LA-UR-20-20324
Approved for public release; distribution is unlimited.

Title: Robust Bayesian change detection for cyber-security applications

Author(s): Hallgren, Karl Lars Yvon
Turcotte, Melissa

Intended for: Student presentation

Issued: 2020-01-13

Robust Bayesian change detection for cyber-security applications

Karl Hallgren

Mentor at LANL: Melissa Turcotte†
PhD supervisors: Nicholas Heard* and Niall Adams*

† Advanced Research in Cyber-Systems (A-4), Los Alamos National Laboratory
* Department of Mathematics, Imperial College London (UK)

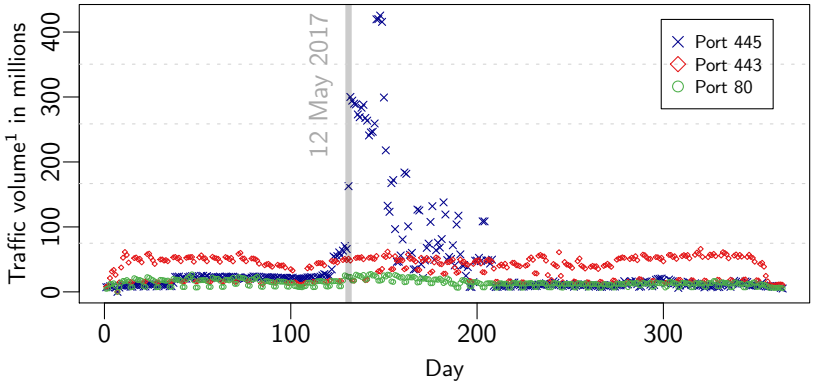
August, 2019

Los Alamos NATIONAL LABORATORY EST. 1943

Motivation for change detection

A cyber attack typically changes the behaviour of the target network.

For example, by exploiting a Windows vulnerability the 2017 WannaCry ransomware attack led to spikes of activity on port 445.



[1] recorded by a network router at Imperial College London

Los Alamos NATIONAL LABORATORY EST. 1943

Motivation for change detection

To detect the presence of a network intrusion, it can be informative to monitor for changes in the high-volume data sources which are collected inside an enterprise computer network

- NetFlow - summaries of connection between devices
- Authentication events
- Host event logs

Large data set derived from the operational network environment at LANL

- Unified Host and Network Data Set (Turcotte et al., 2017)
  `https://csr.lanl.gov/data/2017.html`

Los Alamos NATIONAL LABORATORY EST. 1943

Bayesian changepoint analysis

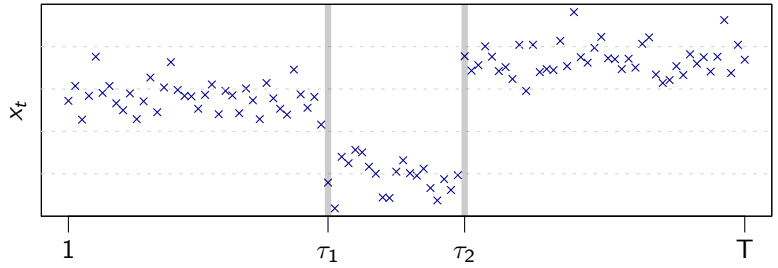
An unknown number k of changepoints, whose positions are denoted by $\tau_{1:k} = (\tau_1, \ldots, \tau_k)$, split the data $x_1, \ldots, x_T$ into $k+1$ segments such that within each segment j

$$x_t \overset{\text{iid}}{\sim} f(\cdot \mid \theta_j) \quad \forall t \in \{\tau_{j-1}, \ldots, \tau_j - 1\}$$

for some segment parameter θ_j.

