SUPPLEMENT TO "MINING EVENTS WITH DECLASSIFIED DIPLOMATIC DOCUMENTS" BY GAO, GOETZ, CONNELLY AND MAZUMDER

1. Simplifications for $\mathcal{T}$. Note that

$$\widehat{p_i^{\text{ave}}} = \sum_{j \in \mathrm{N}(\Delta;i)} y_j / \sum_{j \in \mathrm{N}(\Delta;i)} n_j.$$

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

$$\mathrm{Var}(\widehat{p_i^{\text{ave}}}) = \sum_{j \in \mathrm{N}(\Delta;i)} \mathrm{Var}(y_j)/(\sum_{j \in \mathrm{N}(\Delta;i)} n_j)^2 = \frac{p_{H_0}(1-p_{H_0})}{(\sum_{j \in \mathrm{N}(\Delta;i)} n_j)}$$

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.,

$$\hat{p}_{H_0} = \sum_{j=1}^{N} y_i / \sum_{j=1}^{N} n_i.$$

This leads to:

$$\hat{\sigma}_t^2 = \frac{\hat{p}_{H_0}(1-\hat{p}_{H_0})}{\sum_{j \in \mathrm{N}(\Delta;t)} n_j},$$

which is used in the computation of $\mathcal{T}_t$ and $\mathcal{T}$ in the main paper.

Significance level	$\Delta = 2$	$\Delta = 5$	$\Delta = 10$
< 0.1	863	914	962
< 0.01	651	768	869
< 0.001	469	622	754
< 0.0001	317	509	638

TABLE 1

Table showing how many TAGS-specific cables survive at different significance levels of $\mathcal{T}$ test for different values of Δ, 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 Δ 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 Δ within this range—in our experiments, we took $\Delta = 5$ to understand localized heightened activity within a span of $1-2$ weeks.

2. Sensitivity of global testing results to Δ. Table 1 presents sensitivity analysis of the global testing results (Section 2.1 of the main paper) wrt the window width Δ.

# jumps	method	jump locations
3	PELT	441 479 596
	this paper	441 479 596
5	PELT	441 457 472 505 602
	this paper	441 457 472 479 596
6	PELT	441 457 472 505 596 853
	this paper	441 457 472 479 506 596
7	PELT	383 441 457 472 505 596 853
	this paper	441 457 472 479 506 596 853
8	PELT	51 383 441 457 472 505 596 853
	this paper	383 441 457 472 479 506 596 853
10	PELT	51 383 441 457 461 462 472 505 596 853
	this paper	383 441 457 461 462 472 479 506 596 853
11	PELT	51 383 441 457 461 462 472 479 506 596 853
	this paper	51 383 441 457 461 462 472 479 506 596 853
12	PELT	51 383 441 457 461 462 466 472 479 506 596 853
	this paper	51 383 441 457 461 462 466 472 479 506 596 853
14	PELT	51 383 402 430 441 457 461 462 466 472 479 506 596 853
	this paper	51 383 402 430 441 457 461 462 466 472 479 506 596 853
15	PELT	51 383 402 430 431 441 457 461 462 466 472 479 506 596 853
	this paper	51 383 402 430 441 457 461 462 466 472 479 506 553 596 853
16	PELT	51 383 402 430 431 441 457 461 462 466 472 479 505 559 601 853
	this paper	51 383 402 430 441 457 461 462 463 466 472 479 506 553 596 853
18	PELT	51 291 380 394 401 430 431 441 457 461 462 466 472 479 505 559 601 853
	this paper	51 291 383 394 402 430 441 457 461 462 463 466 472 479 506 553 596 853
19	PELT	51 291 380 394 401 430 431 441 457 461 462 463 466 472 479 505 559 601 853
	this paper	51 291 383 394 402 430 441 457 461 462 463 466 472 479 495 506 553 596 853

TABLE 2

Comparison between our proposed algorithm for ℓ_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.

3. Comparison with PELT. We compare the results of our method (i.e., the ℓ_0 segmentation approach for Problem (2.4) in main paper) versus PELT [4] (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 `cpt.mean` to detect change points. We study the locations of the change points identified by both methods for a sequence of regularization parameters on the communication stream for VS (see Table 2). 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.

