MINING EVENTS WITH DECLASSIFIED DIPLOMATIC DOCUMENTS

BY YUANJUN GAO*,†, JACK GOETZ‡,§ MATTHEW CONNELLY¶,∥ AND RAHUL MAZUMDER‡,∥

Columbia University*,¶, Massachusetts Institute of Technology‡,∥ and University of Michigan§,∥

Since 1973 the State Department has been using electronic record 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 identify from a large corpus of documents a handful that historians would consider more interesting. Our approach brings together techniques from nonparametric signal estimation, statistical hypothesis testing and modern optimization methods—leading to a set of tools that help us identify and analyze various geometrical aspects of the communication streams. Dominant periods of heightened activities, 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 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 [4]. 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 what reporters write about, which can vary over time and geography depending on many extraneous factors [22]. 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 [16]. 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. The sheer volume of these records can make it difficult even for the diligent researcher to identify individual events and assess their relative importance.

*Department of Statistics, Columbia University
†Department of Statistics, University of Michigan
‡Sloan School of Management, Operations Research Center and Center for Statistics, MIT.
§Department of History, Columbia University
∥Department of History, Columbia University

Keywords and phrases: U. S. History, National Archives, signal estimation, fused lasso, optimization
In this paper we study a new dataset of declassified documents and use statistical methods to identify political events; and explore how these heterogeneous events manifest themselves in the form of different geometric characteristics of these diplomatic communication streams. Since 1973, the State Department has been using electronic records systems to preserve classified communications. The National Archives 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—creating many opportunities for statistical analyses.

Our goal is to explore a mix of classical and 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 of additional 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—our work explores how this relates to heterogeneous political events. To provide some intuition, 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 heightened traffic intensity that plays out over longer periods, such as an increase over time.

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 based on the traffic of communication streams. Unfortunately, reliable text data is unavailable for many thousands of the cables in this corpus, either because State Department storage systems failed to preserve it, or because only the metadata has been declassified for cables that have been deemed to contain sensitive information. In this context, the TAGS-specific features are found to be particularly useful and effective in identifying events of importance to a social scientist.

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 to be smaller compared

1Website: https://aad.archives.gov/aad/series-list.jsp?cat=WR43
to later years, due to fewer records. We study the time series at a 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; in panel (c) multiple bursts 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)). Changes in the proportion of a particular TAGS appearing in a communication stream seem to be 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 holidays—as a pre-processing step, we filtered out the days where the total number of cables being communicated were very small.

1.2. Scope of this work. A first goal of our work is to quantitatively define traits that separate communications like panel (d) from panels (a)–(c). We develop statistical methods that can mine these (TAGS-specific) time-series and identify communication streams that exhibit statistically interesting activities. Once they are identified, we develop algorithms (Sections 2 and 3) that perform a deeper investigation of each series and identify time intervals where the signal undergoes abrupt localized changes in communication traffic. We also present methods to quantify and contrast these various geometric patterns.

Our general methodological approach is inspired by statistical principles pioneered in the context of signal segmentation and change point modeling, with origins in 1950s [35]—see [39, 10] for excellent overview(s) of the topic. On the algorithmic front we employ ideas from modern first order methods in continuous optimization [32, 33], that complements state-of-the-art approaches in change point detection [24, 25]. Statistical models for change point detection have enjoyed a great deal of success across several application domains spanning speech processing, financial analysis, bio-informatics, climatology, network traffic, gait analysis, text processing, among others. Similar models are also employed in the context of burstiness analysis [27]—such ideas are enhanced in important ways for identifying events in Twitter communication streams; see [1, for example] for a nice survey. Our main goal in this paper is to build upon classical and modern statistical signal estimation/inference tools and enhance them in suitable ways so that they can provide insights to a historian on a new dataset available from the National Archives. A recurring theme of this work is to establish a synergy between statistics and social science perspectives. As we discuss in Section 4, analysis of diplomatic documents using TAGS-based features relevant to a historical scientist presents a unique set of challenges. It is quite different from the analysis of text-based features to mine events in Twitter communication streams. This makes our approaches different in scope.

2. Statistical Methodology. We first present a brief outline of the main statistical approaches pursued in this paper. Section 2.1 addresses how we can use a global testing approach to determine whether a TAGS-specific communication stream, among several hundreds, is interesting or not. To further explore the geometry of the underlying signal we use a regularized negative log-likelihood criterion based on the fused lasso penalty [37]; and also its $\ell_0$-counterpart [25, 9, 24]. For efficient computation, we propose a unified framework for these optimization problems, which seem to be promising alternatives to prior approaches [24, 25]. We use hypothesis testing ideas based on sample splitting [40] to associate p-values to the
detected jumps—see Section 2.3. Inspired by [27], the jumps in an individual series are aggregated to obtain “bursts”, leading to a rank-ordering of political events across the corpus. Finally in Section 3, we discuss how to estimate the underlying proportions with models that are more flexible than piecewise constant segments.

2.1. Identifying Interesting Communication Streams. 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 independent across \(t\). We are interested in the following question:

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 \(2\Delta\) and consider all the points in the \(\Delta\) neighborhood of a time point \(i\), given by:

\[
N(\Delta; i) = \{ j : 1 \leq j \leq N, |j-i|\leq \Delta \}\]

The average proportion in this neighborhood:

\[
p_{i}^{\text{ave}} := \frac{\sum_{j \in N(\Delta; i)} y_j}{\sum_{j \in N(\Delta; i)} n_j}
\]

is a measure of communication traffic around the reference time point \(i\). We say that a large value of \(p_{i}^{\text{ave}}\) compared to a baseline value \(\hat{p}_{H_0}\) indicates the presence of an intense localized activity of some form\(^2\). We hypothesize such an activity 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 global testing approach with \(H_0: p_t = p \ \forall t\) versus \(H_1:\) there exists an \(i\) such that \(p_{i}^{\text{ave}}\) is larger than the (global) average proportion. Inspired by popularly used scan statistics [15], we propose the following test statistic:

\[
(2.1) \quad \mathcal{T} = \max_{i} T_i \quad \text{where,} \quad T_i := \frac{(\hat{p}_{i}^{\text{ave}} - \hat{p}_{H_0})}{\hat{\sigma}_t},
\]

where, \(\hat{p}_{H_0}\) is the (estimated) global proportion of the signal estimated under the null hypothesis; \(\hat{p}_{i}^{\text{ave}}\) is an estimate of \(p_{i}^{\text{ave}}\) i.e., \(\hat{p}_{i}^{\text{ave}} = \sum_{j \in N(\Delta; i)} y_j / \sum_{j \in N(\Delta; i)} n_j\). Furthermore, \(\hat{\sigma}_t\) is the estimated standard deviation of \(\hat{p}_{i}^{\text{ave}}\) evaluated under the null \((H_0)\): If \(\hat{p}_{H_0}\) denotes the estimate of \(p\) under the null, then \(\hat{\sigma}_t^2 = (\hat{p}_{H_0}(1 - \hat{p}_{H_0})) / (\sum_{j \in N(\Delta; i)} n_j)\) (see Section A.1 for a derivation). \(T_i\) measures the strength of a locally contiguous period of heightened activity—we take the supremum over all time points \(t\) to get \(\mathcal{T}\). Larger the value of \(\mathcal{T}\), more pronounced is the localized traffic compared to the baseline value \(\hat{p}_{H_0}\). We use a permutation based approach to compute the null distribution of \(\mathcal{T}\). Figure 2 (see also Figure 10 for additional examples) shows different communication streams with their associated p-values. A large\(^3\) p-value for panel (a), Figure 2 (corresponding to TAGS FI) signifies a lack of interesting activity in this series—this aligns with an expert’s understanding that during this period, there was limited diplomatic activity at the international scale.

\(^2\)At this point, we do not offer an explanation of the exact geometric reason behind such an heightened activity—we address this at a later stage in the paper.

\(^3\)We note the choice of a p-value threshold (i.e., whether it is deemed to be large or small) often depends upon the subjective intuition of a practitioner.
related to TAGS FI.

To understand the sensitivity of results with $\Delta$, a summary of how many cables survive different $p$-value thresholds for different choices of $\Delta$, are provided in Table 2 (see Appendix). While the TAGS-specific $p$-values are found to change with $\Delta$ the overall results remain quite stable\textsuperscript{4}. As we use this step to remove a small fraction of the communication streams from further downstream analysis, this sensitivity to $\Delta$ does not affect our subsequent analysis. The results in Figure 2 suggest an important limitation of statistic (2.1)—this framework in itself does not offer much insight into the geometry of the signal. This motivates the methods presented in Section 2.2.