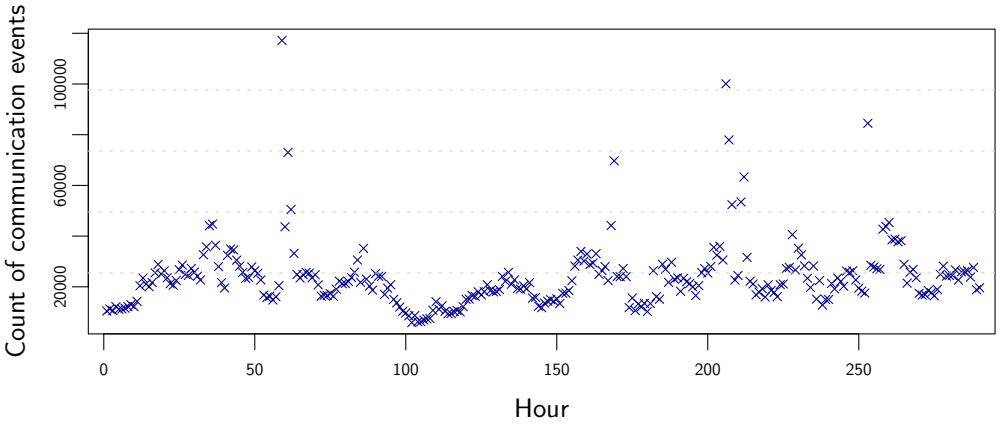
A priori each time point is assumed to be a changepoint with probability p so that

$$\pi(k, \tau_{1:k}) = p^k (1-p)^{T-1-k} \propto \{p/(1-p)\}^k$$



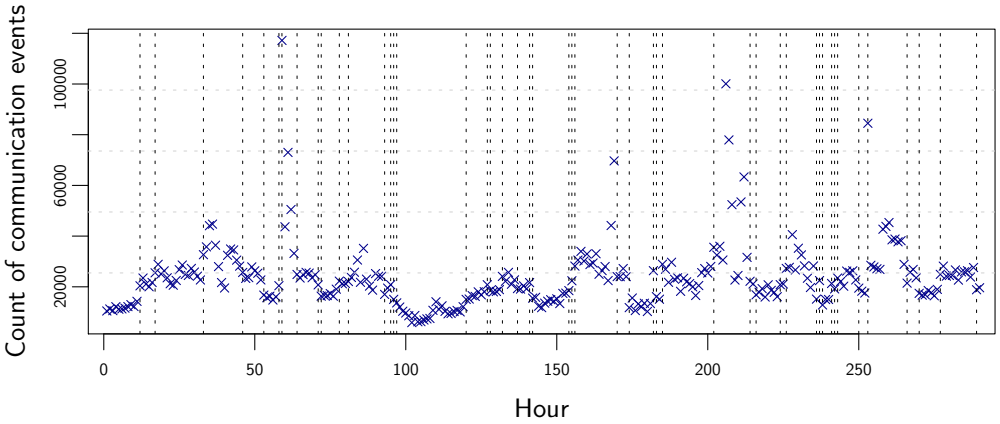
Los Alamos NATIONAL LABORATORY EST. 1943

Difficulties applying traditional change detection methods



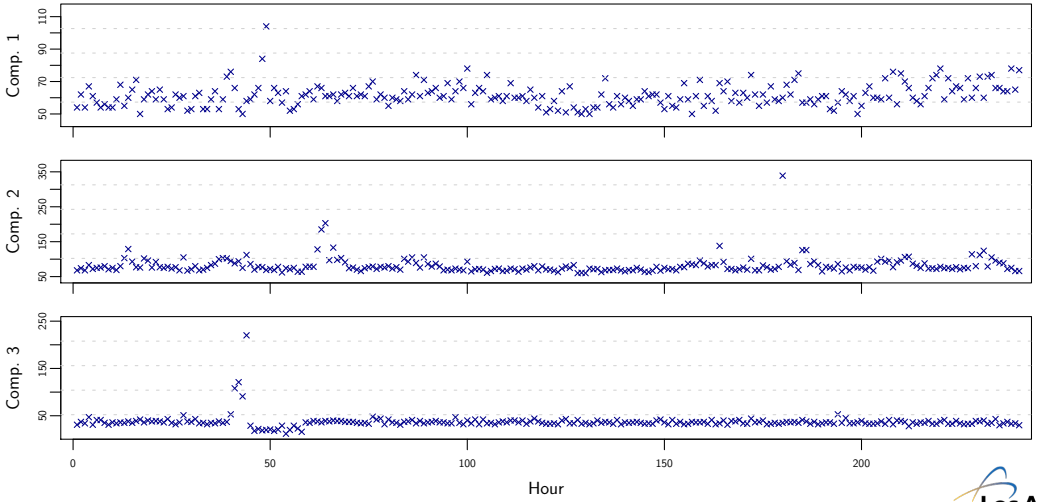
Los Alamos NATIONAL LABORATORY EST. 1943

