Statistical methods for separating human and automated activity in computer network traffic

Imperial College London

Francesco Sanna Passino
francesco.sanna-passino16@imperial.ac.uk

Supervisor:
Dr Nicholas Heard

PhD project
"Latent factor representations of dynamic networks in cyber-security"

1. Problem

Most datasets used for cyber-security can be considered as mixtures of human and automated events. For example, it is estimated that the proportion of automated traffic in Network Flow data is approximately 95%. For statistical purposes, it is essential to correctly separate these two types of activity, in order to build sound models of normal behaviour of the network.

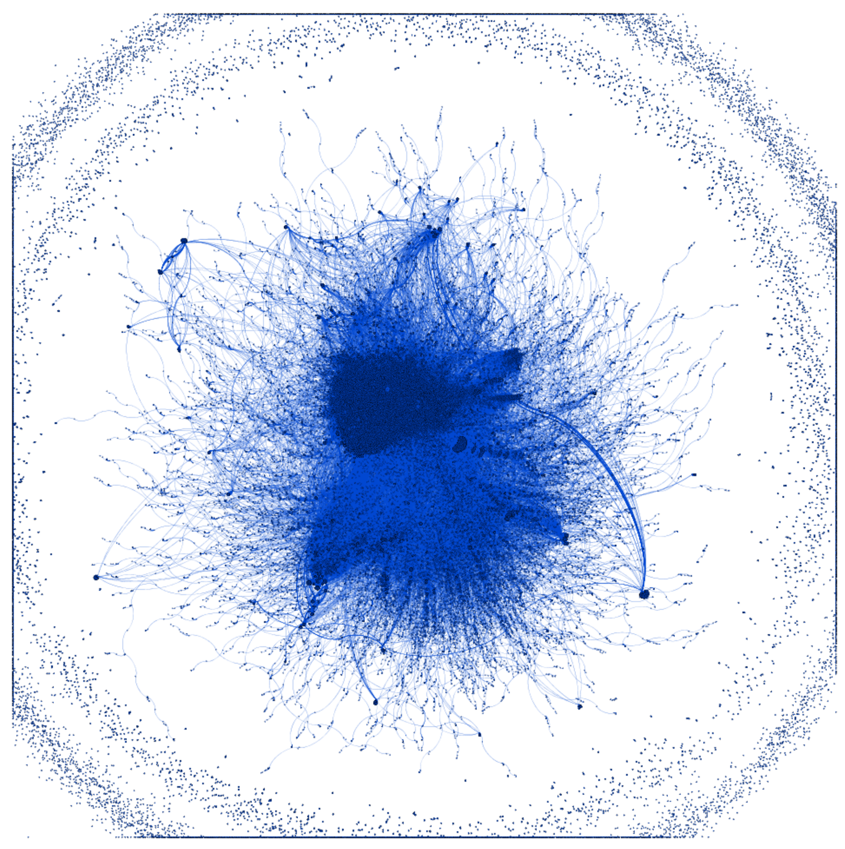

Figure 1: *Imperial College network graph on June 7, 2017, 11:15 – 11:16am. Each node corresponds to an IP address, an edge is drawn if the two IPs have connected within the observation period.*

2. Detection of periodicities

Methodology developed in **Heard, Rubin–Delanchy and Lawson (2014)**:

- $t_1, t_2, \dots, t_N \to$ timestamps of the NetFlow events involving a client X and a server Y,
- $N(t), t \geq 0 \to$ counting process: number of NetFlow records involving the client X and the server Y at each time point t, starting from $t=0$,
- *Periodogram* $\hat{S}(f)$ at frequency $f > 0$:

$$\hat{S}(f) = \frac{1}{T}\left|\sum_{t=1}^{T}\left(\mathrm{d}N(t) - \frac{N(T)}{T}\right)e^{-2\pi\imath f t}\right|^2$$

where $\mathrm{d}N(t) = N(t) - N(t-1)$.

- Fourier's g-test for the null H_0 of no periodicities:

$$g = \frac{\max_{1\leq k\leq \lfloor T/2\rfloor}\hat{S}(f_k)}{\sum_{1\leq j\leq \lfloor T/2\rfloor}\hat{S}(f_j)},\ f_k = \frac{k}{T\Delta t}$$

- Setting $\lambda = \min\{\lfloor 1/g\rfloor, \lfloor T/2\rfloor\}$, the p-value is:

$$\mathbb{P}(g > g_\star) = \sum_{j=1}^{\lambda}(-1)^{j-1}\cdot\frac{m}{j}\cdot(1 - jg_\star)^{m-1}$$

3. Transforming the data

Suppose that an edge is periodic at significance level α with periodicity $p = T\Delta t/\mathrm{argmax}_{1\leq k\leq \lfloor T/2\rfloor}\hat{S}(f_k)$. Let $t_1, \dots, t_N$ be the sequence of **arrival times** on the edge. The quantity of interest for inference is a **latent assignment** z_i, defined as follows:

$$z_i = \begin{cases} 0 & \text{if } t_i \text{ is human} \\ 1 & \text{if } t_i \text{ is automated}\end{cases}$$

where $\mathbb{P}(z_i = 1) = \theta$ and $\mathbb{P}(z_i = 0) = 1 - \theta$.
Two quantities are used to model the arrival times:

- the **wrapped arrival time** x_i:

$$x_i = (t_i \mod p) \times 2\pi/p$$

- the **daily arrival time** y_i:

$$y_i = (t_i \mod 86400) \times 2\pi/86400$$

where 86400 is the number of seconds in one day.