2.2. Identifying Jumps in Communication Streams. We propose methods to explore finer structural properties of the series that are not informed by the methods in Section 2.1. Inspired by popularly used signal segmentation/estimation meth-

\textsuperscript{4}Usually, the $p$-values smaller than 0.001 become smaller with increasing values of $\Delta$; larger $p$-values remain large.
ods [25, 37, 24, 29], we seek to identify breaks or jumps in a piecewise constant approximation of the signal \( t \mapsto p_t \).

**Regularized Maximum Likelihood.** Using the notation of Section 2.1, we assume \((y_t|n_t, p_t) \sim \text{Bin}(n_t, p_t)\) for \( t = 1, \ldots, N \) where, \( p_t \) denotes the probability of success and \( N \) denotes the total number of time points. This leads to a joint likelihood (conditional on \( \{(n_t, p_t)\}_{t \geq 1} \)) given by:

\[
P(N \{y_t\} | \{(n_t, p_t)\}_1^N) = \prod_{t=1}^{N} \binom{n_t}{y_t} p_t^{y_t} (1 - p_t)^{(n_t - y_t)}.
\]

Note that an unconstrained maximum likelihood estimator will lead to \( \hat{p}_t = y_t/n_t \) for all \( t \), which overfits the data. Therefore additional structural constraints on \( p_t \) are needed for interpretable models. Using the standard logistic parametrization: \( p_t = \exp(\theta_t)/(1 + \exp(\theta_t)) \), the negative log-likelihood (2.2) in terms of the variables \( \{\theta_t\}_1^N \) is:

\[
\sum_{t=1}^{N} \{ -y_t \theta_t + n_t \log(1 + \exp(\theta_t)) \} - \sum_{t=1}^{N} \log \left( \binom{n_t}{y_t} \right)
\]

where, the second term above does not depend upon \( \theta_t \), \( t \geq 1 \). The expression in (2.3) is convex in \( \{\theta_t\}_1^N \)—we consider \( \theta_t, t \geq 1 \) to be our natural parameters. We will approximate \( t \mapsto \theta_t \) by a piecewise constant signal (generalizations are discussed in Section 3). Locations where the underlying signal \( t \mapsto \theta_t \) exhibits a discontinuity will be called a “jump” in the communication stream. We consider the following regularized criterion:

\[
\min_{\theta_t, 1 \leq t \leq N} \sum_{t=1}^{N} \left( -y_t \theta_t + n_t \log(1 + \exp(\theta_t)) \right) + \lambda H(\theta),
\]

where, \( L(\theta) := \sum_{t=1}^{N} \{ -y_t \theta_t + n_t \log(1 + \exp(\theta_t)) \} \) is the part of (2.3) depending upon \( \{\theta_t\}_1^N \) (i.e., the data-fidelity term) and \( H(\theta) \) is the regularizer. \( H(\theta) \) encourages the estimated \( \theta_t \)’s (and hence the proportion \( p_t \)'s) to be piecewise constant and the regularization parameter \( \lambda > 0 \) controls the amount of shrinkage. Two examples of \( H(\theta) \) we study appear in earlier works in signal estimation [37, 29, 9, 24, 25, 39]—our intent is to present a simultaneous analysis for both these choices:

- \( \ell_1 \)-segmentation (fused lasso): Here \( H(\theta) = H_{\ell_1}(\theta) = \sum_{t=1}^{N-1} |\theta_{t+1} - \theta_t| \)—this penalizes the total variation of a signal; and may also be thought as a soft-version of the number of jumps in \( \theta_t, t \geq 1 \).

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

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.5.
Choice of regularizer. For the $\ell_1$ penalty $H_{\ell_1}(\theta)$, Problem (2.4) is a convex optimization problem. This is commonly referred to as the fused lasso or total-variation penalty \cite{37, 29} and used in the context of signal estimation wherein the underlying signal is assumed to have a small total variation norm. $H_{\ell_1}(\theta)$ shrinks the successive coefficient differences $\{\theta_{t+1} - \theta_t\}_t$; and due to the presence of the $\ell_1$-norm, encourages sparsity in $\theta_{t+1} - \theta_t$’s leading to a piecewise constant signal $t \mapsto \theta_t$. The shrinkage effect of the $\ell_1$-penalty severely penalizes large values of the jumps $\theta_{t+1} - \theta_t$. Hence, this penalty leads to a model with many jumps, especially when the tuning parameter is chosen so as to obtain a model with good data-fidelity (e.g., if the tuning parameter is chosen based on a held out test set to minimize test error). To obtain a model with fewer jumps, the regularization parameter needs to be made larger — in the process, important jumps may be missed. These observations are well-known in the context of the usual Lasso estimator in regression—see for example \cite{6, 30}. An alternative is to use an $\ell_0$-based penalty \cite{25, 9} which directly penalizes the number of jumps and is agnostic to the precise value of the jump. Both these penalty functions are popularly used in change point detection in statistics—see for example \cite{39} for a recent review. The rich literature on $\ell_0$ and $\ell_1$-based approaches seems to have grown somewhat independently of each other, with curious links and differences between the two approaches. Indeed the $\ell_0$ and $\ell_1$-based estimators have different operating characteristics—this is also seen in our numerical experiments. For example, if the amount of shrinkage for the $\ell_0$-penalty is not very high, one can obtain a signal with short segments. Furthermore, being agnostic to the magnitude of the jumps, the $\ell_0$-based estimator may lead to a signal estimate that is “spiky” — this is often ameliorated with $\ell_1$-based estimators, which shrinks the magnitude of a jump. For additional discussion on the delicate differences between $\ell_1$-and-$\ell_0$-based estimators in the regression context\textsuperscript{5}, we refer the reader to the recent works of \cite{31, 19}. The signal estimation problem we study here differs from the regression problem and poses a unique set of challenges. In this paper, we present a unified framework for both the $\ell_1$ and $\ell_0$-based estimators. We hope that this will allow us to gather interesting insights into the differences between the resulting estimators in the context of the problems studied here. More importantly, our framework will allow practitioners to make an informed decision regarding what might be more suitable in their context.

Other regularizers beyond the $\ell_0$ and $\ell_1$ penalties alluded to above, are also used in the context of change point models—see for example \cite{25, 39}. In Section 3 we discuss another regularizer that encourages a piecewise linear description of $t \mapsto \theta_t$.

\textsuperscript{5}The differences of these estimators depend upon a multitude of factors, such as: signal to noise ratios, feature correlations, model sparsity, sample size, number of features, etc.


end, we rely on efficient dynamic programming solvers \([24, 25]\) for the least squares loss with the \(H_{\ell_1}(\theta)\) or \(H_{\ell_0}(\theta)\) penalty.

The framework presented below applies to problems more general than Problem (2.4). In particular, they apply to a more general class of problems than that can be handled via the dynamic programming methods in \([24, 25]\)—they are complementary to the suite of algorithms used in change point models\(^6\) and may be of independent interest. We present proximal gradient descent methods \([3]\) for problems of the composite form \([33]\):

\[
\min_{\theta} \phi(\theta) := \mathcal{L}(\theta) + \lambda H(\theta),
\]

where, \(\mathcal{L}(\theta)\) is a function with Lipschitz continuous gradient:

\[
\|\nabla \mathcal{L}(u) - \nabla \mathcal{L}(v)\| \leq \ell \|u - v\|, \quad \forall u, v \in \mathbb{R}^N.
\]

In the case of Problem (2.4) we have \(\ell = \frac{1}{4} \max_{i=1}^{N} n_i\). This follows by noting that the \(i\)th coordinate of \(\nabla \mathcal{L}(u)\) is: \(\{\nabla \mathcal{L}(u)\}_i = -y_i + n_i \exp(u_i)/(1 + \exp(u_i))\) and \(\nabla^2 \mathcal{L}(u)\) is a diagonal matrix with the \(i\)th diagonal entry satisfying:

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

Hence, the largest eigenvalue of \(\nabla^2 \mathcal{L}(u)\), i.e., \(\lambda_{\max}(\nabla^2 \mathcal{L}(u)) \leq \frac{1}{4} \max_{i=1}^{N} n_i\), which justifies the choice of \(\ell\), as above. For a fixed \(L \geq \ell\), the proximal gradient algorithm performs the following updates (for all \(k \geq 0\)):

\[
\theta_{k+1} \in \arg \min_{\theta} \frac{L}{2} \| \theta - \left( \theta_k - \frac{1}{L} \nabla \mathcal{L}(\theta_k) \right) \|_2^2 + \lambda H(\theta).
\]

This leads to a decreasing sequence of objective values \(\phi(\theta_{k+1}) \leq \phi(\theta_k)\) for \(k \geq 0\). If \(\phi(\theta)\) is bounded below (which is true for Problem (2.4) as soon as \(n_i > 0\) for all \(i\)), then \(\phi(\theta_k)\) converges to a finite value. We now study the fate of the sequence \(\phi(\theta_k)\) (and \(\theta_k\)) depending upon the choice of \(H(\theta)\).