Difficulties applying traditional change detection methods



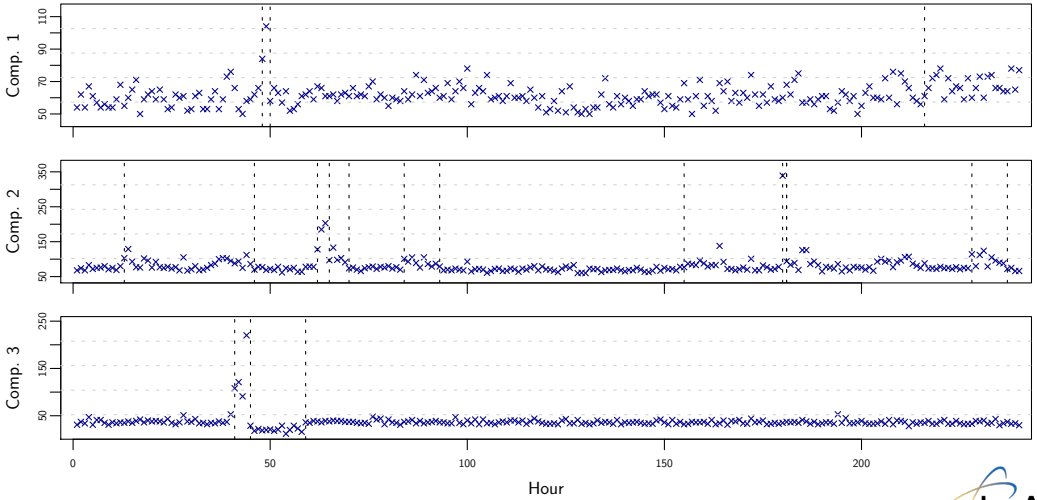
Los Alamos NATIONAL LABORATORY EST. 1943

Difficulties applying traditional change detection methods



Los Alamos NATIONAL LABORATORY EST. 1943

Difficulties applying traditional change detection methods



Los Alamos NATIONAL LABORATORY EST. 1943

Difficulties applying traditional change detection methods

Most traditional changepoint detection methods will fail to characterise what cyber-security analysts mean by a change and consequently fit many more changepoints than preferable.

1. Classical changepoint models are not robust to seasonal variations, gradual drifts and other *normal* dynamic phenomena one may observe in cyber data.

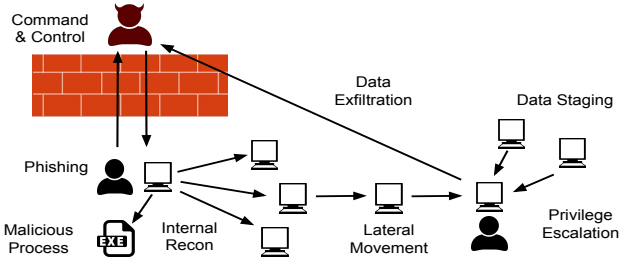
2. Suspicious changes in the behaviour of a computer within a network can be observed under non-attack conditions.

Los Alamos NATIONAL LABORATORY EST. 1943

Therefore it is relevant to develop methods to reduce false positive detections in changepoint analysis for cyber-security applications.

1. Building changepoint models which are robust to *normal* dynamic phenomena one may observe in cyber data, yet still tractable.

2. Combining evidence from multiple sources to identify patterns of changes which are a priori likely to correspond to the kill chain of an attack (Hutchins et al., 2011; Sexton et al., 2015).



Los Alamos NATIONAL LABORATORY EST. 1943

Part 1. Building robust segment models

Relaxing the assumption of exchangeability

For each segment, assuming the data are iid conditional on some segment-specific parameter is equivalent to assuming that the data are infinitely exchangeable by De Finetti's representation theorem (Bernardo and Smith, 1993).

A potentially infinite sequence $x_1, x_2, \ldots$ is said to be *infinitely exchangeable* under a probability measure F if, for every n and any set of indices $I_n = \{i_1, \ldots, i_n\}$, we have

$$F(x_{i_1}, \ldots, x_{i_n}) = F(x_{\sigma(i_1)}, \ldots, x_{\sigma(i_n)}),$$

for any permutation σ on I_n.

$\longrightarrow$ This assumption of complete symmetry within segments is too restrictive for cyber-security applications.

