ with a piecewise constant signal – this is performed by using a regularized negative log-likelihood criterion based on the fused Lasso penalty [23] or its ℓ_0-counterpart [6, 15]. For efficient computation, we develop new fast algorithms for these penalties by adapting existing work for the least squares loss function to the case of the generalized linear model likelihood, studied herein – see Section 2.2. The piecewise constant segments lead to localized changes in the underlying signal $t \mapsto p_t$ – we use hypothesis testing ideas based on sample splitting [25] to rank-order the strengths of the jumps based on the associated p-values – see Section 2.3. The locations of discontinuities or jumps of the signal

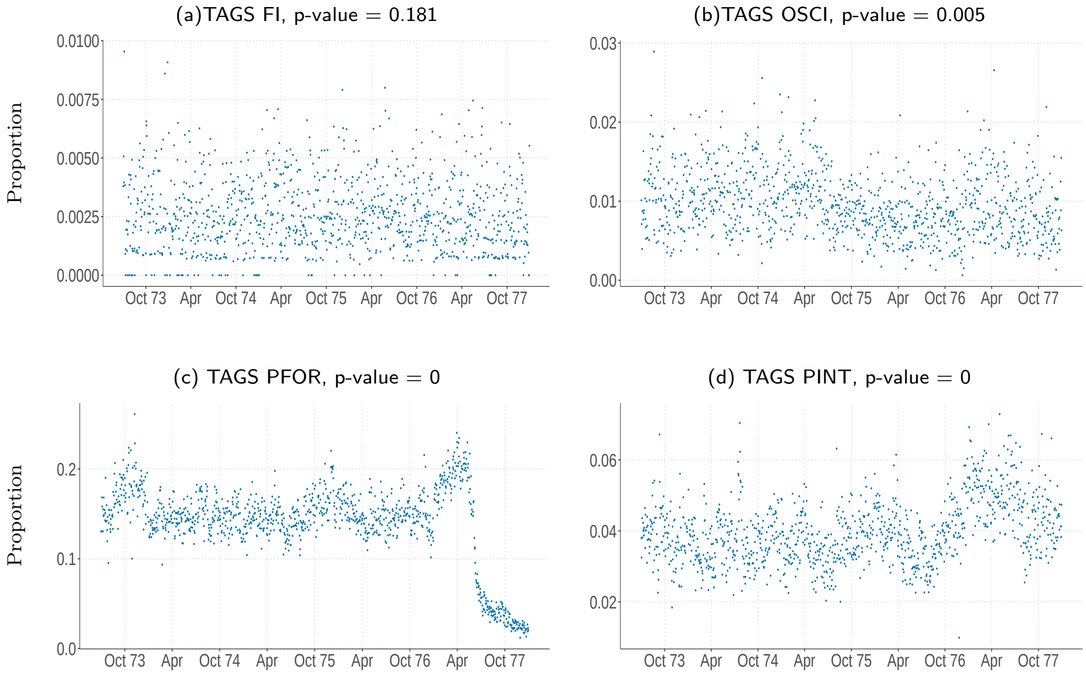


Figure 2: Communication streams with different significance scores in the spirit of Section 2.1: (a): relating to Finland, corresponds to a null model (b): relating to scientific grants, shows a weak deviation from the null model – perhaps due a slow decreasing trend in the series (c): relating to foreign policy – this generally shows significant deviation from the null model, which is due to sudden changes/jumps after Oct '76 (d): relating to internal political affairs, shows deviation from the null model but not due to a jump as prominent as (c) – this series seems to exhibit some systematic pattern of heightened activity after Oct '76, leading to a small p-value. The small p-values suggest the presence of a statistically interesting event in each series, and can be used to identify interesting communication streams – the p-value however, does not provide additional insights into the finer structural patterns of the streams. Additional examples can be found in Figure 10.

are aggregated to obtain locally contiguous subintervals of heightened diplomatic communications, which we call bursts, adopting the terminology coined by [17] in a different context (Section 2.4). Finally in Section 3, we discuss preliminary results on estimating the underlying signal with more flexible local models, beyond piecewise constant segments.

2.1 Identifying Interesting Communication Streams

We develop a framework to identify communication streams that exhibit some form of a heightened or in other words, statistically interesting activity. Consider a TAGS-specific series $(y_t, n_t), t = 1, \ldots, N$, where, y_t denotes the number of documents containing the specific TAGS among n_t cables, with proportion p_t. We will assume that the conditional distributions of $(y_t | n_t, p_t)$'s are all independent across t. We ask:

Is there any evidence of (localized) heightened intensity of the proportions, compared to a baseline model, where all proportions are the same?

To measure a localized change (increase) in intensity, we fix a window of size 2Δ and consider all the points in the Δ neighborhood of a time point i, given by: $\mathrm{N}(\Delta; i) = \{j : 1 \leq j \leq N, |j - i| \leq \Delta\}$ (we pick $\Delta = 5$ days in our experiment). The average proportion in this neighborhood: $p_i^{\text{ave}} := \sum_{j \in \mathrm{N}(\Delta;i)} n_j p_j / \sum_{j \in \mathrm{N}(\Delta;i)} n_j$, is a measure of communication traffic around the reference time point i. We say that a large value of p_i^{ave} compared to a baseline value p (for example, the global proportion), indicates the presence of an intense activity of some form; and declare such an occurrence to be statistically *interesting*. We hypothesize such an event to be associated with an event of historical interest; and subsequently validate this belief by factoring in the insights of a historian or social scientist.

Formally, we consider a simple testing problem with H_0: $p_t = p \ \forall t$ versus H_1 : there exists an i such that p_i^{ave} is larger than the (global) average proportion. Inspired by popularly used scan statistics [10], we propose to use the following test statistic:

$$\mathcal{T} = \max_t \ T_t \quad \text{where,} \quad T_t := (\widehat{p_t^{\text{ave}}} - \hat{p}_{H_0})/\hat{\sigma}_t, \tag{1}$$

where, $\hat{p}_{H_0}$ is the (estimated) global proportion of the signal estimated under the null hypothesis; $\widehat{p_t^{\text{ave}}}$ is an estimate of p_t^{ave}; and $\hat{\sigma}_t$ is the standard deviation of $\widehat{p_t^{\text{ave}}}$ evaluated under the null (H_0). $\mathcal{T}$ measures the strength of a locally contiguous period of heightened activity–the larger this value, the more pronounced is the localized traffic. We use a permutation based approach to compute the distribution of $\mathcal{T}$ under the null. Figure 2 shows different communication streams with their associated p-values computed using the framework described above; and varying levels of activity in the different representative communication streams – a large p-value for panel (a) (corresponding to TAGS FI) signifies a lack of interesting activity in this series – this aligns with an expert's understanding of historical events. A table providing a summary of how many cables survive different p-value thresholds are provided in Table 1 (Appenfix).

We also used a test-statistic depending upon the likelihood ratio test: we replaced T_t (in display (1)) by the local likelihood ratio derived from the binomial distribution; assuming that $(y_t|n_t, p_t)$ are independently distributed $\mathrm{Bin}(n_t, p_t)$ for $t = 1, \ldots, N$. The results obtained via this likelihood based approach was quite similar to that obtained from the model-free testing procedure described above; and are hence omitted for the sake of brevity.

2.2 Identifying Jumps in Communication Streams

The method in Section 2.1 simply identifies whether something statistically interesting, in the form of a heightened localized activity (for example), is occurring somewhere in a communication stream. It does not inform us about the structure or precise location of such an activity. We proceed to explore some finer structural properties of the series. Inspired by popularly used signal segmentation/estimation methods [23, 15, 20] in statistical signal processing, we seek to identify breaks or jumps in a piecewise constant approximation of the signal $t \mapsto p_t$ – these changes are more localized than the (global) deviations studied in Section 2.1.

We use the notation of Section 2.1. We assume herein that $(y_t|n_t, p_t) \overset{\text{ind}}{\sim} \text{Bin}(n_t, p_t)$, where, p_t denotes the probability of success. Our model assumes that the parameters p_t's are piecewise constant. Locations where the underlying signal $t \mapsto p_t$ exhibits a discontinuity will be called a "jump" in the communication stream. Instead of working directly with the proportion p_t's we will use the logistic-link function: $p_t = \exp(\theta_t)/(1 + \exp(\theta_t))$ and will treat θ_t as a natural parameter. This leads to the following regularized estimation problem:

$$\underset{\theta_t, 1 \le t \le N}{\text{minimize}} \quad \phi(\boldsymbol{\theta}) := \sum_{t=1}^{N} \Big(-y_t\theta_t + n_t \log(1 + \exp(\theta_t))\Big) + \lambda H(\boldsymbol{\theta}), \tag{2}$$

where, $\mathcal{L}(\boldsymbol{\theta}) := \sum_{t=1}^{N} (-y_t\theta_t + n_t \log(1 + \exp(\theta_t)))$, the negative logarithm of the likelihood is the data-fidelity term and $H(\boldsymbol{\theta})$ is the regularizer. $H(\boldsymbol{\theta})$ encourages the estimated θ_t's (and hence the proportion p_t's) to be piecewise constant and the regularization parameter $\lambda > 0$ controls the amount of shrinkage. Two common examples of $H(\boldsymbol{\theta})$ that we use herein are [23, 20, 6, 15]:

- ℓ_1-segmentation (Fused Lasso): $H(\boldsymbol{\theta}) = H_{\ell_1}(\boldsymbol{\theta}) = \sum_{t=1}^{N-1} |\theta_{t+1} - \theta_t|$, which penalizes the total variation of a signal, which may also be thought as a soft-version of the number of the number of jumps in $\theta_t, t \ge 1$.