**The fused lasso penalty** \((H_{\ell_1}(\theta))\). Due to the convexity of \(H(\theta)\), the function \(\phi(\theta)\) is convex in \(\theta\). Using standard results in proximal gradient methods, \(\theta_k\) converges to a minimum of Problem (2.5) with the penalty function \(H_{\ell_1}(\theta)\). In terms of convergence rates of objective values, it follows from \([3]\) that:

\[
\phi(\theta_k) - \phi(\theta^*) \leq \frac{L}{2k} \|\theta_0 - \theta^*\|_2^2,
\]

where, \(\theta^*\) is an optimal solution to Problem (2.5)—hence, the sequence \(\phi(\theta_k)\) converges to the minimum of Problem (2.5) at a worst case rate of \(O(\frac{1}{k})\). If \(\mathcal{L}(\theta)\) is strongly convex in \(\theta\), the sequence \(\phi(\theta_k)\) exhibits a linear rate \([33]\) of convergence.

\(^6\)Due to the generality of our methods, they may not lead to optimal solutions to Problem (2.4) if \(H(\theta)\) is nonconvex.
of $\mathcal{L}(\theta_k)$. Since, $\theta_k$’s are uniformly bounded\footnote{Note that for Problem (2.5), $\min_{\theta} \mathcal{L}(\theta)$ has a finite minimizer (as $n_i > 0$ for all $i$) — hence, an optimal solution to Problem (2.5) is finite; and the sequence $\{\theta_k\}$ is uniformly bounded.} and $\min_{\theta} n_i > 0$ then $\mu := \inf_k \mu_k > 0$. The convergence rate is given by [33]:

\[
(2.10) \quad \phi(\theta_k) - \phi(\theta^*) \leq \left(1 - \frac{\mu}{4L}\right)^k (\phi(\theta_0) - \phi(\theta^*)).
\]

The rates in (2.9) and (2.10) are interesting to interpret. If $\frac{\mu}{4L}$ is small, the sub-linear rate (2.9) explains the convergence speed of $\phi(\theta_k)$ in the initial stages of the algorithm, after which the linear rate (2.10) will kick in. If $\frac{\mu}{4L}$ is large, the linear rate of convergence dominates and the algorithm converges very fast. Curiously, the convergence speed of the algorithm adapts to the better of the linear or sublinear rates—the algorithm does not need to be modified.

We note that accelerated variants [33] of proximal gradient descent can also be used — they improve the convergence rate in (2.9) to $O(1/k^2)$.

Note that sub-problem (2.8) is a problem of the form (for some $\lambda' > 0$)

\[
(2.11) \quad \text{minimize}_{u \in \mathbb{R}^N} \frac{1}{2} \|u - \bar{u}\|^2_2 + \lambda' \sum_{i=1}^{N-1} |u_{i+1} - u_i|.
\]

This can be solved very efficiently via dynamic programming [24] with a worst case cost of $O(N)$—in fact for $N \approx 10^6$ the solver of [24] takes usually around 0.03 seconds on a modest desktop computer. Hence, obtaining a solution to Problem (2.4) for a similar size usually takes at most a second.

The $\ell_0$-segmentation penalty ($H_{\ell_0}(\theta)$). The algorithm above can also be applied for the penalty $H_{\ell_0}(\theta)$. In update (2.8) we set $H(\theta)$ to $H_{\ell_0}(\theta)$. Instead of Problem (2.11), we solve the following jump penalized least squares problem:

\[
(2.12) \quad \text{minimize}_{u \in \mathbb{R}^N} \frac{1}{2} \|u - \bar{u}\|^2_2 + \lambda' \sum_{i=1}^{N-1} 1(u_{i+1} \neq u_i),
\]

which can be computed efficiently using the dynamic programming algorithm(s) of [24, 25]. Intuitively, our proposed proximal method approximates the smooth part of the loss function (by a quadratic objective as in a Newton method). The main non-convexity of the optimization problem lies in solving (2.12), which can be solved to optimality via dynamic programming. The proximal gradient algorithm outlined above is different from the dynamic programming algorithm [24, 25] that can be used to solve Problem (2.4) to optimality. However, as we mentioned before, the authors do not provide efficient implementations for Problem (2.4) in their packages. When direct comparisons are possible, our proposed algorithm seems to be faster than [24] (See Section 2.2.2 for details). While our algorithm (in theory) leads to a local solution of Problem (2.4), in our numerical experiments, solutions were found to be near-optimal. For Problem (2.4), our method delivers better solutions compared to the methods of [24, 25] for the $\ell_0$-based Gaussian model.

Describing the properties of the sequence $\theta_k, k \geq 1$ is subtle for Problem (2.4) (with penalty function $H_{\ell_0}(\theta)$) due to non-convexity of the optimization problem.
Following [6] it can be shown that the sequence $\phi(\theta_k)$ is decreasing, bounded below* and it converges to $\phi^*$ (this may depend upon the initialization). We say that $\tilde{\theta}$ is a first-order stationary point for Problem (2.4) if setting $\theta_k = \tilde{\theta}$ leads to $\theta_{k+1} = \tilde{\theta}$. We say that $\theta_k$ is an $\epsilon$-accurate first order stationary point for Problem (2.5) if $\left\|\theta_{k+1} - \theta_k\right\|^2 \leq \epsilon$. Following the convergence analysis in [6][Theorem 3.1], we obtain the following finite-time convergence rate of $\theta_k$ to a first order stationary point:

$$\min_{0 \leq k \leq K} \left\|\theta_k - \theta_{k-1}\right\|^2 \leq \frac{2(\phi(\theta_0) - \phi^*)}{K(L - \ell)}.$$  

The above convergence rate is conservative — in practice, $\phi(\theta_k)$ is found to converge much faster (usually, within 10 iterations). In fact following [6] one can establish that the support of the successive differences in the vector $\theta_k$ will converge after finitely many iterations — then the proximal gradient updates will correspond to usual gradient updates on a smooth function. Hence the sequence $\theta_k$ will converge, and exhibit an asymptotic linear rate of convergence.

A constrained variant of Problem (2.5). The above framework for the penalized problem can be extended to a constrained version of the form:

$$\min_{\theta} L(\theta) \quad \text{s.t.} \quad H(\theta) \leq \kappa,$$

where, $\kappa \geq 0$ is the regularization parameter (in constrained form). An important instance of the above corresponds to the choice $H(\theta) = H_{\ell_0}(\theta)$ — in which case we constrain the total number of jumps to a pre-set value $\kappa$. To obtain solutions to Problem (2.14), we will need to modify the proximal operator (2.8) to:

$$\theta_{k+1} \in \arg \min_{\theta} \left\|\theta - \left(\theta_k - \frac{1}{L} \nabla L(\theta_k)\right)\right\|^2 \quad \text{s.t.} \quad H(\theta) \leq \kappa.$$  

For $H(\theta) = H_{\ell_0}(\theta)$ this proximal operator can (once again) be solved using the dynamic programming framework of [24, 25] (at a slightly higher cost than the jump penalized least squares problem). When $H(\theta) = H_{\ell_1}(\theta)$, the solutions obtained from (2.14) are in one-to-one correspondence to solutions from (2.5). This is no longer true when $H(\theta) = H_{\ell_0}(\theta)$—it may be possible that for certain choices of $\kappa$, Problem (2.5) does not lead to a solution with $H(\theta) = \kappa$ for any value of $\lambda$. In this sense, formulation (2.14) with $H(\theta) = H_{\ell_0}(\theta)$ may be more favorable than the penalized version.

2.2.2. Related work (Algorithms). The origins of change point modeling in statistics date back to 1950s [35]—see [39, 10] for excellent overview(s) of the topic. Dynamic programming based segmentation for curve-fitting appeared in 1960s [5]. Approximate segmentation methods e.g., based on binary segmentation appear in [36, 34]—these are popular heuristics that may not lead to an optimal solution to the nonconvex $\ell_0$-penalized problem, as pointed out by [25] and others. Some more recent exact segmentation algorithms appear in [2, 21]—for a sequence of length $N$

*This is satisfied under minor conditions as discussed earlier.
these methods have a computational cost of $O(N^2)$. In [25], the authors use a pruning step to improve the algorithm of [21]. Related algorithms based on improvements of dynamic programming also appear in [24].