Los Alamos NATIONAL LABORATORY EST. 1943

Segment models for non-exchangeable data

It is of interest to build segment models which may admit exchangeability but also other weak forms of dependence, so that the corresponding changepoint models are more flexible (yet still tractable).

Los Alamos NATIONAL LABORATORY EST. 1943

Moving-sum changepoint model

Within a generic segment, suppose the observed data $x_1, \ldots, x_n$ satisfy

$$x_t = \sum_{i=0}^{m} y_{t-i},$$

where (y_t) are latent random variables such that

$$y_t \overset{\text{iid}}{\sim} f(\cdot \mid \theta, m)$$

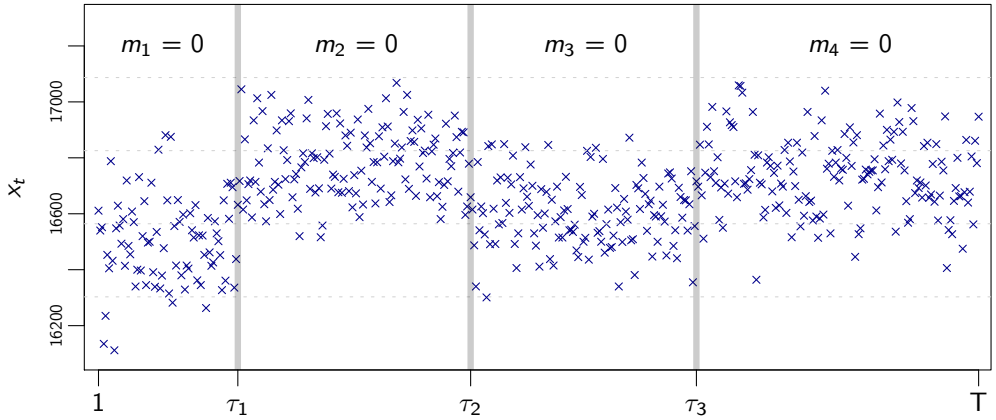
for some *unknown* segment-specific parameters $\theta \in \Theta$ and $m \geqslant 0$.

This approach allows m-dependence within segments, whilst maintaining a desired marginal distribution in the class of convolution-closed infinitely divisible distributions.

- E.g., if $y_t \overset{\text{iid}}{\sim} \text{Poisson}(\theta/(m+1))$ then $x_t \sim \text{Poisson}(\theta)$ with m-dependence.

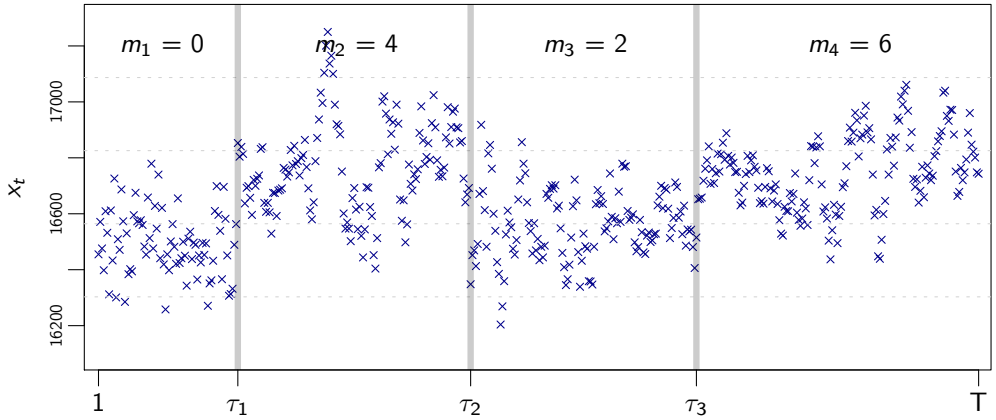
Los Alamos NATIONAL LABORATORY EST. 1943

