Bursts Detection in Call Trains for Identifying Fraud in Telecommunications

Miguel Pebes-Trujillo*; Daniel Manrique-Vallier

Department of Statistics, Indiana University. Bloomington, Indiana, USA. Contact*: mpebestr@indiana.edu; http://pages.iu.edu/∼mpebestr/

1. Introduction

Standard worldwide telephony-contracts specify some prohibited activities that abuse the service and cause enormous financial losses in the industry, e.g. **scam, making autodialed calls, transmitting pre-recorded audio, or telemarketing**.

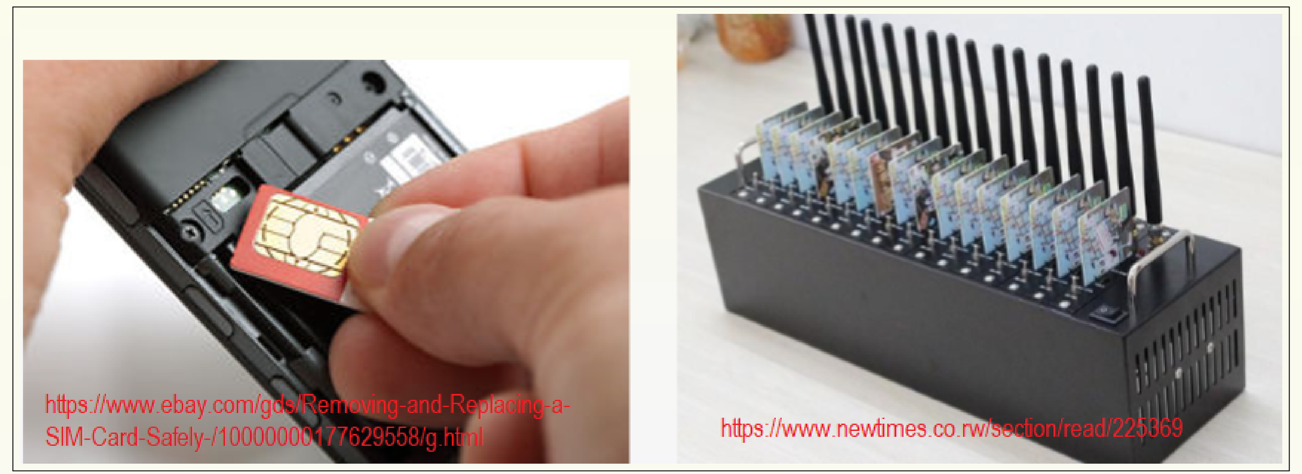

Fig. 1. Prohibitions involve the automated generation of calls.

Our **goal** is to build a tool to detect users engaged in abuse of services, considered in this context as fraud.

1.1 Call-Trains:

Companies store the users calls records (CDR's) for billing purposes, e.g.

userID	date-time	direction	secs	...
TLF_01	10/15/2018-13:04:00	outgoing	43	
TLF_01	10/19/2018-09:46:21	outgoing	209	
TLF_01	10/20/2018-11:31:08	outgoing	161	
TLF_01	10/25/2018-17:06:00	outgoing	45	
TLF_01	10/25/2018-17:08:28	outgoing	10	

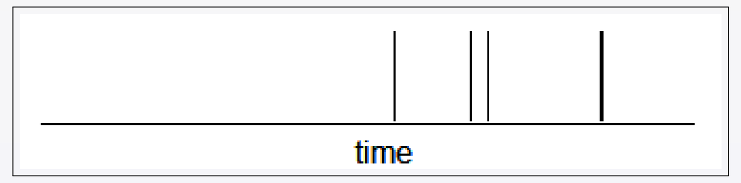

Fig. 2. We represent CDR's as *call-trains* (spikes over time).

1.2 Patterns of Automatic Calls:

fraudulent outgoing call-trains exhibit "bursty" behavior.