The two works most related to the segmentation problem discussed in Section 2.2.1, are [25, 24]. The \texttt{R} package \texttt{changepoint} provides a user-friendly interface for mean and/or variance changes for Gaussian data, and some other distributions (though not the Binomial distribution considered in this paper). In Section A.3 we present a comparison of the change points obtained by the algorithm in \texttt{changepoint} (for mean changes in Gaussian data) versus the method presented here for the $\ell_0$ jump penalized problem.

Under some assumptions, the method of [25] has an expected cost of $O(N)$, though the worst-case cost is $O(N^2)$. [24] describe dynamic programming algorithms for the $\ell_1$ and $\ell_0$-penalties with a variety of loss functions (separable across time points). The \texttt{R} package of [24] provides an user friendly interface for the squared error loss for both these penalty functions—their cost is $O(N)$. In our experience, the algorithm of [24] was found to be faster than [25] for large instances of Problem (2.12) (Gaussian $\ell_0$ segmentation): For problems with $N = 10^5$ and $N = 10^6$ [24] exhibits a 300x-fold and 2000x-fold improvement (respectively) in runtimes over the algorithm of [25]. The signals we consider are much smaller (usually, $N \approx 1500$), hence the time difference between [24] and [25] are less pronounced ([24] leads to a 5x-fold improvement) for computing a solution to Problem (2.12). However, for every TAGS-specific time-series, we use multiple calls to this function via (2.8) and consider multiple values of $\lambda$. When aggregated across all the TAG-specific time series, there is an overall time-benefit in using the framework of [24].

For the $\ell_1$-based segmentation problem with the Binomial likelihood, we compared our algorithm with the code of [24]. For signals of length $N = 10^4, N = 10^5$ and $N = 10^6$ the implementation\footnote{We note that the paper [24] presents an implementation in \texttt{R} for this problem, unlike the least squares problem which is written in \texttt{C} with a \texttt{R} wrapper.} of [24] takes around 1.5, 18 and 197 seconds (respectively)—our method exhibits 187, 267 and 540x fold speedups (respectively).

Since [25, 24] rely on dynamic programming, the class of loss functions considered is rather limited. Our proposed framework can address a larger family of loss functions. Since we rely on existing efficient solvers for Problem (2.8), our algorithms are also efficient — with a cost of $O(N)$ per-iteration. It is not clear if and how the dynamic programming algorithms [25, 24] can be adapted to more flexible penalties—e.g., incorporating unequal time spacing for the $\ell_1$-based problem (cf Section A.5) and obtaining piecewise linear segments (cf Section 3) — we can handle both these generalizations. Thus, Section 2.2.1 presents an easy-to-implement alternative approach to obtain good solutions to problem (2.5) by leveraging excellent solvers for the least squares problem and promises to be an interesting alternative for signal segmentation tasks.

Signal segmentation problems also appear in the computer science/data-mining literature. In a seminal paper, Kleinberg [27] formalize models for detecting bursts in events in the context of a continuous steam of events (e.g., email messages over time); and discrete time events (events arriving in batches as in conference papers)—they
have close ties to change-point models in statistics. For the first case, Kleinberg models email messages arriving over time with an exponential inter-arrival rate. These rates can take values in a finite set (aka states) and are of the form: \( q_i = q_0 s^t \) for some scale factor \( s \), and \( q_0 \) is a pre-specified base-rate. The rates can change over time, but there is a penalty for an increase in the current rate to a newer one. He uses dynamic programming algorithms for hidden Markov models to perform the estimation. The computational cost increases with the number of states — the R package bursts (that implements the algorithms of [27]) is usually found to be much slower than the implementations of [25, 24] alluded to above. For the case of events occurring in batches over discrete periods of time, Kleinberg uses a Binomial model for every time point—there are two models corresponding to success probabilities \( q_1, q_0 \), with \( q_0 \) denoting the known baseline and \( q_1 = q_0 s \) for some pre-specified scale factor \( s \). Our framework is different: While Kleinberg allows \( p_t \) to take two pre-specified values, we allow \( p_t \) to take a continuum of values in \([0, 1]\). We allow for a flexible family of penalty functions \( H_{\ell_0} \) and \( H_{\ell_1} \) that penalize any increase/decrease in \( p_t \), while Kleinberg penalizes only an increase. Our algorithms are also much faster as we leverage the efficient solvers of [25, 24]. It is also known that Kleinberg’s model may lead to under-smoothing—to this end, localized window averaging is often recommended as a smooth pre-processing step. The regularized likelihood framework considered in this paper, systematically addresses the smoothing/under-smoothing tradeoff by adjusting the regularization parameter \( \lambda \) (e.g., based on cross-validation).

Fig 3. Estimators obtained from Problem (2.4) with \( \ell_1 \) (upper panel) and \( \ell_0 \) (lower panel) regularization. The data is synthetic and the underlying signal contains two sharp jumps (on the left) and a gradual increasing trend (on the right). We use cross validation to select a value of \( \lambda \). The \( \ell_1 \) penalty shrinks the estimated probability during a big burst \((501 \leq t \leq 550)\) and leads to more jumps during the gradual increase period \((551 \leq t \leq 1203)\). The \( \ell_0 \)-based estimator leads to a better estimate of the signal burst, and leads to fewer jumps during the gradual increase period.

\(^{10}\)The penalty for a change in rate is proportional to: \((j - i)\) if \( q_j > q_i \)—there is no penalty for a decrease in rate.

\(^{11}\)In the special case, \( p_t \in \{q_0, q_1\} \), the penalty function choice should not make much of a difference, with the exception of sudden decreases towards the boundary of the signal.
2.2.3. Estimated Signal. To gather some intuition about the behavior of the estimators described above, we consider a synthetic example in Figure 3 and some real datasets in Figures 4 and 5.

Illustration with a synthetic dataset. In this synthetic example (Figure 3), 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. Note that the underlying signal is not piecewise constant—there is model misspecification due to the linearity on the right part of the signal. This example is chosen to shed light into the behaviors of the $\ell_0$ and $\ell_1$-based estimators for the real datasets studied herein, where there is obvious model mis-specification. Figure 3 presents the signal estimates (for both the $\ell_0$ and $\ell_1$ penalties) at the cross-validated choices of the tuning parameter— we use $k$-fold (with $k=10$) cross validation [18] 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 $k$th point into the same fold). For both 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 $\ell_1$-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 $\ell_0$-scheme. In addition, the estimates for the linear component (at the right) also differ across the $\ell_0$ and $\ell_1$ schemes. The $\ell_0$ regularizer leads to a fewer number of segments (here three) compared to the $\ell_1$-penalty which has several smaller jumps.

Illustration on TAGS series. Figures 4 (TAGS UNGA) and 5 (TAGS VS) show the estimated signal proportions obtained via estimator (2.4). Both the penalty functions do a good job in estimating a piecewise constant version of the underlying signal—the $\ell_0$ scheme leads to fewer jumps than its $\ell_1$ counterpart, for a comparable data-fidelity. The figures also show fitted signals for a few other values of $\lambda$ 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 [18] (see also the R package genlasso). We can see that as $\lambda$ decreases, the algorithm captures a more granular structure of the data and estimates more jumps.

Figure 4 shows communication traffic corresponding to TAGS-specific to the UN General Assembly. The cyclical jumps correspond to the regular Fall meetings of the General Assembly. In addition, our signal estimate suggests additional peaked activities, for example a jump in April-May 1974. It seems that this jump is of smaller intensity compared to the other regular peaks corresponding to the annual meetings. Further investigation revealed that this jump occurred when Algeria called a special session to demand UN support for a “New International Economic Order.”

More specifically, data is generated by $y_t \sim \text{Bin}(n_t = 200, p_t), t = 1, \ldots, 1203$, 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 $551 \leq t \leq 1203$. (Note that a constant value of $n_t$ is a simplification—in the real dataset $n_t$ depends upon $t$.)

For Figure 4, the λ values were (a) for the $\ell_0$-penalty: 2.2, 4.8 and 13.7 (top to bottom); and (b) for the $\ell_1$-penalty: 39.2, 66.3 and 189 (top to bottom). For Figure 5, the λ values were (a) for the $\ell_0$-penalty: 2.9, 8.1 and 23.2 (top to bottom); and (b) for the $\ell_1$-penalty: 86.1, 245.7 and 701 (top to bottom).
In Figure 5 (TAGS related to South Vietnam), we observe that the most intense jump corresponds to the fall of Saigon and the end of the Vietnam War. The signal estimate suggests that this event is accompanied by a peak in communication traffic in April 1975—this happened with the collapse of the South Vietnamese regime and the rush to evacuate American personnel. A social scientist might also be interested to identify and explore smaller jumps; here they correspond to the refugee crisis that continued into the following year.
Fig 5. Figure showing the raw proportions (in blue dots) for TAGS VS (South Vietnam) and the estimated proportions $p_t$, $t \geq 1$, as obtained from the regularization framework in Problem (2.4). The left panel shows the results for the $\ell_0$-segmentation penalty and the right panel the $\ell_1$-penalty. The middle rows correspond to the optimal $\lambda$ (as discussed in the text), and we show a few additional choices of the regularization parameter for each example. The biggest burst corresponds to the fall of Saigon and the end of the Vietnam War. A social scientist 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.