Simulations from the moving-sum changepoint model



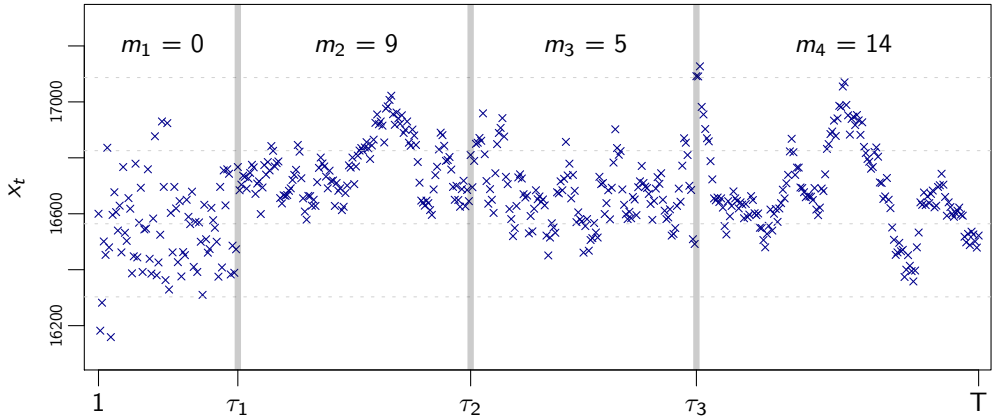
Los Alamos NATIONAL LABORATORY EST. 1943

Simulations from the moving-sum changepoint model



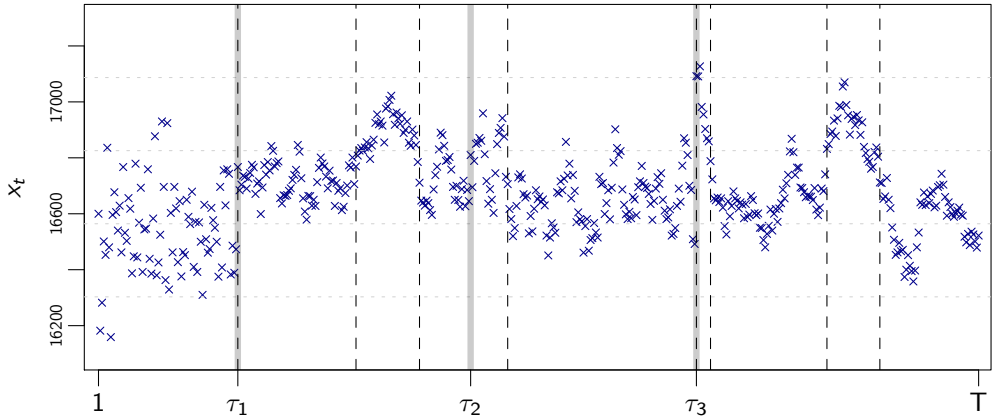
Los Alamos NATIONAL LABORATORY EST. 1943

Simulations from the moving-sum changepoint model



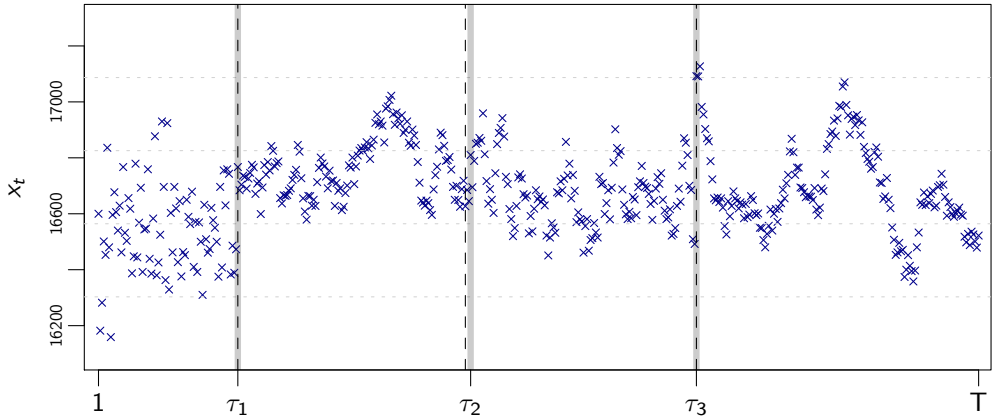
Los Alamos NATIONAL LABORATORY EST. 1943