- ℓ_0-segmentation: Here we take $H(\boldsymbol{\theta}) = H_{\ell_0}(\boldsymbol{\theta}) = \sum_{t=1}^{N-1} \mathbf{1}(\theta_{t+1} \ne \theta_t)$, which penalizes the number of jumps in the signal θ_t.

We assume above and in the discussion below, that the time points are equally spaced. If they are not equispaced, the penalty function needs to be adjusted in a straightforward fashion as discussed in Section A.1.

For the ℓ_1 penalty function $H_{\ell_1}(\boldsymbol{\theta})$, Problem (2) is a convex optimization problem. The ℓ_1-penalty on the successive differences in $\boldsymbol{\theta}$, is commonly referred to as the fused lasso or total-variation penalty [23, 20]. This semi-norm induces shrinkage along with sparsity on the coefficient differences $\theta_{t+1} - \theta_t$. However, the ℓ_1-based penalty often over-estimates the number of jumps when the tuning parameter is chosen so as to obtain a model with good data-fidelity – this is especially true if the tuning parameter is chosen based on a held out test set to minimize test error. This is due to the shrinkage effect of the ℓ_1-penalty, which severely penalizes large values of the jumps $\theta_{t+1} - \theta_t$. To obtain a model with fewer jumps, the regularization parameter needs to be made larger – this however, may lead to a model where some of the important jumps are missed. Many of these limitations can be overcome by using a ℓ_0-based penalty [6] which directly penalizes the number of jumps and is less agnostic to the precise value of the jump. The caveat however, is that the resulting optimization problem becomes non-convex and discrete optimization methods are required to obtain the global minimum of such problems [6, 15]. Developing global optimization algorithms for Problem (2) (for the penalty $H_{\ell_0}(\boldsymbol{\theta})$) along the lines of [21, 3] is beyond the scope of the current paper. Herein, we describe fast and robust algorithms to obtain good quality solutions to Problem (2). Our proposal is motivated by first-order optimization based algorithms [22] pioneered in the convex optimization community. Loosely speaking, these are iterative methods that can be used to obtain high quality approximate solutions for convex optimization tasks, compared to off-the-shelf interior point methods that are difficult to scale to large problems. These methods apply to both $H_{\ell_1}(\boldsymbol{\theta})$ and $H_{\ell_0}(\boldsymbol{\theta})$ – however, there are certain subtleties as we describe below.

2.2.1 Model Fitting: Optimization Algorithm

Problem (2) is of the composite form, i.e., the objective function can be written as the sum of a smooth convex function $\mathcal{L}(\boldsymbol{\theta})$ and a non-smooth penalty function $H(\boldsymbol{\theta})$. We will apply proximal gradient descent methods [1] for this problem. The negative log-likelihood function $\mathcal{L}(\boldsymbol{\theta})$ is continuously differentiable and satisfies:

$$\|\nabla\mathcal{L}(\mathbf{u}) - \nabla\mathcal{L}(\mathbf{v})\| \leq \ell\|\mathbf{u} - \mathbf{v}\|, \qquad \forall \mathbf{u}, \mathbf{v}; \tag{3}$$

with $\ell = \frac{1}{4}\max_{i=1} y_i$. This follows by noting that the ith coordinate of $\nabla\mathcal{L}(\mathbf{u})$ is: $\{\nabla\mathcal{L}(\mathbf{u})\}_i = -y_i + n_i \exp(u_i)/(1+\exp(u_i))$ and $\nabla^2\mathcal{L}(\mathbf{u})$ is a diagonal matrix with the ith diagonal entry being:

$$\left\{\nabla^2\mathcal{L}(\mathbf{u})\right\}_{ii} = n_i \exp(u_i)/(1+\exp(u_i))^2 \leq \frac{1}{4}n_i, \quad i = 1, \ldots, N. \tag{4}$$

Hence, the largest eigenvalue of $\nabla^2\mathcal{L}(\mathbf{u})$, i.e., $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{u})) \leq \frac{1}{4}\max_{i=1}^N n_i$, which justifies the choice of ℓ, as above. For a fixed $L \geq \ell$, our algorithm performs the following updates (for all $k \geq 1$):

$$\boldsymbol{\theta}_{k+1} \in \arg\min_{\boldsymbol{\theta}} \ \frac{L}{2}\left\|\boldsymbol{\theta} - \left(\boldsymbol{\theta}_k - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right)\right\|_2^2 + H(\boldsymbol{\theta}). \tag{5}$$

This leads to a decreasing sequence of objective values $\phi(\boldsymbol{\theta}_{k+1}) \leq \phi(\boldsymbol{\theta}_k)$ for $k \geq 1$. We now study the fate of the sequence $\boldsymbol{\theta}_k$, depending upon the choice of $H(\boldsymbol{\theta})$.

The fused lasso penalty ($H_{\ell_1}(\boldsymbol{\theta})$) We first consider the case where the regularizer and hence the function $\phi(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. The sequence $\boldsymbol{\theta}_k$ converges to a minimum of Problem (2) with the penalty function $H_{\ell_1}(\boldsymbol{\theta})$. More precisely, it can be shown that [1]

$$\phi(\boldsymbol{\theta}_k) - \phi(\boldsymbol{\theta}^*) \leq \frac{L}{2k}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_2^2, \tag{6}$$

where, $\boldsymbol{\theta}^*$ is an optimal solution to Problem (2). Display (6) states that the sequence $\phi(\boldsymbol{\theta}_k)$ converges to a minimizer of Problem (2) with a rate of $O(\frac{1}{k})$. In fact, this rate can be improved further under a minor additional assumption. Let $\mu_k = \min_{i=1,\ldots,N} \ \left\{\nabla^2\mathcal{L}(\boldsymbol{\theta}_k)\right\}_{ii}$, i.e., the smallest diagonal entry of the Hessian of $\mathcal{L}(\boldsymbol{\theta})$. As long as $\boldsymbol{\theta}_k$'s are uniformly bounded and $\min_i \ n_i > 0$ then $\mu := \min_k \mu_k > 0$. The rate of convergence is linear; and is given by:

$$\phi(\boldsymbol{\theta}_k) - \phi(\boldsymbol{\theta}^*) \leq \gamma^k \left(\phi(\boldsymbol{\theta}_0) - \phi(\boldsymbol{\theta}^*)\right), \tag{7}$$

where,

$$\gamma = \begin{cases} \frac{L}{\mu} & \text{if } \frac{\mu}{L} \geq 2 \\ (1 - \frac{\mu}{4L}) & \text{otherwise.} \end{cases} \tag{8}$$

Note that sub-problem (5) requires solving a problem of the form:

$$\underset{\mathbf{u} \in \Re^N}{\text{minimize}} \;\; \frac{1}{2}\|\mathbf{u} - \bar{\mathbf{u}}\|_2^2 + \lambda' \sum_{i=1}^{N-1} |u_{i+1} - u_i|, \tag{9}$$

which can be solved very efficiently via Dynamic Programming [15] with cost $O(N)$ for $\lambda' > 0$ – one can solve instances with N a few thousands in a few milliseconds.