2.3. A deeper investigation of Jumps. The framework of Section 2.2 can be used to obtain a simple piecewise constant approximation of the underlying signal. Upon further investigation, they offer insights into the behavior of communication streams. Section 2.3.1 discusses how to quantify the intensity of a jump using sample-splitting ideas [40]. In Section 2.4 we show how these jumps can be aggregated to obtain the notion of a “burst” [27] in a communication stream. We also present related social science perspectives.
2.3.1. How intense is a Jump?. A jump estimated by the $\ell_0$ or $\ell_1$-segmentation procedure may reflect (a) a discontinuity in the signal, as we saw in the first half of Figure 3— in this case, the signal is well approximated by locally constant segments with pieces adapting to the data; and/or (b) a localized trend in the signal, as we saw in the second half of Figure 3. A jump in (b) is a consequence of the slope of $t \mapsto p_t$ and not a discontinuity. A piecewise constant signal can be thought of approximating the underlying (linear) trend in $t \mapsto p_t$. 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 as follows: given an estimate $\{\hat{p}_t\}$ and a set of candidate jumps, can we obtain a scoring for their strengths and sizes? This would lead to a smaller set of jumps that merit closer scrutiny. Towards this end, we use a sample splitting procedure: 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 that is identified in the first stage. In other words, the training set is created by randomly choosing half the cables for each day, with the remaining half set aside for testing purposes.

Suppose $\hat{t}$ is a candidate change point based on the first part of the sample (used for estimating the signal). We denote the time points on the left of $\hat{t}$ as $L(\hat{t})$ and those on the right of $\hat{t}$ as $R(\hat{t})$—these segments $L(\hat{t})$ and $R(\hat{t})$ do not contain any jumps. We assume that $p_t$ for $t \in L(\hat{t})$ are all equal to $p(L, \hat{t})$; and $p_t$ for $t \in R(\hat{t})$ are all equal to $p(R, \hat{t})$. We 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 where, the null distribution is computed based on a permutation test.

Note that a candidate jump obtained at the cross-validated choice of the tuning parameter need not have a low p-value. The p-values thus obtained, 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 significant difference in proportions between the left and right intervals. Scheme (b) is useful for estimators obtained by the $\ell_1$-segmentation scheme, as it is known to make false discoveries of change-point locations (even if the underlying signal is piecewise constant). In our application, the underlying signal is not piecewise constant—it simply serves as an useful approximation. In this case, the p-value scores serve as an useful metric to measure the strength of a jump.

Figure 6 shows the communication stream for TAGS CVIS (Consular Affairs-
Visas) and the estimated signal obtained via $\ell_0$-segmentation. We also computed the p-value scores for each potential jump location as suggested via the $\ell_0$-segmentation fit. The sharp jump in TAGS CVIS that persists across the panels, starts around mid September, 1975 and ends around early October, 1975. This spike is rather curious from a social science perspective. Prior to this period, the number of communications related to visa applications sharply decreased not because this kind of activity diminished, but because archivists decided to stop retaining these records. The exception is a two-week period in September 1975, when the number of CVIS records is comparable or even higher than before. Some concern Indochinese refugees, but many others involve the FBI, especially visas for people coming from the USSR and other communist states. Few if any of these FBI communications have message text. This is intriguing given the (limited) information available to us from the declassified cable documents; and suggests the need for further investigation. This is the kind of anomaly that a social scientist might miss without using this kind of statistical framework.
Figure 7 presents two examples for TAGS ENRG (energy) and TAGS US (for cables relating to the United States). The jumps in TAGS ENRG are much sharper and indicate rapid (though not instantaneous) changes in mean, starting with the 1973 OPEC oil embargo. For the TAGS US figure, 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). The intuition conveyed above is validated via a synthetic example in Figure 9 (with the same data as in Figure 3).

2.4. From Jumps to Bursts. We discuss how to summarize a single communication stream (corresponding to a specific TAGS) with a score that aggregates different jumps into a “burst” — this terminology was introduced by Kleinberg [27] in the context of event detection. Informally speaking, a burst corresponds to a stretch of time where a communication stream depicts traffic larger than a baseline value. The approach we present here differs from Kleinberg [27] (See also discussion in Section 2.2.2), who uses a Binomial model with two states at every time point. Kleinberg defines the weight of a burst to be the aggregated sum of the differences in the loss function (data fidelity term) across these two states. In our approach, we do not restrict $p_t$ to two apriori specified states—instead, we allow for a continuum of values that are obtained by solving a regularized signal segmentation problem.

2.4.1. Computation of the strength of a Burst. Suppose we are given an estimate of a baseline proportion $p_0$ (we discuss how to compute this below) for 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 $T = [t_{\text{start}}, t_{\text{end}}]$, where $\hat{p}_t > p_0, \forall t \in T$. Generalizing the notion of weight of a burst presented in [27], we define the strength $S(T)$ 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: $S(T) = \sum_{t \in T} (\log L(\hat{p}_t | n_t, y_t) - \log L(p_0 | n_t, y_t))$, where $L(\hat{p}_t | n_t, y_t)$ denotes the likelihood at time $t$. As the baseline $p_0$ is specific to a communication stream, the score $S(T)$ represents a deviation from this global baseline. $S(T)$ is different than the magnitude of a jump given by $\hat{p}_{t+1} - \hat{p}_t$—it takes into account the deviation of $\hat{p}_t$ from the baseline $p_0$ as well as the duration of the burst given by the length of $T$. A large value of $S(T)$ 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-specific communication stream can have multiple bursts leading to multiple intervals $T$ — each with an assigned strength $S(T)$.

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

$$p_0 = \bar{p} + \sqrt{\frac{\bar{p}(1 - \bar{p})}{\bar{n}}}, \quad \text{where,} \quad \bar{p} = \frac{\sum_{t=1}^{N} y_t}{\sum_{t=1}^{N} n_t}, \quad \bar{n} = \frac{1}{N} \sum_{t=1}^{N} n_t.$$
A robust estimate like the median can also be used instead of the average. In our experiments we found that the top-ranked slots (cf Table 1) were relatively agnostic to the choice of the baseline $p_0$.