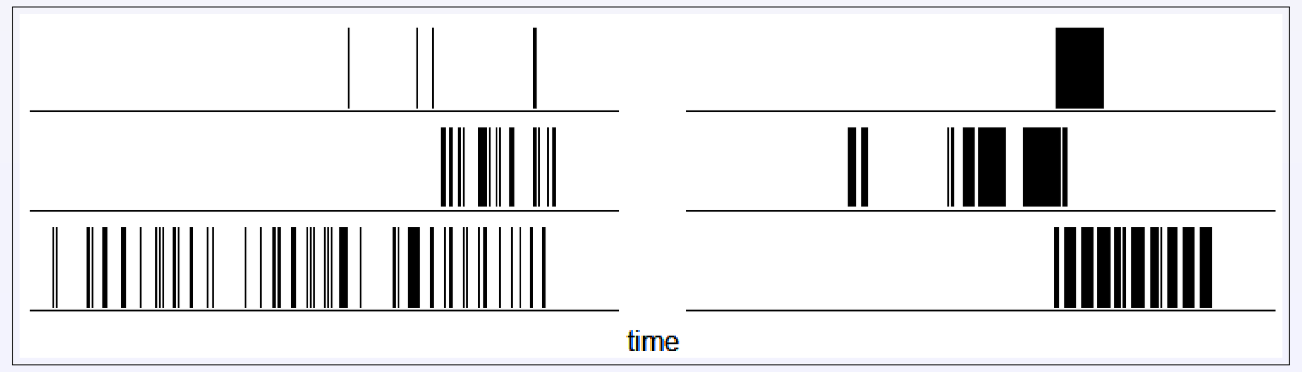

Fig. 3. Left: regular use. Right: fraudulent use.

1.3. Current Practice in Fraud Detection:

Mainly based on aggregated data, either as
-**descriptive statistics** (outliers & thresholds), or
-**classification methods** (building feature vectors over which supervised techniques are applied). *They fail when the user does not have a high calls-traffic.*

2. Our Method

2.1 Idea:

call-trains evolve according to either one or two different **latent** processes: *non-bursting (N) and bursting (B)*, which randomly alternate, e.g.

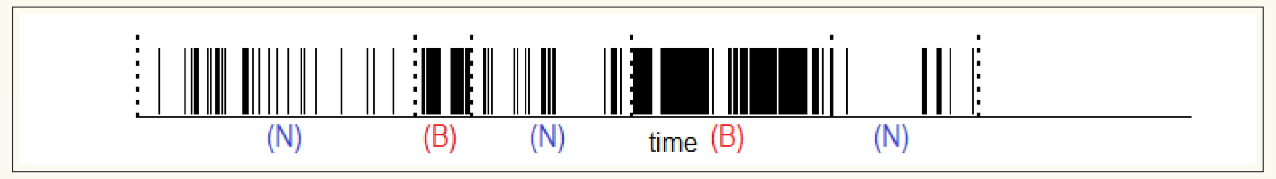

Fig. 4. Observable inter arrival times are governed by unobservable states.

A two-state machine $\boldsymbol{\mathcal{M}}$ represents this dynamics:

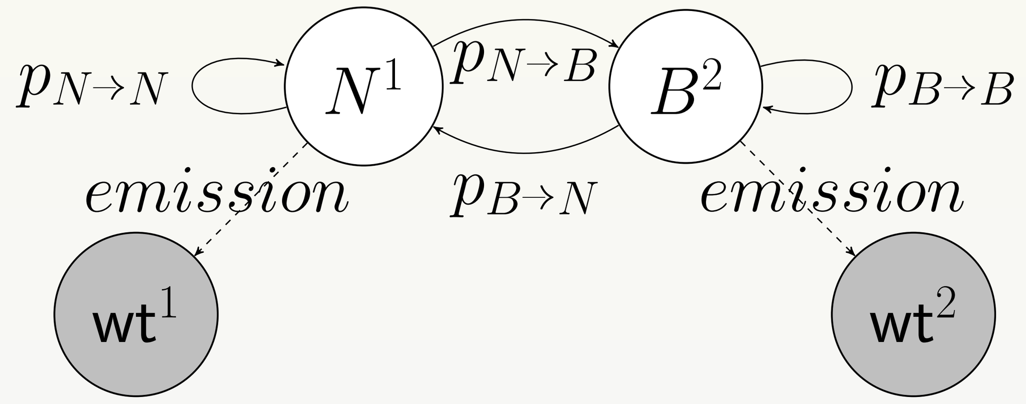

Fig. 5. Emitted inter-arrival times depends on its latent states, which evolve based on transition probabilities.

with time-homogeneous transition probabilities $\boldsymbol{\mathcal{A}}$,

$\boldsymbol{\mathcal{A}} = [p_{k\to l}] : k, l \in \{1,2\}, \sum_l p_{k\to l} = 1, k \in \{1,2\}.$

We model inter-arrival times to detect bursts in the call-train, viewed as a two-state stochastic process.

2.2 Model:

For each user i with k_i calls, we unfold $\boldsymbol{\mathcal{M}}_i$ and define the random variables X_t, the tth inter-arrival time; and Z_t, its latent state. Assumptions $X_t \perp\!\!\!\perp Z_{-t}|Z_t$ and $p(Z_t|Z_1,...,Z_{t-1}) = p(Z_t|Z_{t-1})$ lead us to the likelihood of a *Hidden Markov Model (HMM)*,

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\mathcal{A}}) = p(z_1|\boldsymbol{\pi}) \prod_{t=1}^{k_i} p(x_t|z_t, \boldsymbol{\theta}) \prod_{t=2}^{k_i} p(z_t|z_{t-1}, \boldsymbol{\mathcal{A}});$$

where $\boldsymbol{\pi} = \{\pi = P(Z_1 = 1), 1-\pi\}$, $\boldsymbol{\theta} = \{\lambda_1, \lambda_2\}$, $X_t|Z_t = 1 \sim Exp(\lambda_1)$ and $X_t|Z_t = 2 \sim Exp(\lambda_2)$.

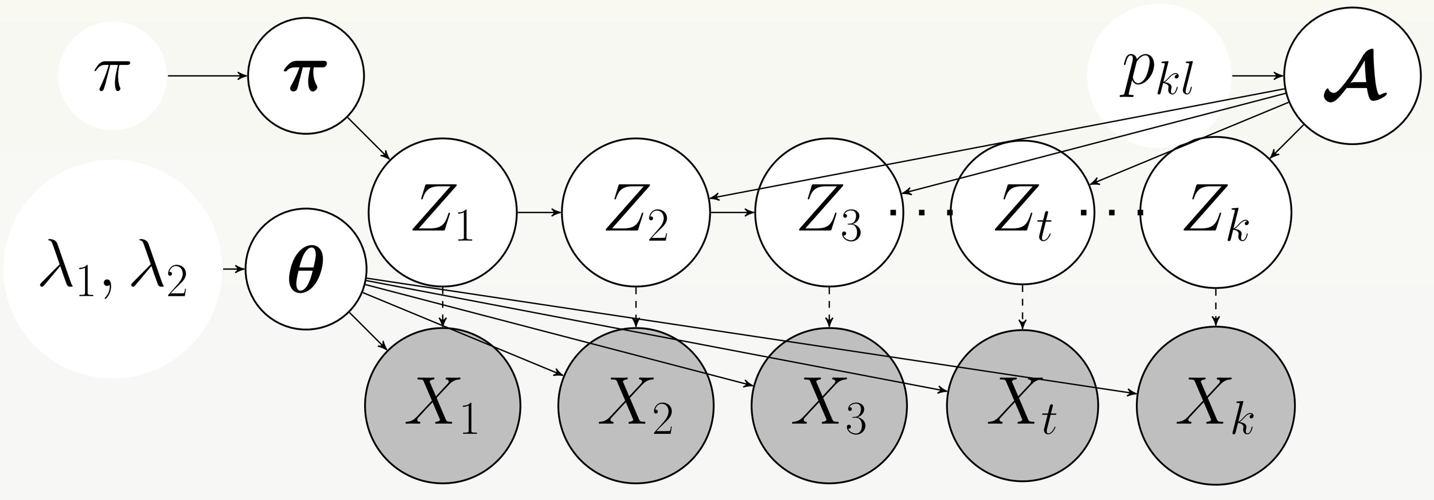

Fig. 6. Graphical model of the Hidden Markov Model.