4. The model

- For simplicity, assume $T \bmod 86400 = 0$ and $T \bmod p = 0$. Then the density of an arrival time can be decomposed as:

$$f(t_i|z_i) \propto f_A(x_i)^{z_i} f_H(y_i)^{1-z_i}$$

- Human events are modelled using the daily arrival time y_i, automated events using the wrapped arrival time x_i.
- *Fixed phase polling*: event times occur every p seconds plus a random zero-mean error.

$$x_i|(z_i = 1, \mu, \sigma^2) \overset{d}{\sim} \mathbb{WN}_{[0,2\pi)}(\mu, \sigma^2)$$

- Unknown density of the daily arrival times $\to$ step function:

$$p(y_i|z_i = 0, \boldsymbol{h}, \boldsymbol{\tau}, B) = \sum_{j=1}^{B}\frac{h_j}{\tau_{(j+1)} - \tau_{(j)}}\mathbb{1}\{y_i \in [\tau_{(j)}, \tau_{(j+1)})\}$$

where B is the number of bins, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_{B+1})$ are the bin locations, and $\boldsymbol{h} = (h_1, \dots, h_B)$, $\sum_j h_j = 1, h_j \geq 0\ \forall\ j$ are the bar heights.

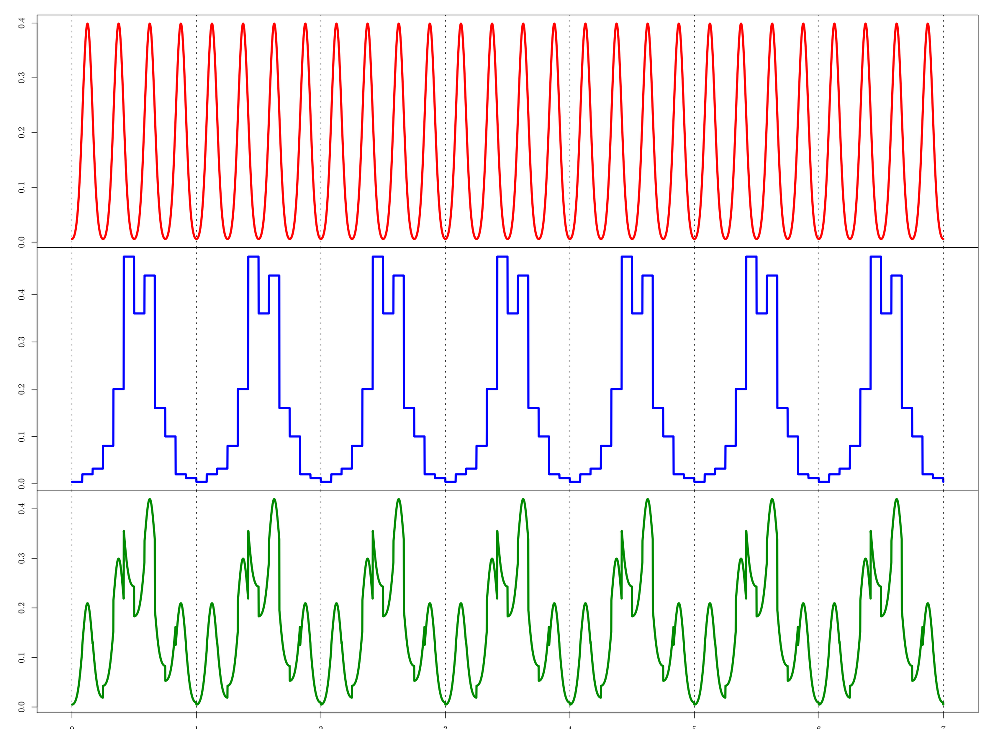

Figure 2: *Densities used in the model, $p = 6$ hours, $\mu = \pi$, $\sigma^2 = 1$, $\theta = 0.5$, $B = 12$, $\tau_j = \frac{2\pi j}{B}$, $j = 0, \dots, B$. Top plot (**red**): unnormalised density of the automated events. Middle plot (**blue**): unnormalised density of the human events. Bottom plot (**green**): unnormalised density of the 50-50 mixture.*

- General framework: $B, \boldsymbol{\tau}, \boldsymbol{h}$ unknown $\longrightarrow$ inference is easier (conjugate priors available!) when considering $\tau_j = 2\pi j/B,\ j = 0, \dots, B$ for each possible value of an unknown $B \in \{1, \dots, B_{\max}\}$.
- The resulting model, for $T \bmod 86400 = 0$ and $\lfloor T/p\rfloor \gg T \bmod p$, is a mixture of the two components:

$$f(t_i|z_i) \underset{\sim}{\propto} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{k=-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(x_i + 2\pi k - \mu)^2\right\}\right)^{z_i}\left(\sum_{j=1}^{B}\frac{h_j}{\tau_{(j+1)} - \tau_{(j)}}\mathbb{1}\{y_i \in [\tau_{(j)}, \tau_{(j+1)})\}\right)^{1-z_i}$$