The ℓ_0-segmentation penalty $(H_{\ell_0}(\boldsymbol{\theta}))$ The algorithm above, also applies to the regularizer $H_{\ell_0}(\boldsymbol{\theta})$. In update (5) we set $H(\theta)$ to $H_{\ell_0}(\boldsymbol{\theta})$. This requires solving

$$\underset{\mathbf{u} \in \Re^N}{\text{minimize}} \;\; \frac{1}{2}\|\mathbf{u} - \bar{\mathbf{u}}\|_2^2 + \lambda' \sum_{i=1}^{N-1} \mathbf{1}(u_{i+1} \neq u_i), \tag{10}$$

which can also be computed efficiently using the dynamic programming algorithm of [15]. Describing the properties of this sequence $\boldsymbol{\theta}_k, k \geq 1$ is subtle since the associated optimization problem (2) is non-convex. Following [3] it can be shown that the sequence $\phi(\boldsymbol{\theta}_k)$ is decreasing, bounded below[2] and it converges to ϕ^* (say), which may not be a global minimum. We say that $\tilde{\boldsymbol{\theta}}$ is a first-order stationary point for Problem (2) if it satisfies the following fixed point equation:

$$\tilde{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}}{\arg\min} \;\; \frac{L}{2}\left\|\boldsymbol{\theta} - \left(\tilde{\boldsymbol{\theta}} - \frac{1}{L}\nabla\mathcal{L}(\tilde{\boldsymbol{\theta}})\right)\right\|_2^2 + H_{\ell_0}(\boldsymbol{\theta}).$$

$\boldsymbol{\theta}_k$ is said to be an ϵ-accurate first order stationary point, if $\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \leq \epsilon$. Following the convergence analysis in [3][Theorem 3.1], we obtain the following finite-time convergence rate of $\boldsymbol{\theta}_k$ to a first order stationary point:

$$\min_{1 \leq k \leq K} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \leq \frac{2(\phi(\boldsymbol{\theta}_1) - \phi^*)}{K(L - \ell)}. \tag{11}$$

Note that this algorithm may not reach the global minimum of the ℓ_0 version of Problem (2). However, in practice, it reaches a high quality solution quite fast.

2.2.2 Estimated Signal

To gather some intuition about the behavior of the segmentation methods described above, we consider a synthetic example in Figure 3 – here the underlying (true) signal is piecewise constant with three levels up to time point t_0, there is a right discontinuity at t_0 after which it becomes linear [3]. The figure presents the signal

[2] We will assume that the function $\phi(\boldsymbol{\theta})$ is bounded below and the minimum exists.

[3] More specifically, data is generated by $y_t \overset{\text{ind}}{\sim} \text{Bin}(n_t = 200, p_t), t = 1, ..., 1203$, where $p_t = 0.5$ for $1 \leq t \leq 200$; $p_t = 0.6$ for $201 \leq t \leq 500$; $p_t = 0.8$ for $501 \leq t \leq 550$; $p_t = 0.55 + (t - 550)/3000$ for $551 \leq t \leq 1203$.

estimates (for both the ℓ_0 and ℓ_1 penalties) at the cross-validated choices of the tuning parameter – we use k-fold cross validation [12] which is also used in the R package `genlasso` (Since we want to ensure each fold is representative of the time series, instead of randomly assigning points to a fold, we systematically assign points by placing every kth point into the same fold). For both segmentation schemes, the estimated signals $\{\hat{p}_t\}$ serve as good (overall) approximations of $\{p_t\}$ – however, there are some subtle differences. First of all, the ℓ_1-segmentation scheme leads to biased estimates and the bias becomes quite prominent in estimating the jump at the centre of the signal. This behavior is not present for the ℓ_0-scheme. In addition, the estimates for the linear component (at the right) also differ across the ℓ_0 and ℓ_1 schemes. The ℓ_0 regularizer leads to a fewer number of segments (here three) compared to the ℓ_1-penalty which has several smaller jumps.

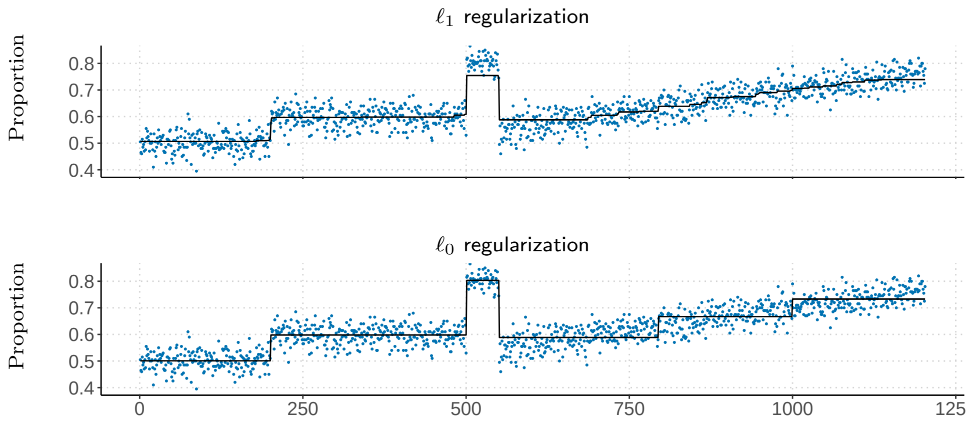


Figure 3: Estimators obtained from Problem (2) with ℓ_1 (upper panel) and ℓ_0 (lower panel) regularization. The data is synthetic and the underlying signal contains two sharp jumps and a gradual increasing trend. We use cross validation to select a value of λ. ℓ_1 penalty shrinks the estimated probability during a big burst ($501 \leq t \leq 550$) and gives more jumps during the gradual increase period ($551 \leq t \leq 1203$).

Figures 4 and 5 show the estimated signal proportions $\hat{p}_t = \exp(\hat{\theta}_t)/(1+\exp(\hat{\theta}_t))$, where the $\hat{\theta}_t$'s are obtained from Problem (2). Both the penalty functions do a good job in estimating a piecewise constant version of the underlying signal – the ℓ_0 scheme leads to fewer jumps than its ℓ_1 counterpart, for a comparable data-fidelity. The figures also show fitted signals for a few other values of λ around the cross-validated choice at the center (λ increases as one moves down the rows): we include the tuning parameter selected by the one-standard error rule [12] (see also the R package `genlasso`). We can see that as λ decreases, the algorithm captures a more granular structure of the data and estimates more jumps.

2.3 A deeper investigation of Jumps

2.3.1 How intense is a Jump?

The procedure in Section 2.2 provides an estimated piecewise constant signal $\{\hat{p}_t\}_{t\geq 1}$. In particular, this can be used to identify locations where the signal changes: $\hat{p}_{t+1} \neq$