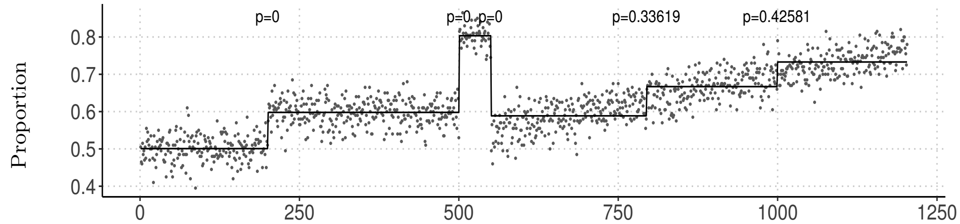

FIG 1. *Synthetic data: data description is the same as Figure 3 of the main paper, which contains three real jumps and a linear increasing trend. The first three detected jumps (from left to right) have small p-values (close to zero) – they correctly correspond to the jumps in the underlying signal. The other two potential jumps have p-values* ~ 0.33 *and* ~ 0.42 *respectively – these jumps are a consequence of the linear trend. (The notation* $p = x$ *is a shorthand for p-value being equal to* x*.)*

4. Discussion of local p-values on synthetic data. We consider a synthetic example in Figure 1 (with the same data as in Figure 3 of the main paper) – here we observe that the p-values (based on sample splitting) 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.

5. 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 for the ℓ_1-penalty needs to be modified to:

$$\underset{\mathbf{u} \in \Re^N}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{u} - \bar{\mathbf{u}}\|_2^2 + \lambda' \|D\mathbf{u}\|_1, \tag{5.1}$$

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 (5.1) 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 [1] 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:

$$
\begin{aligned}
\mathbf{u} &\leftarrow \arg\min_{\mathbf{u}} \ \mathcal{L}(\mathbf{u},\boldsymbol{\alpha};\boldsymbol{\nu}) \\
\boldsymbol{\alpha} &\leftarrow \arg\min_{\boldsymbol{\alpha}} \ \mathcal{L}(\mathbf{u},\boldsymbol{\alpha};\boldsymbol{\nu}) \\
\boldsymbol{\nu} &\leftarrow \boldsymbol{\nu} + \rho(\boldsymbol{\alpha} - D\mathbf{u}),
\end{aligned} \tag{5.2}
$$

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 [1] for details pertaining to the convergence of this algorithm and choices of ρ. We note that the update wrt $\mathbf{u}$ in display (5.2) can be solved quite easily via solving a system of linear equations:

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

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) [2, 5] – 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 (5.2) requires a solving the following problem:

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

where, $\mathbf{z} = (D\mathbf{u} - \boldsymbol{\nu}/\rho)$. A solution to the above problem is given by the familiar soft-thresholding [3] operation where, $\alpha_i = \operatorname{sgn}(z_i)\max\{|z_i| - \lambda'/\rho, 0\}$. The sequence of updates in (5.2) are performed till some form of convergence criterion is met [1].