<table>
<thead>
<tr>
<th>TAGS</th>
<th>meaning</th>
<th>start</th>
<th>end</th>
<th>peak</th>
<th>Burst Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ETRN</td>
<td>Economic Affairs-Transportation</td>
<td>1973-07-02</td>
<td>1974-08-09</td>
<td>1973-09-28</td>
<td>5146.05</td>
</tr>
<tr>
<td>2 CVIS</td>
<td>Consular Affairs-Visas</td>
<td>1973-07-02</td>
<td>1974-01-02</td>
<td>1974-06-28</td>
<td>4839.35</td>
</tr>
<tr>
<td>3 SHUM</td>
<td>Social Affairs-Human Rights</td>
<td>1977-01-19</td>
<td>1977-12-30</td>
<td>1977-11-18</td>
<td>2872.02</td>
</tr>
<tr>
<td>7 SOPN</td>
<td>Social Affairs-Public Opinion and Information</td>
<td>1976-11-26</td>
<td>1977-12-30</td>
<td>1977-08-26</td>
<td>1507.14</td>
</tr>
<tr>
<td>10 XF</td>
<td>Middle East</td>
<td>1973-10-09</td>
<td>1973-12-19</td>
<td>1973-10-16</td>
<td>1453.76</td>
</tr>
<tr>
<td>11 AO</td>
<td>Angola</td>
<td>1975-11-08</td>
<td>1976-02-23</td>
<td>1975-11-10</td>
<td>1439.58</td>
</tr>
<tr>
<td>12 CY</td>
<td>Cyprus</td>
<td>1974-07-15</td>
<td>1974-07-29</td>
<td>1974-07-20</td>
<td>1378.79</td>
</tr>
<tr>
<td>13 VM</td>
<td>Vietnam</td>
<td>1977-10-11</td>
<td>1977-12-30</td>
<td>1977-10-12</td>
<td>1365.45</td>
</tr>
<tr>
<td>14 PDEV</td>
<td>Political Affairs-National Development</td>
<td>1977-06-13</td>
<td>1977-12-30</td>
<td>1977-08-31</td>
<td>1344.70</td>
</tr>
<tr>
<td>15 VS</td>
<td>Vietnam (South)</td>
<td>1973-07-02</td>
<td>1975-06-06</td>
<td>1975-04-25</td>
<td>1150.46</td>
</tr>
<tr>
<td>16 UNGA</td>
<td>UN General Assembly</td>
<td>1975-08-19</td>
<td>1975-12-13</td>
<td>1975-11-07</td>
<td>1044.29</td>
</tr>
<tr>
<td>17 CARR</td>
<td>Consular Affairs-Americans Arrested Abroad</td>
<td>1977-06-01</td>
<td>1977-12-30</td>
<td>1977-06-28</td>
<td>951.98</td>
</tr>
<tr>
<td>18 MCAP</td>
<td>Political Affairs-Military Capabilities</td>
<td>1973-07-02</td>
<td>1974-08-15</td>
<td>1974-07-03</td>
<td>903.26</td>
</tr>
<tr>
<td>19 ENRG</td>
<td>Economic Affairs-Energy</td>
<td>1973-11-08</td>
<td>1974-02-21</td>
<td>1974-01-25</td>
<td>760.64</td>
</tr>
<tr>
<td>20 PBOR</td>
<td>Political Affairs-Boundary and Sovereignty Claims</td>
<td>1977-07-01</td>
<td>1977-12-30</td>
<td>1977-11-09</td>
<td>685.75</td>
</tr>
<tr>
<td>22 RH</td>
<td>Rhodesia</td>
<td>1976-09-01</td>
<td>1977-12-30</td>
<td>1977-08-31</td>
<td>599.89</td>
</tr>
<tr>
<td>26 OREP</td>
<td>Operations-Congressional Travel</td>
<td>1976-10-27</td>
<td>1976-11-18</td>
<td>1976-11-02</td>
<td>481.08</td>
</tr>
<tr>
<td>27 PRG</td>
<td>Provisional Revolutionary Government of South Vietnam</td>
<td>1975-01-16</td>
<td>1975-02-06</td>
<td>1975-02-03</td>
<td>470.51</td>
</tr>
<tr>
<td>29 UNGA</td>
<td>UN General Assembly</td>
<td>1974-04-05</td>
<td>1974-12-05</td>
<td>1974-10-10</td>
<td>417.20</td>
</tr>
<tr>
<td>30 CB</td>
<td>Cambodia (Khmer Republic)</td>
<td>1973-07-02</td>
<td>1975-05-21</td>
<td>1975-04-16</td>
<td>370.14</td>
</tr>
</tbody>
</table>

Table 1
Top 30 bursts identified using $\ell_0$ 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.
2.4.2. **Interpretation of Bursts.** Table 1 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 event 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 [28]. 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).

To validate these results, we consulted four standard reference works on U.S. foreign relations [11, 13, 23, 12]. The editors are all domain experts, but the varying content of each one reflects the different ways social scientists evaluate historical significance. Nevertheless, all of the aforementioned events appear in every one of these reference works, suggesting our framework succeeds in identifying events that are broadly recognized as historically important.

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 denouncement 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). This suggests that the ranking of the bursts – while not necessarily corresponding to historical importance – does reflect the likelihood that each one will correspond to events historians have already recognized as significant. But among the unrecognized events, like a 1975 UN General Assembly debate over the command of foreign military forces in South Korea, there are some that 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. **A generalization 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 (i.e., piecewise constant) approximation of the communication streams. However, as we discussed before many of the signals are not piecewise linear. 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 subtle technical issues in incorporating this structure into our likelihood framework, as we discuss below.

Let us consider the usual signal denoising problem with data: \( \tilde{y}_i = \mu_i + \epsilon_i \), for \( i = 1, \ldots, N \) where, \( \epsilon_i \sim N(0, \sigma^2) \). Suppose we would like to estimate \( \mu \) such that it is piecewise linear. A method to achieve this is by using the recently popularized \( \ell_1 \) trend-filtering approach [26, 38]. Here, one uses a convex regularizer \( H_{\ell_1}^{tf}(\mu) = \sum_t |\mu_{t+2} - 2\mu_{t+1} + \mu_t| \) to obtain a signal with piecewise linear segments:

\[
(3.1) \quad \text{minimize} \quad \frac{1}{2} \sum_{i=1}^{N} (\tilde{y}_i - \mu_i)^2 + \lambda H_{\ell_1}^{tf}(\mu).
\]

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

Our situation is different from the denoising example (with least squares loss) as 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 \( p_t \) will lead to a difficult non-convex optimization problem due to the nonlinear dependence of \( p_t \) on \( \theta_t \). Instead, we let the latent parameters \( t \mapsto \theta_t \) to be piecewise linear—this leads to a computationally tractable estimation framework based on convex optimization. Towards this end, we propose an adaption of Problem (2.4) by using the regularizer \( H(\theta) = H_{\ell_1}^{tf}(\theta) \):

\[
(3.2) \quad \text{minimize} \quad \sum_{t=1}^{N} \left( -y_t \theta_t + n_t \log(1 + \exp(\theta_t)) \right) + \lambda H_{\ell_1}^{tf}(\theta).
\]

Figure 8 shows the results of estimates obtained from some communication streams using the \( \ell_1 \)-trend filtering penalty. If the time points are not equally spaced, then this penalty needs to be modified—see for example [26] and also Section A.5.

**Computation.** The proximal gradient-stylized update (2.8) can be adapted to the setting described above with \( H(\theta) = H_{\ell_1}^{tf}(\theta) \). However, unlike the fused \( \ell_0 \) and \( \ell_1 \) penalties used in Section 2.1, the current proximal operator can no longer be computed via dynamic programming. To this end we use existing specialized solvers for the \( \ell_1 \)-trend filtering problem for the least squares loss function — in particular, we found the specialized interior point solver\(^{16}\) of [26] to work quite nicely for the problem sizes encountered in this paper. We note however, that solving Problem (3.2) is computationally more demanding than the piecewise constant segmentation approach — the main difference is due to the proximal operator: Problem (2.11) can be solved much more efficiently than Problem (3.1). As Problem (2.4) is convex, the sequence (2.8) converges to a minimum of the optimization problem (note that the minimum exists under minor assumptions). The convergence rates outlined in (2.9) and (2.10) will also apply to this problem. If we set \( H(\theta) = H_{\ell_0}^{tf}(\theta) \), the resulting

\(^{16}\)We use the R package wrapper available from [https://github.com/hadley/l1tf](https://github.com/hadley/l1tf).
Figure 8. Figure showing the estimates obtained from Problem (2.4) with the $\ell_1$-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.

Problem (2.4) becomes a challenging nonconvex optimization problem – in this case, there is no analogue of the highly efficient dynamic programming implementation that was available for $H_{\ell_0}(\theta)$. Hence, in this paper, we limit our attention to the convex $\ell_1$-trend filtering regularizer.

Social Science Interpretation. The shapes of the estimated signal (see Figure 8) capture key differences between different kinds of events in the history of U.S. foreign relations — these would have been less obvious using the piecewise constant signal approximation framework of Section 2.2. Some crises, like the Cyprus coup, occur with no warning, but also make little difference in the longer-term level of attention and activity. Others, like the OPEC oil embargo, are similarly unexpected, but signal the beginning of a period marked by moments of heightened activity well over the previous baseline. Still others, like the Fall of Saigon, build to a climax, and then gradually subside. Finally, the rise of human rights as a concern for policymakers is gradual but seemingly inexorable. The taxonomy of these different patterns provide useful insights to a social scientist in identifying and classifying different kinds of events. There are a few broad categories of patterns: from the
sudden and unexpected, to the gradual rise and fall, to longer-term trends—these patterns can help social scientists develop a taxonomy of historical events to gain a deeper understanding of heterogenous data.

4. Related work (event detection literature). We present a brief contextualization of our work in regard to the event detection literature within the computer science/data mining community. Event detection in text communication streams such as news or social media platforms (e.g., Twitter) is a fairly rich area of research. See for example, the nice review by [1] for an overview of event detection in Twitter communication streams. Starting with the seminal work of [27], important contributions have been made in the field of feature-pivot techniques, where one attempts to identify context-specific words that depict a sharp rise in frequency as an event emerges. An important line of research in this area (see for example [14, 20]) focuses on how to understand contributions from different words towards describing an event. A key focus in this line of work is an elaborate design of specialized features to describe the content of messages (keywords, hashtags, advanced text/context based features, etc), interactions among users, etc. For many of these approaches text-processing methods play an important role. It is important to note that the dataset we analyze in this paper and its application context is different from event detection in Twitter streams — the corpus we study does not have rich text-based data as in the case of modern Twitter applications. So we present a preliminary analysis of text-based data for our problem in Section 5.1. The communication streams we study are naturally characterized by TAGS, which were defined by the State Department when they were entered into the system. Our emphasis is on the use of statistical methods that make our models and results interpretable to a social scientist—this encourages a synergy between both statistics and social science perspectives (See Section 5 for a discussion). As mentioned by [1], many of the complex methods used in the context of analyzing Twitter streams are based on an ad hoc selection of thresholds—this is understandable given the complexity of the problem. In contrast, such heuristics are mostly avoided in our methods due to the nature of the statistical models pursued in this paper.