Simulations from the moving-sum changepoint model



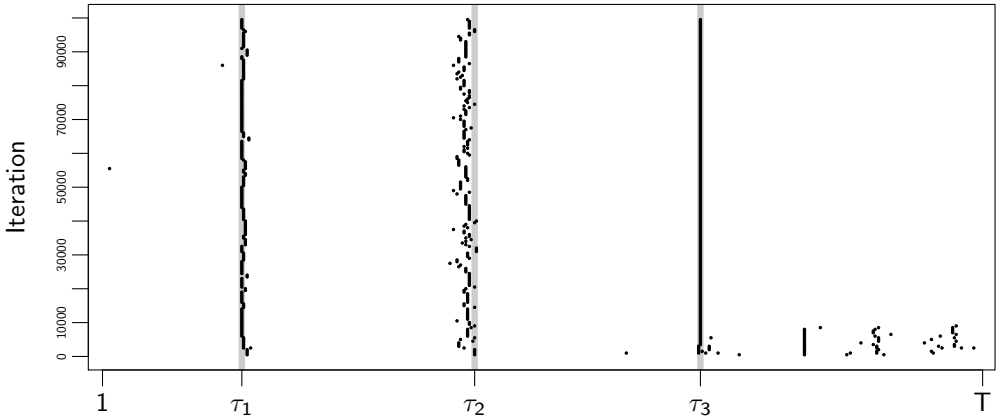
Los Alamos NATIONAL LABORATORY EST. 1943

Simulations from the moving-sum changepoint model



Los Alamos NATIONAL LABORATORY EST. 1943

Simulations from the moving-sum changepoint model



Los Alamos NATIONAL LABORATORY EST. 1943

Part 2. Towards kill chain detection

Inadequate prior distribution for the changepoints

The aim is to perform changepoint analysis on L time series of length T, denoted by $\boldsymbol{y} = (y_{\ell,t})$.

Let $\boldsymbol{s} = (s_{\ell,t})$ be a matrix indicating the positions of the changepoints,

$$s_{\ell,t} = \begin{cases} 1 & \text{if } t \text{ is a changepoint for the } \ell\text{-th time series} \\ 0 & \text{otherwise.} \end{cases}$$

A priori assumed that

$$\pi(\boldsymbol{s}) = \prod_{\ell,t} \left\{ p^{s_{\ell,t}} (1-p)^{1-s_{\ell,t}} \right\} \propto \prod_{\ell,t} \left\{ \frac{p}{1-p} \right\}^{s_{\ell,t}}$$

One issue is that $\pi(\boldsymbol{s}) = \pi(\boldsymbol{s}')$, where

$$\boldsymbol{s} = \begin{pmatrix} 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{s}' = \begin{pmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Los Alamos NATIONAL LABORATORY EST. 1943

Encoding prior knowledge on cyber-attack propagation

Cyber-analysts can specify prior knowledge on cyber-attack propagation by defining the neighbourhoods of the indices ℓ and t, denoted by $\partial(\ell)$ and $\partial(t)$, respectively.

For example, if $\ell, \ell_1, \ell_2, \ldots, \ell_L$ denote the indices of time-series representing the process activity of computers which are known to be *peers* on the network, we may have

$$\{\ell_1, \ldots, \ell_L\} \subset \partial(\ell) \quad \text{and} \quad \partial(t) = \{t-1, t, t+1\}.$$

To encode the structure in the neighbourhood of (ℓ, t), define

$$w_{\ell,t} \equiv w_{\ell,t}(\boldsymbol{s}) = \left(1 + \sum_{\ell' \in \partial(\ell)} \mathbb{1}\{1 \in s_{\ell', \partial(t)}\}\right) \Big/ (1 + |\partial(\ell)|)$$

For example,

$$w_{\ell,t} = 1/3 \quad \text{if} \quad \boldsymbol{s} = \begin{pmatrix} 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \textcircled{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 \end{pmatrix}$$

Encoding prior knowledge on cyber-attack propagation

Cyber-analysts can specify prior knowledge on cyber-attack propagation by defining the neighbourhoods of the indices ℓ and t, denoted by $\partial(\ell)$ and $\partial(t)$, respectively.