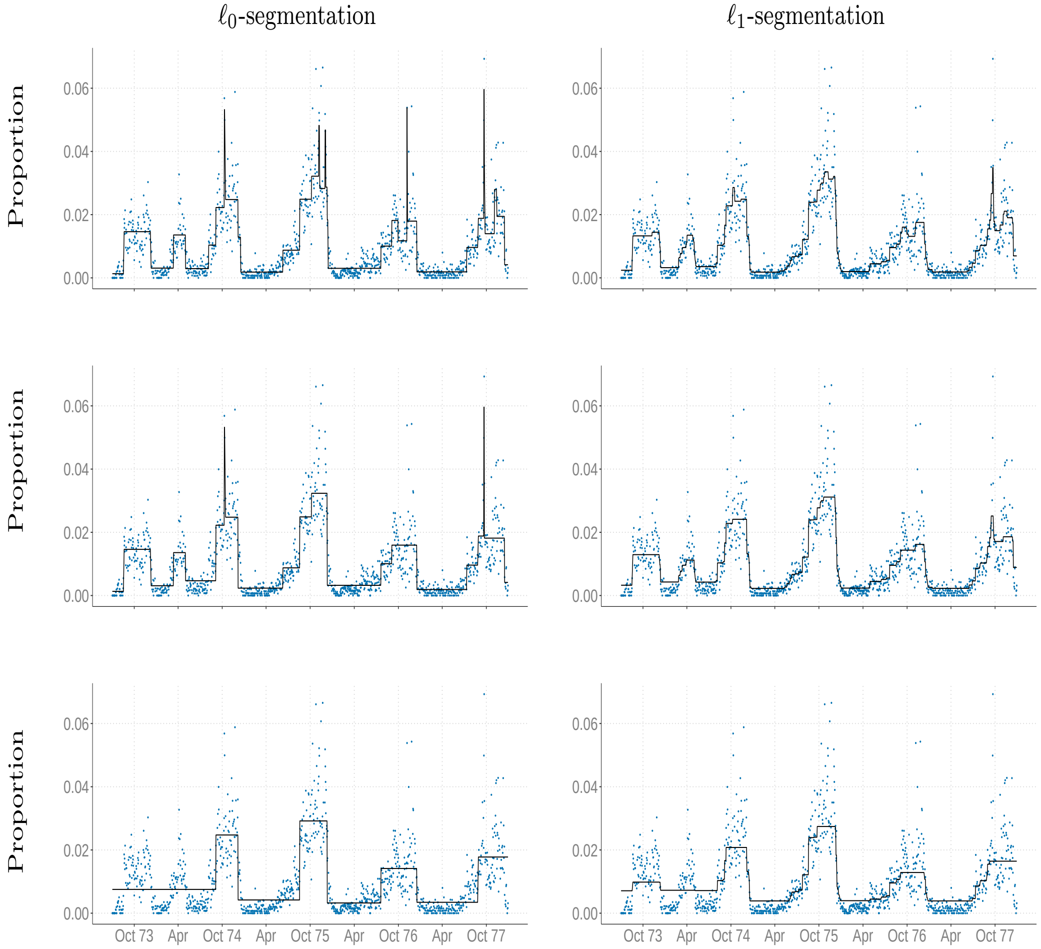


Figure 4: Figure showing the raw proportions (in blue dots) for TAGS UNGA (U.N. General Assembly) and the estimated proportions $\hat{p}_t, t \geq 1$, as obtained from the regularization framework in Problem (2). The left panel shows the estimates obtained with the ℓ_0-segmentation penalty and the right panel shows the estimates with the ℓ_1-segmentation penalty. The middle rows correspond to the optimal λ chosen by 10-fold cross-validation. It shows how, in between the cyclical jumps in UN-related communications relating to the regular Fall meetings of the General Assembly, there was also a jump in April-May 1974. This occurred when Algeria called a special session to demand UN support for a "New International Economic Order." We show a few additional choices of the regularization parameter for each example (see text for details). The figure (bottom right panel) shows the large bias incurred by the ℓ_1-penalization method in the estimation process – this behavior is less pronounced with the ℓ_0 penalty. The ℓ_1-penalty also exhibits a stair-casing effect – with many small jumps. Unlike the ℓ_1-penalty, the ℓ_0-penalty selects some segments that are *very* short, these segments disappear upon increasing the penalty parameter.

$\hat{p}_t$. Of course, this jump depends upon the choice of the tuning parameter, the penalty used and also the underlying signal. A jump estimated by the ℓ_0 or ℓ_1-segmentation procedure may reflect (a) a discontinuity in the signal – in this case, the signal is well approximated by locally constant segments with pieces adapting

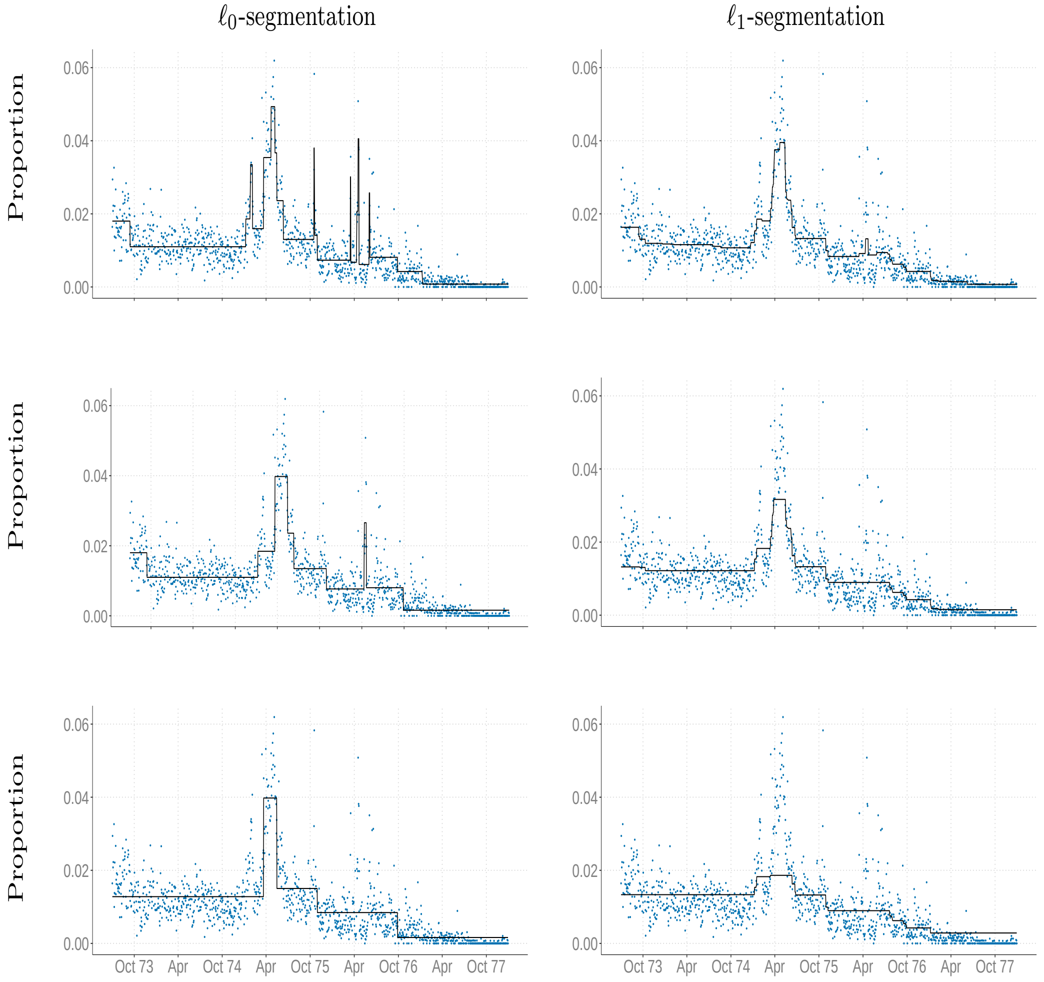


Figure 5: Figure showing the raw proportions (in blue dots) for TAGS VS (South Vietnam) and the estimated proportions $\hat{p}_t, t \geq 1$, as obtained from the regularization framework in Problem (2). The left panel shows the results for the ℓ_0-segmentation penalty and the right panel the ℓ_1-penalty. The middle rows correspond to the optimal λ chosen by cross-validation, and we show a few additional choices of the regularization parameter for each example. A practitioner might select one or another depending on whether they would want to identify smaller jumps that correspond, in this case, to the refugee crisis that followed the defeat of South Vietnam.

to the data (b) a localized trend in the signal, as we saw in Figure 3 – a jump here is a consequence of the slope of $t \mapsto p_t$ and not a discontinuity in the signal $t \mapsto p_t$.

Given an estimate of $\{\hat{p}_t\}$ a scholar accustomed to analyzing events through a close reading of historical documents may ask:

- Which of these jumps might be important or are indicative of a historical event of interest?
- Can one obtain a rank ordering of the jumps based on their intensities?

We formalize this question as follows: given an estimate of $\{\hat{p}_t\}$ and a set of candidate jumps, can we identify jumps that are *strong enough*? Ideally, we would like a simple measure that associates a score to the strength and size of a jump selected by the estimation procedure – this would help us select a smaller set of jumps that merit closer scrutiny. Towards this end, we use a sample splitting[4] procedure [25]: a subsample of size 50% of the data is used for estimating the location of the jumps and the remaining held out part of the data is used to associate a p-value score (the method is described below) to each jump identified in the first stage. This method is simple, intuitive and provides a natural method to rank the jumps in a communication stream. Using this scheme, one can potentially reduce the number of jumps selected by the fitting procedure by screening out jumps with p-values larger than a user-defined pre-specified threshold.

We describe our approach with reference to the ℓ_0-penalization procedure, though the idea will also apply for the ℓ_1-penalized estimate. Suppose a candidate location for the change point $\hat{t}$ is estimated based on the first part of the sample (used for estimating the signal). The test statistic is evaluated on the held out sample. We take a neighborhood of size 2Δ centered at $\hat{t}$, and denote the time points on the left of $\hat{t}$ as $L(\hat{t}, \Delta)$ and those on the right of $\hat{t}$ as $R(\hat{t}, \Delta)$. Let us assume that p_t for $t \in L(\hat{t}, \Delta)$ are all equal to $p(L, \hat{t})$; and p_t for $t \in R(\hat{t}, \Delta)$ are all equal to $p(R, \hat{t})$. We then test the null hypothesis (H_0) that the proportions on the left and right parts of $\hat{t}$ are equal: $p(L, \hat{t}) = p(R, \hat{t})$; versus the alternative (H_1) that $p(L, \hat{t}) \neq p(R, \hat{t})$. We use the likelihood ratio test statistic for this purpose. To compute the null distribution, we used a two step procedure. We first identified segments of the time series which did not overlap with any candidate change point location (i.e., parts of the series where the estimated signal was constant for stretches of size at least $2\Delta + 1$) of the time series. Based on the regions thus selected (i.e., the locally constant stretches of the signal), we simulated the null distribution of the test statistic by using a permutation test. These 2-sided p-values were consequently used as a measure of intensity of every jump.