5. Concluding remarks and discussion. In this paper, we present statistical methods to analyze diplomatic cable communications during 1973-1977 recently made available by the U.S. National Archives. The complexity and heterogeneity of different historical events present themselves in the form of various geometric patterns in the communication streams. An important challenge of this work has been in identifying and proposing a suitable suite of statistical tools useful to a social scientist to glean insights from this newly available data. Our study focuses on the TAGS-specific communication streams; and understanding how geometric characteristics of these streams relate to events of historical significance.

We present a global testing framework to identify which among potentially thousands of communication streams exhibit interesting statistical activity that merit further downstream analysis. We propose signal segmentation methods based on $\ell_1$ and $\ell_0$ penalization to identify structural breaks aka jumps in the communication streams — the proposed algorithms are complementary to the area of change point
models in statistics; and are found to be much faster than existing implementations. We present a sample-splitting framework to perform statistical inference on the detected jumps; and discuss a simple but effective notion of combining jumps to bursts (following [27]) to obtain a rank ordering of them. Finally, we present extensions of piecewise constant signals to model the underlying process of communication stream traffic.

5.1. A social science perspective. Results available from our statistical analysis in some cases correspond to well known events; while in other cases they lead to curious findings that a social scientist might have missed in the absence of statistical tools such as the ones presented herein. The tools proposed here may be used (with suitable modifications) in other contexts, beyond the dataset analyzed in this paper. With the ever growing volume of digital content, it will become increasingly important for social scientists to devise a range of new methods to identify patterns and anomalies. When a corpus consists of hundreds of millions of emails – as is the case with the Obama White House files – in the absence of a suitable statistical framework, it may be challenging to filter out less interesting communications without obscuring unexpected and potentially important information. Historical events are quite heterogeneous, and social scientists need methods, such as the ones presented herein, that can help them prioritize which communications they should examine most closely.

An important characteristic of our framework—of particular appeal to a social scientist—is that it permits efficient algorithmic processing of large corpora. Our methodology does not require prior knowledge of the content of the communications—it is thus agnostic to specific biases that may be imposed by a practitioner. At the same time, our proposed framework is flexible enough to be able to capture very different kinds of events — our statistical methods also present a clear guideline of what geometric patterns these events might correspond to. Throughout our project, we have observed an interesting synergy between statistics and social science during the interpretation of the findings.

Our results in Section 2.3 and Section 3 suggests how a social scientist may use our framework to identify and classify different kinds of events. Some of these events can be described by jumps (piecewise constant signals), while others are better explained by piecewise linear segments. While each one is unique, there are clearly different classes, from the sudden and unexpected, to the gradual rise and fall, to longer-term trends, to cyclical patterns. Categorizing these different phenomena provides a useful heuristic in analyzing heterogeneous data, and could help social scientists develop a taxonomy of historical events. Developing a comprehensive statistical framework to perform shape-based grouping of these taxonomies is an interesting topic for future research.

Our framework is perhaps most important in supporting a new, more inductive approach to historical analysis, where it may not be possible to know in advance what communications should be examined more closely. Since the “archive” is increasingly digital, and archivists are no longer able to create traditional finding aids to guide researchers, it is increasingly difficult to decide which events or trends were most important. Our framework makes it possible to instead start with the data, and
then conduct a statistical analysis of communications that deviate from established communications patterns.

Inductive and deductive approaches are not mutually exclusive. For instance, we asked a historian who had just published a history of the 1970s to independently identify and rank the most important events. There was substantial overlap between his list and the one generated by our methodology. But there were also interesting differences, which raised what he called “crucial questions” about how we judge historical significance:

“Is that most vital quality to be assessed only in [the] perspective of hindsight, or can we use quantitative aggregation of contemporary data to achieve novel perspectives?”

Our framework can already help researchers meet the challenge of exponentially larger archives and do what machines can never do: interpret complex data, assess historical significance, and determine what this history really means for the present and the future.

Using text data to understand regime changes. While our work has focused on cable metadata and communication volume, similar methods could be applied to textual data. Unfortunately, the text is unavailable for many thousands of the cables in this corpus, either because State Department storage systems failed to preserve it, or because only the metadata has been declassified for cables that have been deemed to contain sensitive information. But we performed a small scale analysis to explore what might be learned from what data is available.

We studied communications relating to several regime changes in the 1970s, by using the name of the incoming leader as a guiding feature. In this analysis we looked at the weekly proportion of cables with the country in question (CY for Cyprus, CI for Chile, AR for Argentina, and PK for Pakistan) which also contain the name of the leader or leaders of the coup in the text (SAMPSON for CY, PINOCHET for CI, VIDEĽA or GUZZETTI for AR, and ZIA-UL-HAQ or ZIA for PK). Documents with unavailable text data were not included in the analysis. We applied our framework to detect potential change points (by restricting the number of jumps to at most two). Our results appear in Figure 11. In three of the four cases we examined, Pakistan, Cyprus, and Argentina, our framework clearly shows the change in communications patterns that corresponds with a coup. This is as opposed to, for example, South Vietnam, where the peak corresponding to the fall of Saigon is preceded by a long period of increasing activity. Although all three do have cables mentioning the coup leader before the event, these mentions are not persistent enough to shift the location.

Chile presents an interesting exception, in which the change point appears well after the regime change. There is increased interest in Pinochet leading up to it, but U.S. diplomats did not recognize he was a key figure in military plotting against the Allende government. The Nixon administration faced accusations of backing the coup once it succeeded, leading it to be cautious initially in contacts with the new leaders. That did not start to change until the spring of 1974 [17].

Even though the textual data for the cable corpus is incomplete, rich textual data is available for other corpora necessitating an in-depth analysis of text processing
methods. There is already a large body of work in the context of event detection in Twitter data [1] using text processing tools. We leave the combination of statistical analysis of communications streams and techniques from natural language processing as topics of future research.

Acknowledgements. R. Mazumder was supported by ONR (grant # N000141512342) and NSF-IIS (grant # 1718258); M. Connelly was supported by NSF-RIDIR (grant # 1637159). The authors will like to thank: David Blei, David Madigan, Shawn Simpson for helpful comments and encouragement; and Raymond Hicks for his help in providing us access to the text data used in Section 5. The authors will also like to thank the workshop participants of “Famine and Feast–International Historical Research in the Digital Age” (London, UK; 2015); and seminar participants at the US Census Bureau (2018) for comments on the work. Preliminary results from this article appeared in a media article on buzzfeed.com.

APPENDIX A: APPENDIX (SUPPLEMENTARY MATERIAL)

A.1. Simplifications for \( T \). Note that

\[
\hat{p}_{i}^{\text{ave}} = \frac{\sum_{j \in N(\Delta;i)} y_{j}}{\sum_{j \in N(\Delta;i)} n_{j}}.
\]

Under the null (note that we also assume independence across \( i \)) we have:

\[
\text{Var}(\hat{p}_{i}^{\text{ave}}) = \sum_{j \in N(\Delta;i)} \frac{\text{Var}(y_{j})}{\left( \sum_{j \in N(\Delta;i)} n_{j} \right)^{2}} = \frac{p_{H_{0}}(1 - p_{H_{0}})}{\left( \sum_{j \in N(\Delta;i)} n_{j} \right)}
\]

Above, \( p_{H_{0}} \) denotes the value of \( p \) under the null. Now, we replace \( p_{H_{0}} \) by an estimate under the null, i.e.,

\[
\hat{p}_{H_{0}} = \frac{\sum_{j=1}^{N} y_{i}}{\sum_{j=1}^{N} n_{i}}.
\]

This leads to:

\[
\hat{\sigma}_{t}^{2} = \frac{\hat{p}_{H_{0}}(1 - \hat{p}_{H_{0}})}{\sum_{j \in N(\Delta;t)} n_{j}},
\]

which is used in the computation of \( T_{t} \) and \( T \) in (2.1).

A.2. Sensitivity of global testing results to \( \Delta \). Table 2 presents sensitivity analysis of the global testing results (Section 2.1) wrt the window width \( \Delta \).