For example, if $\ell, \ell_1, \ell_2, \ldots, \ell_L$ denote the indices of time-series representing the process activity of computers which are known to be *peers* on the network, we may have

$$\{\ell_1, \ldots, \ell_L\} \subset \partial(\ell) \quad \text{and} \quad \partial(t) = \{t-1, t, t+1\}.$$

To encode the structure in the neighbourhood of (ℓ, t), define

$$w_{\ell,t} \equiv w_{\ell,t}(\boldsymbol{s}) = \left(1 + \sum_{\ell' \in \partial(\ell)} \mathbb{1}\{1 \in s_{\ell', \partial(t)}\}\right) \Big/ (1 + |\partial(\ell)|)$$

For example,

$$w_{\ell,t} = 3/3 \quad \text{if} \quad \boldsymbol{s} = \begin{pmatrix} 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \textcircled{1} & 0 & 0 & 0 & 0 \end{pmatrix}$$

Los Alamos NATIONAL LABORATORY EST. 1943

Encoding prior knowledge on cyber-attack propagation

Cyber-analysts can specify prior knowledge on cyber-attack propagation by defining the neighbourhoods of the indices ℓ and t, denoted by $\partial(\ell)$ and $\partial(t)$, respectively.

For example, if $\ell, \ell_1, \ell_2, \ldots, \ell_L$ denote the indices of time-series representing the process activity of computers which are known to be *peers* on the network, we may have

$$\{\ell_1, \ldots, \ell_L\} \subset \partial(\ell) \quad \text{and} \quad \partial(t) = \{t-1, t, t+1\}.$$

To encode the structure in the neighbourhood of (ℓ, t), define

$$w_{\ell,t} \equiv w_{\ell,t}(\boldsymbol{s}) = \left(1 + \sum_{\ell' \in \partial(\ell)} \mathbb{1}\{1 \in s_{\ell', \partial(t)}\}\right) \Big/ (1 + |\partial(\ell)|)$$

For example,

$$w_{\ell,t} = 3/3 \quad \text{if} \quad \boldsymbol{s} = \begin{pmatrix} 0 & 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \textcircled{1} & 0 & 0 & 0 & 0 \end{pmatrix}$$

Los Alamos NATIONAL LABORATORY EST. 1943

Informative prior distribution for the changepoints

Consider a prior distribution for $\boldsymbol{s}$ which takes into account a priori knowledge on cyber-attack propagation,

$$\pi(\boldsymbol{s}) \propto \prod_{\ell,t} \left\{ \left(\frac{p}{1-p} \right) \phi_\lambda(w_{\ell,t}) \right\}^{s_{\ell,t}}$$

for some non-decreasing function $\phi_\lambda : (0,1] \longrightarrow (0,1]$ such that $\phi_\lambda(1) = 1$.

Now,

$$\pi(\boldsymbol{s})/\pi(\boldsymbol{s}') = \frac{\phi_\lambda^3(1)}{\phi_\lambda^2(2/3)\phi_\lambda(1/3)} \geqslant 0,$$

with

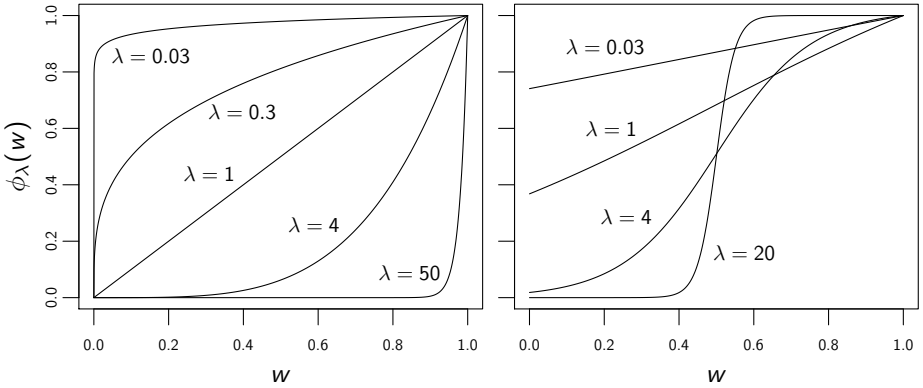
$$\boldsymbol{s} = \begin{pmatrix} 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{s}' = \begin{pmatrix} 0 & 0 & \mathbf{1} & 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Los Alamos NATIONAL LABORATORY EST. 1943

Flexible informative prior distribution for the changepoints

Possible choices for the link function ϕ_λ include:

$$\phi_\lambda(w) = w^\lambda \qquad \phi_\lambda(w) = \frac{1 + \exp(-\lambda)}{1 + \exp(-\lambda[2w - 1])}$$



Los Alamos NATIONAL LABORATORY EST. 1943

Inference via MCMC

Let $\boldsymbol{s}$ be binary matrix with $s_{\ell,t} = 0$, and let $\boldsymbol{s}'$ be identical to $\boldsymbol{s}$ but with $s_{\ell,t} = 1$.