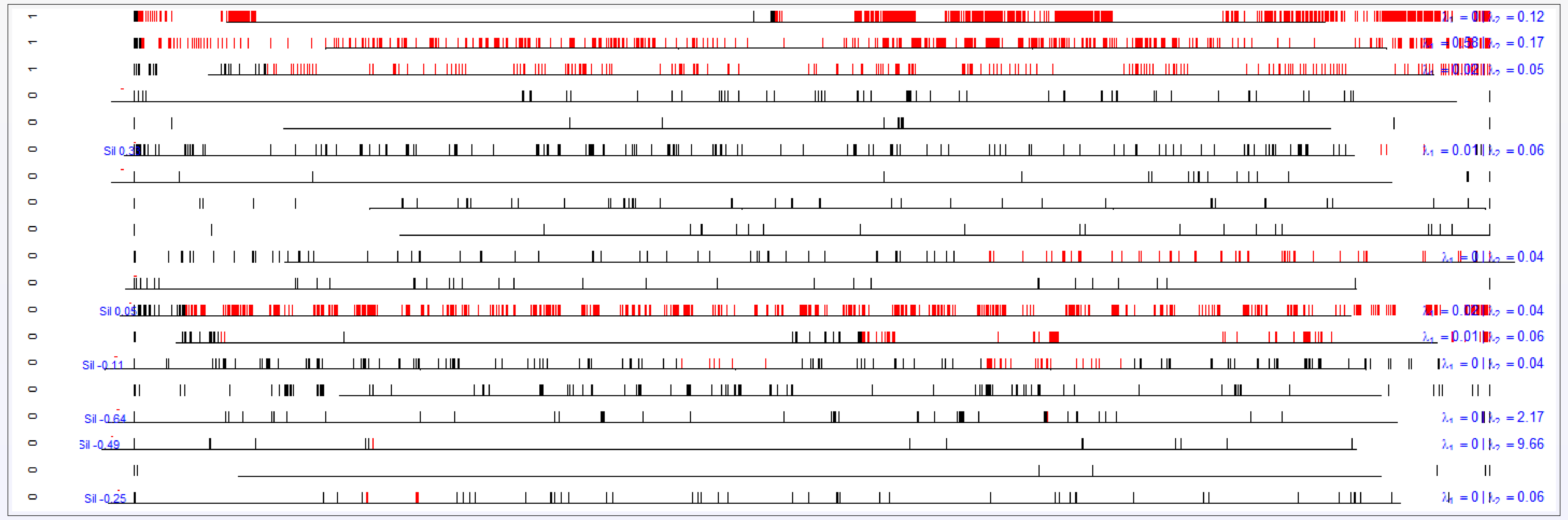

Fig. 7. Burst detection was applied in a real dataset, consisting of all the outgoing calls from a sample of 5,000 subscribers of a Peruvian telco (May 2017). The colors show a posterior point estimate of the sequence of states for 19 call-trains.

2.3 Bayesian Inference:

Learning the parameters $\{\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\mathcal{A}}\} = \{z_1, ..., z_k, \pi, \lambda_1, \lambda_2, p_{N\to B}, p_{B\to N}\}$, answers the question "*What is the sequence of hidden states, initial probabilities, frequency rates, & transition prob. that maximizes $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\mathcal{A}})$?*". Prior dist.: $\pi \sim Beta(a_\pi, b_\pi)$; $\lambda_s \sim Gamma(\alpha_s, \beta_s)$ and $p_{k\to l} \sim Beta(a_p, b_p)$. We use a *Gibbs Sampler* and initialize the states using the *Viterbi algorithm*.