Note that a candidate jump estimated by the signal estimation procedure at the cross-validated choice of the tuning parameter need not necessarily correspond to a jump with a low p-value. A small p-value indicates that the intervals to the left and right of $\hat{t}$ have different proportions[5]. Thus the p-values can be used to (a) devise a scoring mechanism to rank order multiple jumps observed in a series and/or (b) prune out redundant jumps and identify ones that exhibit a strong difference in proportions between the left and right intervals. Figure 6 shows the communication stream for TAGS CVIS and the estimated signal obtained via ℓ_0-segmentation. We also computed the p-values for each potential jump location as suggested via the ℓ_0-segmentation fit. We pruned them and refitted the model based on their p-value scores. It helps to interpret these patterns alongside Figure 9 (TAGS CVIS) which

[4] Due to the large number of samples, sample splitting does not significantly reduce the size of the training dataset.

[5] A jump obtained from the estimated $\{\hat{p}_t\}$ may be due to a linear rise in the signal which need not correspond to a significant change in local proportions. Our experiments indicate that jumps in $\{\hat{p}_t\}$ that correspond to gradual linear rises in the signal, have higher p-values associated with them when compared to sudden or abrupt changes in $\{\hat{p}_t\}$

presents more flexible fits of the underlying signals for this communication stream. Figure 7 shows additional examples interpreting the p-values associated with the different jumps. Figure 7 suggests that the p-values are indicative of whether a jump is due to a shift in the piecewise constant level or a linear trend – the p-values are larger when there is a linear trend rather than a sharp jump (as in a piecewise constant signal). Figure 9 shows a more flexible approximation of this communication stream which provides further insights into the patterns of TAGS US series. To validate the intuition gathered above, we consider a synthetic example in Figure 8 (with the same data as in Figure 3) – here we observe that the p-values tend to be larger for jumps in the right part of the signal – these jumps in the piecewise constant segments result from estimating a linear trend (that appears at the right of the series) with piecewise constant segments. Note that the p-values associated with the first three jumps (at the left of the signal) are quite small – they correspond to jumps in the underlying piecewise constant signal.

In passing we note that it is also possible to perform multiple testing procedures [19] to attach error rates to a family of jumps. In particular, we can use Family Wise Error Rates or False Discovery Rate (FDR) control procedures to return a collection of candidate jumps with a certain prescribed control on the error rate [19]. However, our goal here is to use the sample splitting and p-value framework to perform an exploratory analysis of the strengths of the different jumps – we do not pursue an in-depth study of multiple testing in this work.

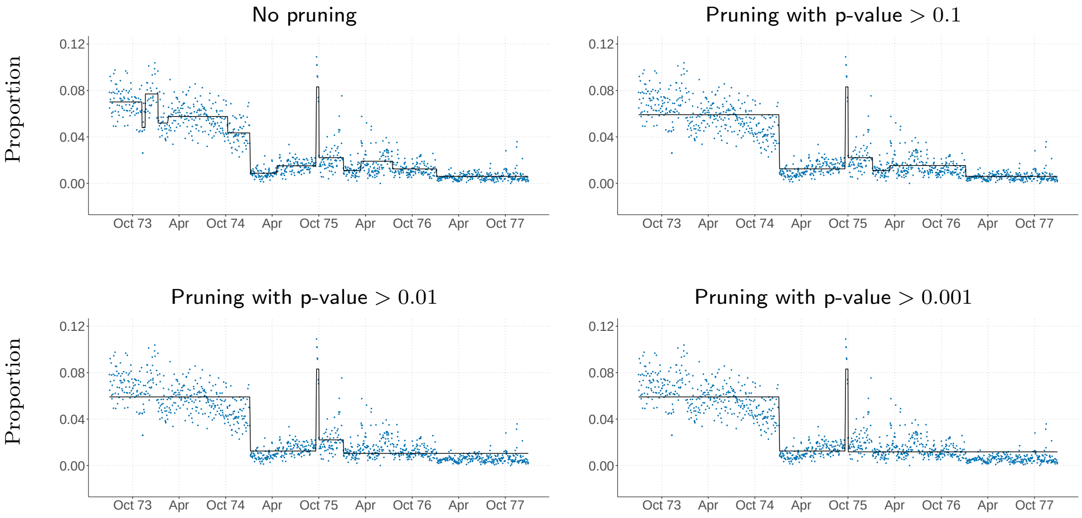


Figure 6: Figure showing the communication stream for TAGS CVIS – the estimated signal is obtained from the ℓ_0-segmentation scheme (at the cross-validated choice of λ). We compute the p-values (based on sample splitting, as described in Section 2.3.1) for every candidate jump location and prune the jump locations (and refit the signal with the new jump locations) based on different thresholds. We observe that a pruning rule based on p-values leads to a fairly robust selection of intense jump locations, where each location corresponds to a sharp change in local means.

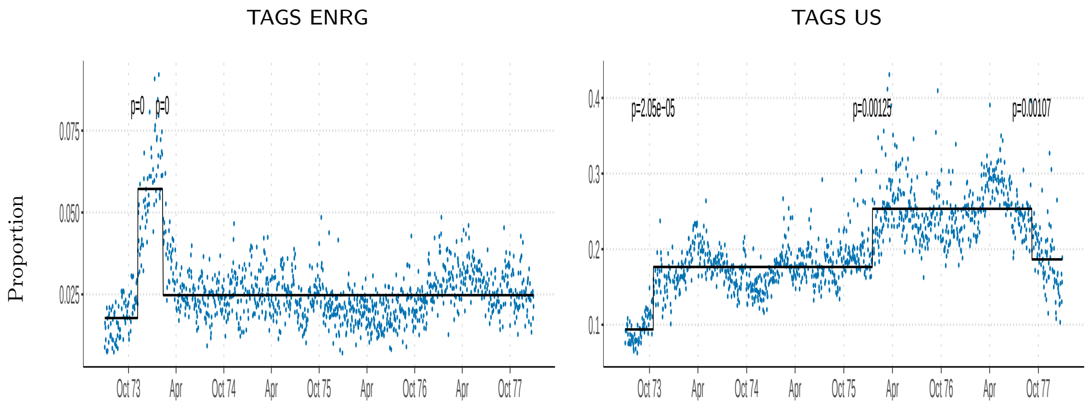


Figure 7: TAGS ENRG (energy) and US (for cables relating to the U.S.) with p-values associated with the estimated jumps (using the framework in Section 2.3.1). The jumps for ENRG are much sharper and indicate rapid (though not instantaneous) changes in mean, starting with the 1973 OPEC oil embargo. This gives them low p-values. In contrast, the two jumps to the right of the signal for US are during less rapid changes in mean and thus have slightly larger p-values ($\sim 10^{-3}$). Figure 9 presents more flexible fits of the underlying signal. (The notation $p = x$ is a shorthand for p-value being equal to x.)

2.4 From Jumps to Bursts

Section 2.2 presents methods to detect several jumps in a signal, and also presents an in-depth investigation of how to associate scores (p-values) to each of the detected jumps using a sample splitting scheme. We now explore if it is possible to summarize a single communication stream (corresponding to a specific TAGS) with a score – a measure that should ideally contain in it information pertaining to the strengths and stretches of the different jumps. Towards this end, following the terminology introduced by Kleinberg [17] we formalize the notion of a "burst". The general approach pursued in this paper and the models used differ from [17]. Informally speaking, a burst corresponds to a stretch of time where a communication stream depicts traffic which is larger than a baseline value. We present a computational scheme to estimate such bursts for a TAGS -specific communication stream.

2.4.1 Computation of the strength of a Burst

As a starting point, we consider an estimate of a baseline proportion p_0 (we discuss how to compute this below) that is specific to a communication stream. A "burstiness period" or simply burst corresponds to a time interval where the estimated signal lies above the baseline value p_0 and is given by $\mathcal{T} = [t_{\text{start}}, t_{\text{end}}]$, where $\hat{p}_t > p_0, \forall t \in \mathcal{T}$. Following [17], we define the strength $S(\mathcal{T})$ of the burst as the logarithm of the likelihood ratio (here, the numerator is the likelihood of the signal and the denominator is that evaluated at the baseline) given by: $S(\mathcal{T}) = \sum_{t \in \mathcal{T}} \left(\log L(\hat{p}_t | n_t, y_t) - \log L(p_0 | n_t, y_t) \right)$, where $L(\hat{p}_t | n_t, y_t)$ denotes the likelihood at time t. Note that the baseline p_0 is specific to a communication stream and the score $S(\mathcal{T})$ represents a deviation from this global baseline. $S(\mathcal{T})$ is different than the magnitude of a jump given by $\hat{p}_{t+1} - \hat{p}_t$ – it takes into account the deviation