A.3. Comparison with PELT. Below we compare the results of our method (i.e., the \( \ell_{0} \) segmentation approach for Problem (2.4)) versus PELT [25] (from R package changepoint). As the latter does not provide functionality for the Binomial likelihood, we used a signal with raw proportions and the function \texttt{cpt.mean} to detect change points. We study the locations of the change points identified by both

---

17 Article www.buzzfeed.com/josephbernstein/can-a-computer-algorithm-do-the-job-of-a-historian?
Table showing how many TAGS-specific cables survive at different significance levels of \( T \) test (2.1) for different values of \( \Delta \), with 0.0001 being the smallest threshold in these experiments. This is out of the first 1000 TAGS, which roughly correspond to the TAGS with more than 50 total cables. There is a big overlap among the cables with small p-values (less than 0.01) across the different \( \Delta \) values — hence, the downstream analysis of these cables remains unaffected. A choice of \( \Delta = 2 \) corresponds to localized activity within a time-window of 1 week, and \( \Delta = 10 \) corresponds to activity of nearly a month. For the purpose of this application, we recommend taking values of \( \Delta \) within this range—in our experiments, we took \( \Delta = 5 \) to understand localized heightened activity within a span of 1 – 2 weeks.

<table>
<thead>
<tr>
<th>Significance level</th>
<th>( \Delta = 2 )</th>
<th>( \Delta = 5 )</th>
<th>( \Delta = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt; 0.1)</td>
<td>863</td>
<td>914</td>
<td>962</td>
</tr>
<tr>
<td>(&lt; 0.01)</td>
<td>651</td>
<td>768</td>
<td>869</td>
</tr>
<tr>
<td>(&lt; 0.001)</td>
<td>469</td>
<td>622</td>
<td>754</td>
</tr>
<tr>
<td>(&lt; 0.0001)</td>
<td>317</td>
<td>509</td>
<td>638</td>
</tr>
</tbody>
</table>

Table 2

Comparison between our proposed algorithm for \( \ell_0 \)-segmentation and the PELT algorithm (R package changepoint) for TAGS VS. We look at the locations of the change points by fixing the number of change points or jumps (shown in the first column). We can see that the solutions are not exactly the same, but are very similar. Rows for which the change point locations detected by PELT and our proposal are different, are indicated by a bold-font representation of the number of change points.

<table>
<thead>
<tr>
<th># jumps</th>
<th>method</th>
<th>jump locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>PELT</td>
<td>441 479 596</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>441 479 596</td>
</tr>
<tr>
<td>5</td>
<td>PELT</td>
<td>441 457 472 505 602</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>441 457 472 506 596</td>
</tr>
<tr>
<td>6</td>
<td>PELT</td>
<td>441 457 472 505 596 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>441 457 472 506 596 853</td>
</tr>
<tr>
<td>7</td>
<td>PELT</td>
<td>383 441 457 472 505 596 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>441 472 479 506 596 853</td>
</tr>
<tr>
<td>8</td>
<td>PELT</td>
<td>51 383 441 457 472 505 596 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>383 441 457 472 479 506 596 853</td>
</tr>
<tr>
<td>10</td>
<td>PELT</td>
<td>51 383 441 457 461 462 472 505 596 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>383 441 457 461 462 472 479 506 596 853</td>
</tr>
<tr>
<td>11</td>
<td>PELT</td>
<td>51 383 441 457 461 462 472 479 506 596 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 441 457 461 462 472 479 506 596 853</td>
</tr>
<tr>
<td>12</td>
<td>PELT</td>
<td>51 383 402 430 441 457 461 506 596 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 402 430 441 457 461 462 472 479 506 596 853</td>
</tr>
<tr>
<td>14</td>
<td>PELT</td>
<td>51 383 402 430 441 559 601 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 402 430 441 559 601 853</td>
</tr>
<tr>
<td>15</td>
<td>PELT</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td>16</td>
<td>PELT</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td>18</td>
<td>PELT</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td>19</td>
<td>PELT</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
<td>51 383 402 430 457 559 601 853</td>
</tr>
</tbody>
</table>

Table 3

Comparison between our proposed algorithm for \( \ell_0 \)-segmentation and the PELT algorithm (R package changepoint) for TAGS VS. We look at the locations of the change points by fixing the number of change points or jumps (shown in the first column). We can see that the solutions are not exactly the same, but are very similar. Rows for which the change point locations detected by PELT and our proposal are different, are indicated by a bold-font representation of the number of change points.
for VS (see Figure 3). The change point locations are given in terms of their location within the stream, as opposed to the corresponding date, for ease of comparison. Note that PELT minimizes a penalized least squares objective function as opposed to using the Binomial likelihood used in our method. Nonetheless the change points detected are quite consistent between the two methods—suggesting that our method is finding very good solutions. When the estimated jumps are found to be different, our algorithm was found to obtain a better solution in terms of a smaller objective value, when compared to PELT.

![Figure 9](image)

Figure 9. 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 \( p \approx 0.33 \) and \( p \approx 0.42 \) 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 \( x \).)

### A.4. Discussion of local p-values on synthetic data.
We consider a synthetic example in Figure 9 (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.

### A.5. Problem (2.8) 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:

\[
H(\theta) = \sum_i |\theta_{i+1} - \theta_i| / \Delta_t
\]

where, \( \Delta_t \) 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 (2.8) needs to be modified to:

\[
\text{minimize}_{u \in \mathbb{R}^N} \frac{1}{2} \| u - \bar{u} \|^2 + \lambda \| Du \|_1,
\]

where, \( \| Du \|_1 = \sum_i |u_{i+1} - u_i| / \Delta_t \). More generally, if we consider the \( \ell_1 \)-trend filtering example with varying time intervals, we get an instance of Problem (A.1) with

\[
\| Du \|_1 := \sum_t \left| \left( \frac{u_{t+2} - u_{t+1}}{\Delta_{t+1}} - \frac{u_{t+1} - u_t}{\Delta_t} \right) \right| \frac{1}{\Delta_t}.
\]
We use the Alternating Direction Method of Multipliers (ADMM) procedure [7] which performs the following decomposition: \( \alpha = Du \) and obtains the Augmented Lagrangian:

\[
\mathcal{L}(u, \alpha; \nu) = \frac{1}{2}\|u - \bar{u}\|^2_2 + \lambda'\|\alpha\|_1 + (\alpha - Du, \nu) + \frac{\rho}{2}\|\alpha - Du\|^2_2,
\]

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

\[
\begin{align*}
  u &\leftarrow \arg\min_u \mathcal{L}(u, \alpha; \nu) \\
  \alpha &\leftarrow \arg\min_\alpha \mathcal{L}(u, \alpha; \nu) \\
  \nu &\leftarrow \nu + \rho(\alpha - Du),
\end{align*}
\]

(A.2)

where, in the update wrt \( u \) other variables remain fixed, and the same applies for the update wrt \( \alpha \). We refer the reader to [7] for details pertaining to the convergence of this algorithm and choices of \( \rho \). We note that the update wrt \( u \) in display (A.2) can be solved quite easily via solving a system of linear equations:

\[
u \leftarrow (\rho D' D + I)^{-1}(\bar{u} + D' \nu + \rho D' \alpha)\]

Note that \((\rho D' D + I)\) 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) \([8, 26]\) – furthermore the inverse can be computed once (at the onset) as the matrix does not change across iterations. The update wrt \( \alpha \) in display (A.2) requires a solving the following problem:

\[
\alpha \leftarrow \arg\min_\alpha \frac{\rho}{2}\|\alpha - z\|^2_2 + \lambda'\|\alpha\|_1,
\]

where, \( z = (Du - \nu / \rho) \). A solution to the above problem is given by the familiar soft-thresholding \([18]\) operation where, \( \alpha_i = \text{sgn}(z_i) \max\{|z_i| - \lambda'/\rho, 0\} \). The sequence of updates in (A.2) are performed till some form of convergence criterion is met [7].
Fig 10. Communication streams with different significance scores in the spirit of the global testing framework presented in Section 2.1. Top row shows the series where the p-values are larger than 0.1—this is similar to TAGS FI (for Finland). The second row has p-values in 0.1-0.001 (such as for OSCI, scientific grants). The third and fourth rows show series that seem to have a high degree of intense activity: all p-values are smaller than 0.001. This includes, for example, SREF (for refugees) and SF (for South Africa). Recall that a strong deviation from the null model under the global testing framework does not necessarily imply a communication stream with a significant change point.
FIG 11. Segmentation based on textual features. We consider weekly proportion of cables with the country in question (CY for Cyprus, CI for Chile, AR for Argentina, and PK for Pakistan) which also contain the name of the leader or leaders of the coup in the text (SAMPSON for CY, PINOCHE for CI, VIDELA or GUZZETTI for AR, and ZIA-UL-HAQ or ZIA for PK). The jumps located by our $\ell_0$-segmentation approach, mostly correspond to regime changes denoted by vertical lines—see text for details.
REFERENCES