Gibbs sampling is possible since the full conditional distribution of $s_{\ell,t}$ is available,

$$s_{\ell,t} \sim \text{Bernoulli}\left(\frac{\mathcal{Q}}{1+\mathcal{Q}}\right),$$

where

$$\mathcal{Q} = \frac{f(\boldsymbol{y} \mid \boldsymbol{s}')}{f(\boldsymbol{y} \mid \boldsymbol{s})} \frac{\pi(\boldsymbol{s}')}{\pi(\boldsymbol{s})},$$

but poor mixing and convergence issues are to be expected since the parameters will typically be highly correlated.

Los Alamos NATIONAL LABORATORY EST. 1943

Conclusion

Acknowledgement

Research presented in this talk was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory (New Mexico, USA) under project number 20180607ECR.

Los Alamos NATIONAL LABORATORY EST. 1943

References

Bernardo, J. M. and Smith, A. F. M. (1993). *Bayesian theory*. Wiley series in probability and mathematical statistics. John Wiley, New York; Chichester.

Hutchins, E., Cloppert, M., and Amin, R. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare and Security Research*, 1.

Sexton, J., Storlie, C. B., and Neil, J. (2015). Attack chain detection. *Statistical Analysis and Data Mining*, 8:353–363.

Turcotte, M. J. M., Kent, A. D., and Hash, C. (2017). Unified Host and Network Data Set. *ArXiv e-prints*. 1708.07518.

Los Alamos NATIONAL LABORATORY EST. 1943

Alternative representation of the moving-sum segment model

Recall $x_t = \sum_{i=0}^{m} y_{t-i}$ for all t. Equivalently,

$$y_t = y_{t-(m+1)} + x_t - x_{t-1}, \quad \text{for } t = 2, \ldots, n, \tag{1}$$

and $y_1 = x_1 - (y_{-m+1} + \cdots + y_0)$.

The first m latent variables $\gamma_{1:m} = (\gamma_1, \ldots, \gamma_m)$, with

$$\gamma_r = y_{-m+r}, \quad \text{for } r = 1, \ldots, m,$$

may be seen as the unknown initial conditions of a stochastic difference equation defined by (1).

Conditional on the $\gamma_{1:m}$, there is a one-to-one deterministic transformation, denoted by Υ, between $x_{1:n}$ and $y_{1:n}$ with unit Jacobian,

$$y_{1:n} = \Upsilon(x_{1:n} \,|\, \gamma_{1:m}),$$

which can be obtained explicitely by iterating (1).

Los Alamos NATIONAL LABORATORY EST. 1943

Conditional likelihood of the observed data

As a result, for each segment resulting from the changepoints, the conditional likelihood of the data

$$f(x_{1:n} \mid \theta, m, \gamma_{1:m}) = f(y_{1:n} \mid \theta, m, \gamma_{1:m}) \times 1.$$

is tractable, and therefore if we treat $(\theta, m, \gamma_{1:m})$ as unknown segment-specific parameters with prior density

$$\pi(\theta)\pi(m) \prod_{r=1}^{m} f(\gamma_r \mid \theta, m),$$

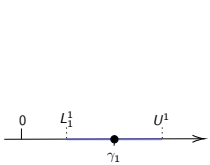
then an expression for the joint posterior distribution of the changepoints and the segment parameters is immediately available via Bayes' theorem.

Los Alamos NATIONAL LABORATORY EST. 1943

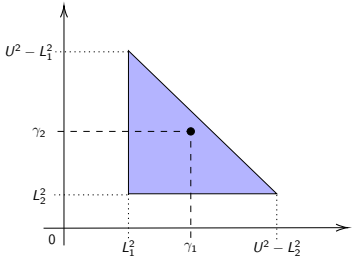
The segment parameter space

Let $\mathcal{Y}_m \equiv \mathcal{Y}_m(x_{1:n})$ denote the set of sequences $\gamma_{1:m}$ such that y_t belongs to the support of $f(y \mid m, \theta)$, for all $t = 1, \ldots, n$.

$\mathcal{Y}_m$ depends on the observed data when the support of $f(y \mid m, \theta)$ is bounded.



$m = 1$



$m = 2$

The structure of the constrained parameter space can be exploited to design a sampling strategy.

Los Alamos NATIONAL LABORATORY EST. 1943