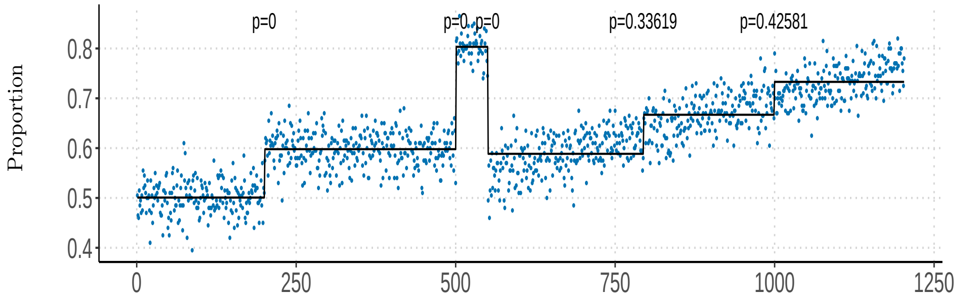


Figure 8: Synthetic data: data description is the same as Figure 3, which contains three real jumps and a linear increasing trend. The first three detected jumps (from left to right) have small p-values (close to zero) – they correctly correspond to the jumps in the underlying signal. The other two potential jumps have p-values ~ 0.33 and ~ 0.42 respectively – these jumps are a consequence of the linear trend (we use the framework in Section 2.3.1). (The notation $p = x$ is a shorthand for p-value being equal to x.)

of $\hat{p}_t$ from the baseline p_0 as well as the duration of the burst given by the length of $\mathcal{T}$. A large value of $S(\mathcal{T})$ means that a large part of the likelihood is explained by deviations from the baseline, and therefore, corresponds to a strong burst. Note that each TAGS can have multiple bursts leading to multiple intervals $\mathcal{T}$ – each with an assigned strength $S(\mathcal{T})$.

Choice of baseline: The baseline value p_0 should be representative of the behavior of the TAG-specific communication stream. The global proportion of a communication stream is a reasonable choice. We set p_0 to be one standard deviation larger than the global proportion

$$p_0 = \bar{p} + \sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}, \quad \text{where,} \;\; \bar{p} = \frac{\sum_{t=1}^{N} y_t}{\sum_{t=1}^{N} n_t}, \quad \bar{n} = \frac{1}{N}\sum_{t=1}^{N} n_t.$$

A robust estimate like the median can also be used instead of the average. In our experiments we found that the top-ranked slots were relatively agnostic to the choice of the baseline p_0.

2.4.2 Interpretation of Bursts

Table 2 presents the top thirty bursts, with the start and end dates, and the date with the highest value. A close study of the content of the cables shows that not all of these bursts correspond with what scholars would recognize as an even of historical importance. After all, the cable TAGS that diplomats used do not necessarily correspond with diplomatic activity. For instance, the second biggest burst is made up of cables related to transportation (ETRN) a TAGS that was commonly used, and overused, from when we begin to have records continuing until 1974, when diplomats' use of this TAGS was largely discontinued. The biggest burst, for CVIS (visas), has a similar pattern (as shown in Figure 6). But in this case, it appears to reflect a decision by archivists to stop preserving records related to visas [18].

To the model, both of these look like bursts, but they simply reflect administrative procedures rather than historical events.

The bursts that follow, on the other hand, appear to correspond well to historical events. The next ten include the Carter administration's prioritization of human rights (SHUM), Anwar Sadat's surprise visit to Israel (PGOV), the Southeast Asian "Boat People" crisis (SREF), the U.S. withdrawal from the International Labor Organization (PORG), the conclusion of the Panama Canal Treaty (PDIP), the 1973 Yom Kippur War (XF, for Middle East), Portugal's withdrawal from Angola (AO), and the 1974 crisis over Cyprus (CY). All are included in each of the four standard reference works we consulted [7, 9, 14, 8].

A systematic evaluation of hundreds of bursts for historical significance lies outside the scope of this paper. But the relative proportion of recognized historical events appears to diminish as one examines smaller bursts, like the ones ranked in the range 13-22. They include the denouement of the Vietnamese War (VM and VS), the OPEC oil embargo (ENRG), the Vladivostok summit (OVIP), and negotiations to end white rule in Rhodesia (RH). But there are also largely unrecognized events, like a 1975 UN General Assembly debate over the command of foreign military forces in South Korea, that would appear to merit closer scrutiny. The identification of such unstudied episodes, no less than rank-ordering well-known events, is valuable for historical scholarship.

3 Beyond Piecewise Constant Segments

A major focus of Section 2 was on approximating a communication stream with a piecewise constant signal. This framework does help us answer some of the major data-driven questions of interest to a political scientist, based on a first order approximation of the communication streams. We now investigate more flexible signal approximations that provide us insights into the finer behavior of the signals. A natural extension of a piecewise constant estimate $\{\hat{p}_t\}$ is a piecewise linear estimate. However, there are subtleties in incorporating this structure into our framework, as we discuss below.

To settle ideas, let us consider the usual signal denoising problem with data: $\tilde{y}_i = \mu_i + \epsilon_i$, for $i = 1, \ldots, N$ where, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We seek to estimate $\boldsymbol{\mu}$ such that it is piecewise linear. In this vein, it is common to use the ℓ_1 trend-filtering approach [16, 24] with regularizer $H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu}) = \sum_t |\mu_{t+2} - 2\mu_{t+1} + \mu_t|$ to obtain a signal with piecewise linear segments:

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{N} (\tilde{y}_i - \mu_i)^2 + \lambda H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu}).$$

The penalty function $H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu})$ encodes the ℓ_1-norm on the discrete second order derivative of the signal $\{\mu_t\}$ assuming that the time points are all equally spaced. $H^{\text{tf}}_{\ell_1}(\boldsymbol{\mu})$ can be interpreted as a convexification of its ℓ_0 version: $H^{\text{tf}}_{\ell_0}(\boldsymbol{\mu}) = \sum_t \mathbf{1}(\mu_{t+2} - 2\mu_{t+1} + \mu_t \neq 0)$ that counts the number of different piecewise linear segments.

Our situation is different from the denoising example outlined above. Since we are working under the modeling assumption: $(y_t|n_t, p_t) \sim \text{Bin}(n_t, p_t)$ with $p_t =$

$\exp(\theta_t)/(1+\exp(\theta_t))$, imposing a trend filtering penalty on p_t so as to maintain piecewise linearity will lead to a non-convex optimization problem due to the nonlinear dependence of p_t on θ_t. Thus, instead of enforcing the sequence $t \mapsto p_t$ to be piecewise linear, we allow the latent parameters $t \mapsto \theta_t$ to be piecewise linear – this leads to a computationally tractable estimation framework. Encouraging $t \mapsto \theta_t$ to be piecewise linear enables $t \mapsto p_t$ to be more flexible than a piecewise constant signal. Towards this end, we propose a simple adaption of the estimation criterion in Problem (2) by setting the regularizer $H(\boldsymbol{\theta}) = H^{\text{tf}}_{\ell_1}(\boldsymbol{\theta})$. Figure 9 shows the results of estimates obtained from some communication streams using the ℓ_1-trend filtering penalty. If the time points are not equally spaced, then this penalty can be modified appropriately – see for example [16] and also Section A.1.

Computation The proximal gradient-stylized algorithm update (5) can be adapted to the setting described above with $H(\boldsymbol{\theta}) = H^{\text{tf}}_{\ell_1}(\boldsymbol{\theta})$. We use the specialized interior point solver[6] of [16] – this works quite nicely for the problem sizes encountered in this paper. The resulting Problem (2) is convex and the sequence (5) leads to the optimum of the problem. The convergence rates outlined in (6) and (7) will also apply to this problem. If we set $H(\boldsymbol{\theta}) = H^{\text{tf}}_{\ell_0}(\boldsymbol{\theta})$, the resulting Problem (2) becomes a challenging nonconvex optimization problem – in this case, there is no analogue of the highly efficient dynamic programming implementation that was available for $H_{\ell_0}(\boldsymbol{\theta})$. Thus, for the case where we desire $t \mapsto \theta_t$ to be piecewise linear, we confine our study to the choice of the convex ℓ_1-trend filtering regularizer.

Acknowledgements

R. Mazumder was supported by ONR (grant # N000141512342) and NSF-IIS (grant # 1718258) for support. The authors will like to thank: David Blei, David Madigan, Shawn Simpson for helpful comments and encouragement. The authors will also like to thank the workshop participants of "Famine and Feast–International Historical Research in the Digital Age" (London, UK; 2015) for comments on the work. Preliminary results from this article appeared in a media article on `buzzfeed.com` [7].

A Appendix

A.1 Problem (5) with irregularly spaced time points

If the time points are irregularly spaced then the penalty functions need to be adjusted accordingly. The fused lasso penalty function becomes: $H(\boldsymbol{\theta}) = \sum_i |\theta_{t+1} - \theta_t|/\Delta_t$ where, Δ_t denotes the time difference between time point t and the next time point, indexed by $t+1$. For this choice, the associated proximal map i.e., Problem (5) needs to be modified to:

$$\underset{\mathbf{u}\in\Re^N}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{u}-\bar{\mathbf{u}}\|_2^2 + \lambda'\|D\mathbf{u}\|_1, \tag{12}$$

[6] We use the R-package wrapper available from https://github.com/hadley/l1tf

[7] Article www.buzzfeed.com/josephbernstein/can-a-computer-algorithm-do-the-job-of-a-historian?