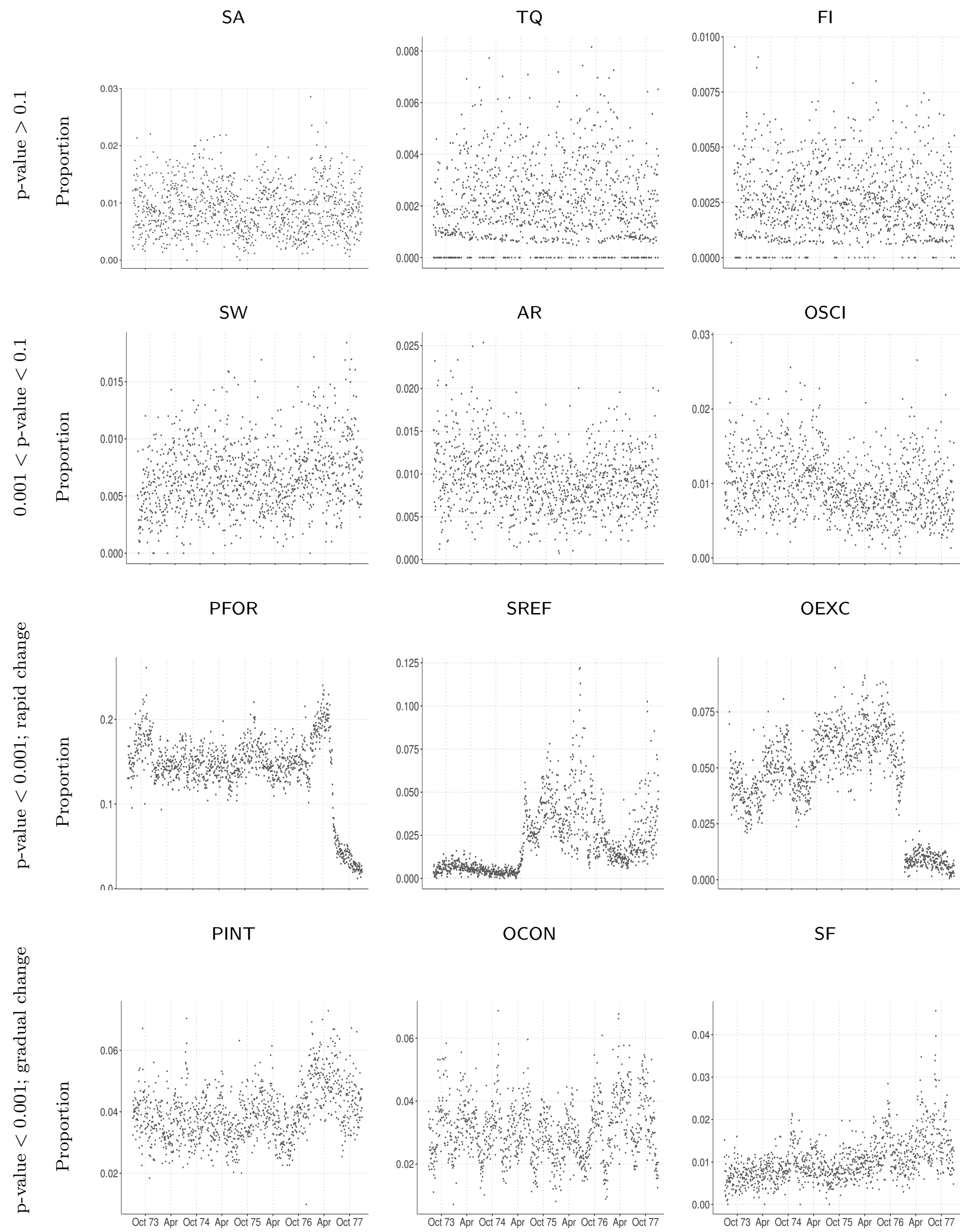

FIG 2. *Communication streams with different significance scores in the spirit of the global testing framework presented in Section 2.1 in main paper. Top row shows the series where the p-values are larger than 0.1—this is similar to TAGS FI (for Finland). The second row has p-values in 0.1-0.001 (such 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).*

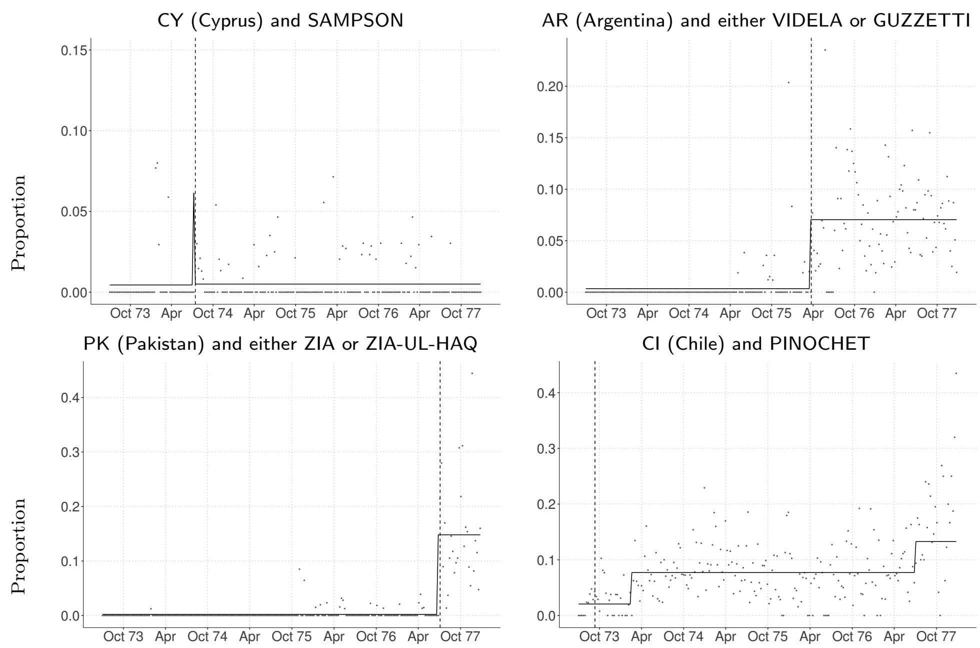

Fig 3. *Segmentation based on textual features. We consider weekly proportion of cables with the country in question (CY for Cyprus, CI for Chile, AR for Argentina, and PK for Pakistan) which also contain the name of the leader or leaders of the coup in the text (SAMPSON for CY, PINOCHET for CI, VIDELA or GUZZETTI for AR, and ZIA-UL-HAQ or ZIA for PK). The jumps located by our ℓ_0-segmentation approach, mostly correspond to regime changes denoted by vertical lines—see discussion in the main paper for details.*

References.

[1] 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.

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

[3] 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.

[4] Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

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