2.4 The probability of fraud:

We need

$$P(\text{"fraud"}\,|\,\text{call-train})$$

Think, "*How would the user behave if he had the opportunity to make infinitely many calls?*". It induces the limiting distribution of its Markov chain, $\tilde{\boldsymbol{\pi}} = [\tilde{\pi} \quad (1-\tilde{\pi})]$, s.t. $\boldsymbol{\pi}\boldsymbol{\mathcal{A}}^n \to \tilde{\boldsymbol{\pi}}$, as $n \to \infty$. We define the event *fraud* by setting a threshold b to the "burstiness": *fraud*$=(1-\tilde{\pi}) > b$. We get

$$P(1-\tilde{\pi} > b\,|\,\text{call-train})$$

3. Main Results!

Fig.7, 8 and 9 show successful identification of bursts and efficient characterization of fraud, respectively.

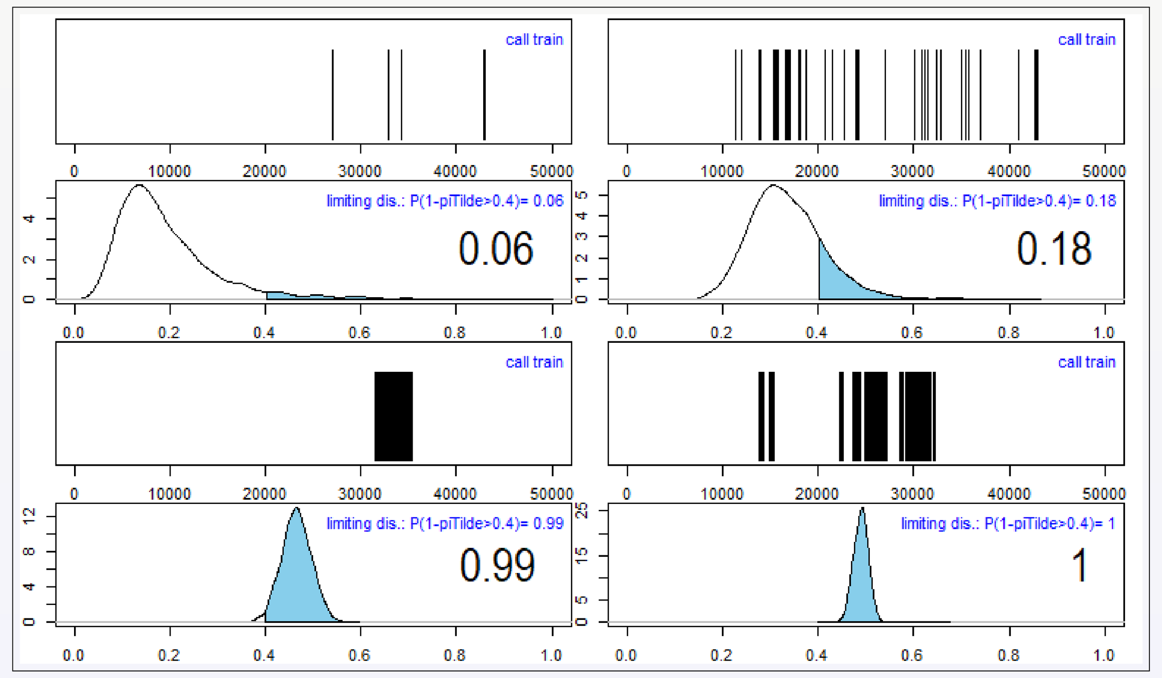

Fig. 8. Probability of fraud for 2 confirmed cases of regular use (top) and 2 confirmed cases of fraud (bottom).

Fig. 9. Effective detection even under lower calls-traffic.