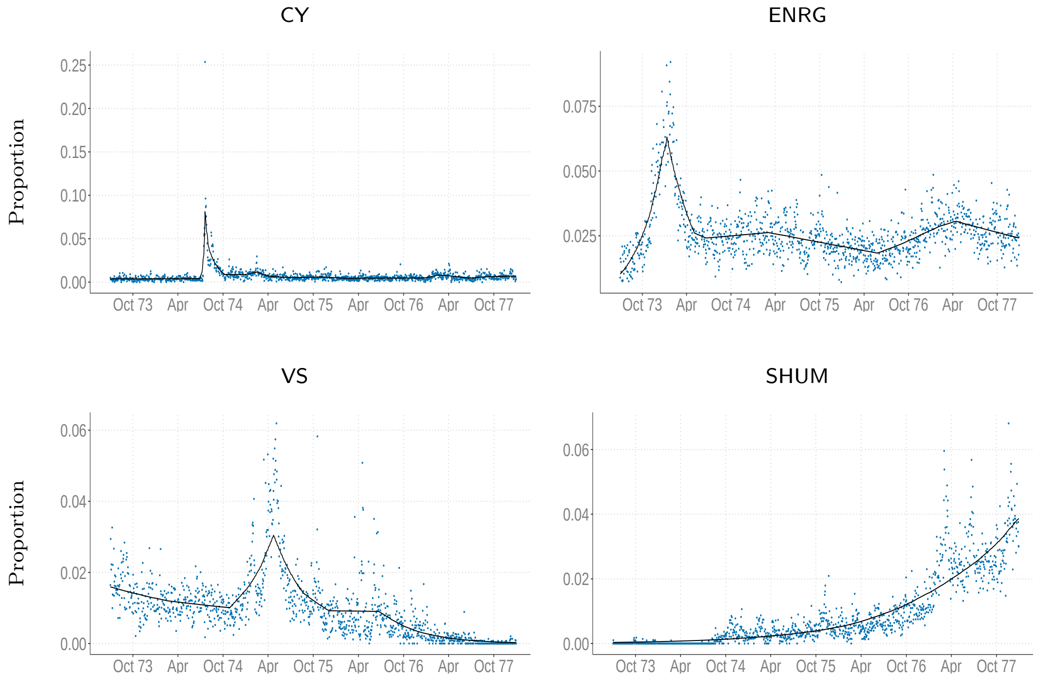


Figure 9: Figure showing the estimates obtained from Problem (2) with the ℓ_1-trend filtering regularizer (See Section 3). The sharp spike in the CY (Cyprus) communication stream corresponds to an unanticipated event, when Greek forces launched a coup with the goal of annexing Cyprus. The first peak for the second stream (ENRG) corresponds to the 1973 energy crisis, after the OPEC oil ministers announced an embargo during the Yom Kippur War. The peak for VS, for South Vietnam, corresponds to the Fall of Saigon in 1975, which marked the end of the Vietnam War. SHUM, for communications related to human rights, shows the increasing attention the State Department gave to this subject, especially after the election of President Jimmy Carter.

Significance level	number of TAGS for $\mathcal{T}$ test (1)
< 0.1	914
< 0.01	768
< 0.001	622
< 0.0001	509
< 0.00001	391

Table 1: Table showing how many TAGS -specific cables survive at different significance levels, with 0.00001 being the smallest detectible level. This is out of the first 1000 TAGS, which roughly corresponded to the TAGS with more than 50 total cables.

where, $\|D\mathbf{u}\|_1 = \sum_i |u_{i+1} - u_i|/\Delta_t$. More generally, if we consider the ℓ_1-trend filtering example with varying time intervals, we get an instance of Problem (12) with

$$\|D\mathbf{u}\|_1 := \sum_t \left| \left(\frac{u_{t+2} - u_{t+1}}{\Delta_{t+1}} - \frac{u_{t+1} - u_t}{\Delta_t} \right) \frac{1}{\Delta_t} \right|.$$

We use the Alternating Direction Method of Multipliers (ADMM) procedure [4] which performs the following decomposition: $\boldsymbol{\alpha} = D\mathbf{u}$ and obtains the Augmented Lagrangian:

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}; \boldsymbol{\nu}) = \frac{1}{2}\|\mathbf{u} - \bar{\mathbf{u}}\|_2^2 + \lambda'\|\boldsymbol{\alpha}\|_1 + \langle \boldsymbol{\alpha} - D\mathbf{u}, \boldsymbol{\nu}\rangle + \frac{\rho}{2}\|\boldsymbol{\alpha} - D\mathbf{u}\|_2^2,$$

for some choice of $\rho > 0$. The usual ADMM approach performs the following sequence of updates:

$$\begin{aligned} \mathbf{u} \leftarrow & \arg\min_{\mathbf{u}} \ \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}; \boldsymbol{\nu}) \\ \boldsymbol{\alpha} \leftarrow & \arg\min_{\boldsymbol{\alpha}} \ \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}; \boldsymbol{\nu}) \\ \boldsymbol{\nu} \leftarrow & \boldsymbol{\nu} + \rho(\boldsymbol{\alpha} - D\mathbf{u}), \end{aligned} \tag{13}$$

where, in the update wrt $\mathbf{u}$ other variables remain fixed, and the same applies for the update wrt $\boldsymbol{\alpha}$. We refer the reader to [4] for details pertaining to the convergence of this algorithm and choices of ρ. We note that the update wrt $\mathbf{u}$ in display (13) can be solved quite easily via solving a system of linear equations:

$$\mathbf{u} \leftarrow (\rho D'D + \mathbf{I})^{-1}(\bar{\mathbf{u}} + D'\mathbf{u} + \rho\mathbf{D}'\boldsymbol{\alpha}).$$

Note that $(\rho D'D + \mathbf{I})$ is a bidiagonal matrix when D corresponds to the weighted fused lasso penalty and a tridiagonal matrix when it corresponds to the weighted trend filtering penalty. The inverses in each of these cases can be computed with cost $O(2N)$ and $O(3N)$ (respectively) [5, 16] – furthermore the inverse can be computed once (at the onset) as the matrix does not change across iterations. The update wrt $\boldsymbol{\alpha}$ in display (13) requires a solving the following problem:

$$\boldsymbol{\alpha} \leftarrow \arg\min_{\boldsymbol{\alpha}} \quad \frac{\rho}{2}\|\boldsymbol{\alpha} - \mathbf{z}\|_2^2 + \lambda'\|\boldsymbol{\alpha}\|_1,$$

where, $\mathbf{z} = (D\mathbf{u} - \boldsymbol{\nu}/\rho)$. A solution to the above problem is given by the familiar soft-thresholding [12] operation where, $\alpha_i = \text{sgn}(z_i)\max\{|z_i| - \lambda'/\rho, 0\}$. The sequence of updates in (13) are performed till some form of convergence criterion is met [4].

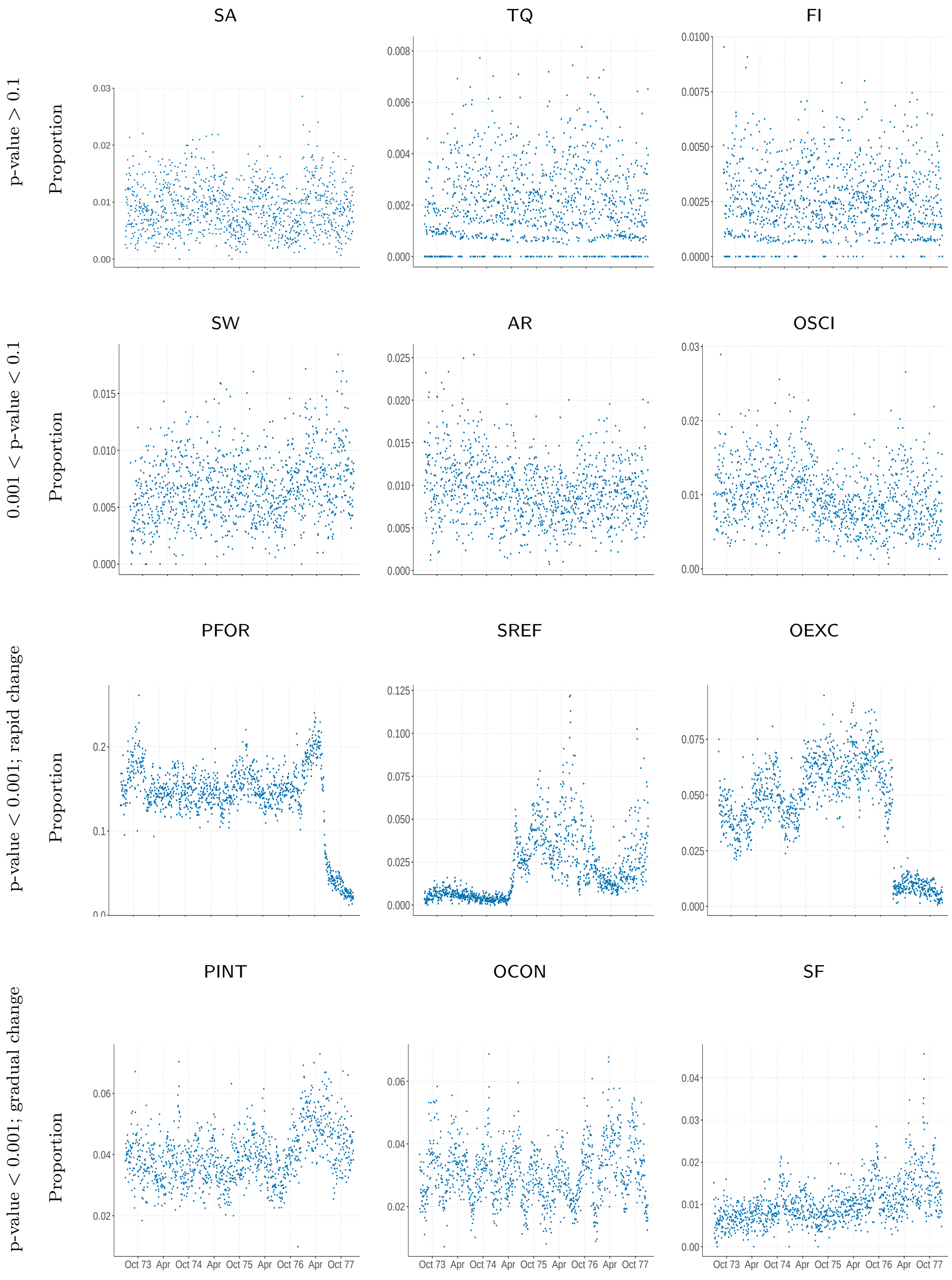


Figure 10: Communication streams with different significance scores in the spirit of Section 2.1. Top row shows the series where, like FI (for Finland), the p-values are larger than 0.1; second row has p-values in 0.1-0.001 (such as for OSCI, scientific grants). The third and fourth rows show series that seemed to have a high degree of intense activity: all p-values were smaller than 0.001. This includes, for example, SREF (for refugees) and SF (for South Africa.) As we will see, a strong deviation from the null model under the global testing framework does not necessarily imply a communication stream with a significant change point.

	TAGS	meaning	start	end	peak	Burst Strength
1	ETRN	Economic Affairs-Transportation	1973-07-02	1974-08-09	1973-09-28	5146.05
2	CVIS	Consular Affairs-Visas	1973-07-02	1975-01-02	1974-06-28	4839.35
3	SHUM	Social Affairs-Human Rights	1977-01-19	1977-12-30	1977-11-18	2872.02
4	US	United States	1976-01-28	1977-09-16	1976-04-15	2516.03
5	PGOV	Political Affairs-Government	1977-06-03	1977-12-30	1977-11-18	2484.57
6	SREF	Social Affairs-Refugees	1975-04-22	1976-07-20	1976-06-02	1662.58
7	SOPN	Social Affairs-Public Opinion and Information	1976-11-26	1977-12-30	1977-08-26	1597.14
8	PORG	Political Affairs-Policy Relations With International Organizations	1977-06-15	1977-12-30	1977-11-11	1547.35
9	PDIP	Political Affairs-Diplomatic and Consular Representation	1977-05-24	1977-12-30	1977-09-02	1462.93
10	XF	Middle East	1973-10-09	1973-12-19	1973-10-16	1453.76
11	AO	Angola	1975-11-08	1976-02-23	1975-11-10	1439.58
12	CY	Cyprus	1974-07-15	1974-07-29	1974-07-20	1378.79
13	VM	Vietnam	1977-10-11	1977-12-30	1977-10-12	1365.45
14	PDEV	Political Affairs-National Development	1977-06-13	1977-12-30	1977-08-31	1344.70
15	VS	Vietnam (South)	1973-07-02	1975-06-06	1975-04-25	1150.46
16	UNGA	UN General Assembly	1975-08-19	1975-12-13	1975-11-07	1044.29
17	CARR	Consular Affairs-Americans Arrested Abroad	1977-06-01	1977-12-30	1977-06-28	951.98
18	MCAP	Political Affairs-Military Capabilities	1973-07-02	1974-08-15	1974-07-03	903.26
19	ENRG	Economic Affairs-Energy	1973-11-08	1974-02-21	1974-01-25	760.64
20	PBOR	Political Affairs-Boundary and Sovereignity Claims	1977-07-01	1977-12-30	1977-11-09	685.75
21	OVIP	Operations-VIP Travel Arrangements	1974-10-09	1974-11-09	1974-10-31	607.17
22	RH	Rhodesia	1976-09-01	1977-12-30	1977-08-31	569.89
23	AEMR	Administration-Emergency and Evacuation	1975-03-28	1975-05-12	1975-04-28	524.59
24	MPLA	Popular Movement for the Liberation of Angola	1975-11-07	1976-02-24	1976-02-18	507.98
25	MSG	Marine Security Guards	1976-09-02	1977-12-30	1977-11-28	507.27
26	OREP	Operations-Congressional Travel	1976-10-27	1976-11-18	1976-11-02	481.08
27	PRG	Provisional Revolutionary Government of South Vietnam	1975-01-16	1975-02-06	1975-02-03	470.51
28	MNUC	Military and Defense Affairs-Military Nuclear Applications	1977-03-11	1977-12-30	1977-08-22	421.53
29	UNGA	UN General Assembly	1974-09-05	1974-12-05	1974-10-10	417.20
30	CB	Cambodia (Khmer Republic)	1973-07-02	1975-05-21	1975-04-16	370.14

Table 2: Top 30 bursts identified using ℓ_0 segmentation algorithm, using the method in Section 2.4 to compute burst strengths. For interpretations regarding the bursts please see the discussion in Section 2.4.2.

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

[2] John Beieler, Patrick T. Brandt, Andrew Halterman, Philip A. Schrodt, and Erin M. Simpson. Generating political event data in near real time: Opportunities and challenges. In R. Michael Alvarez, editor, *Computational Social Science*, pages 98–120. Cambridge University Press, 2016.

[3] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44 (2):23–42, 2016.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Number 3(1). Now Publishers, 2011.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[6] Leif Boysen, Angela Kempe, Volkmar Liebscher, Axel Munk, and Olaf Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, pages 157–183, 2009.

[7] Lester H Brune and Richard D Burns. *Chronological History of U.S. Foreign Relations: 1933-1988*. Routledge, 2003.

[8] Alexander De Conde, Richard D Burns, Fredrik Logevall, and Louise B. Ketz. *Encyclopedia of American foreign policy*. Scriber, 2002.

[9] Stephen A Flanders and Carl N Flanders. *Dictionary of American foreign affairs*. Macmillan Library Reference, 1993.

[10] Joseph Glaz, Joseph I Naus, Sylvan Wallenstein, Sylvan Wallenstein, and Joseph I Naus. *Scan statistics*. Springer, 2001.

[11] Alex Hanna. Assessing gdelt with handcoded protest data. www.badhessian.org, 2014. (accessed: July 29, 2016).

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction (Springer Series in Statistics)*. Springer New York, 2 edition, 2009.

[13] J Craig Jenkins and Thomas V Maher. What should we do about source selection in event data? challenges, progress, and possible solutions. *International Journal of Sociology*, 46(1):42–57, 2016.

[14] Bruce W Jentleson, Thomas G Paterson, and Nicholas X Rizopoulos. *Encyclopedia of US foreign relations*, volume 2. Oxford University Press, USA, 1997.

[15] Nicholas A. Johnson. A dynamic programming algorithm for the fused lasso and l0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

[16] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. \ell_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

[17] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[18] David Langbart, William Fischer, and Lisa Roberson. Appraisal of records covered by N1-59-07-3-P. *National Archives*, 2007.

[19] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[20] Enno Mammen, Sara van de Geer, et al. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

[21] Rahul Mazumder and Peter Radchenko. The discrete dantzig selector: Estimating sparse linear models via mixed integer linear optimization. *IEEE Transactions on Information Theory*, 2017.

[22] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*.

Kluwer, Norwell, 2004.

[23] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67:91–108, 2005.

[24] Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

[25] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.