- Prior distributions for conjugate analysis in the case of a standard histogram ($\tau_j = 2\pi j/B,\ j = 0, \dots, B$):
  - $(\mu, \sigma^2) \overset{d}{\sim} \mathrm{NIG}(\mu_0, \sigma_0^2, \alpha_0, \beta_0)$
  - $\theta \overset{d}{\sim} \mathrm{Beta}(\gamma_0, \delta_0)$
  - $\boldsymbol{h}|B \overset{d}{\sim} \mathrm{Dirichlet}(2\pi\eta/B\mathbf{1}^\top)$
- Straightforward Gibbs sampler available, even when B is unknown $\to$ it is possible to jointly sample $(\boldsymbol{h}, B)$.
- The algorithm successfully separates human and automated activity in synthetic (labelled) datasets.
- Reasonable results on real edges, where the true labels are not available.

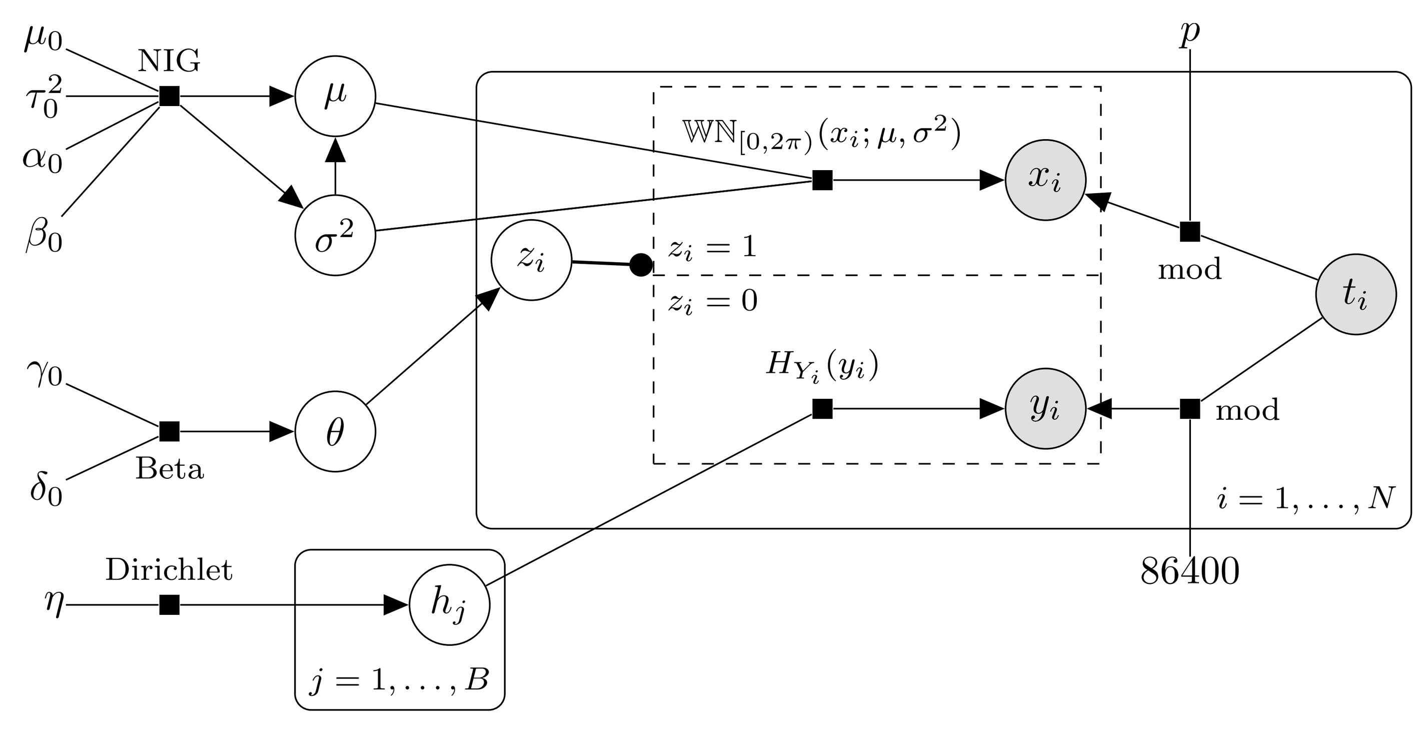

Figure 3: *Representation of the histogram model for a fixed number of bins B.*

5. Results on a real edge

- 2 weeks of connections between an IP X and the Microsoft Live IP 157.56.192.95.
- 13545 events, 1425 filtered human connections.
- The activity slightly increases during the day, suggesting a mixture of human and automated events.
- The distribution of human events obtained from the model shows a clear diurnal pattern, with almost no activity during the night.
- Events are not labeled in this example, but encouraging results have been obtained on synthetic labeled data.

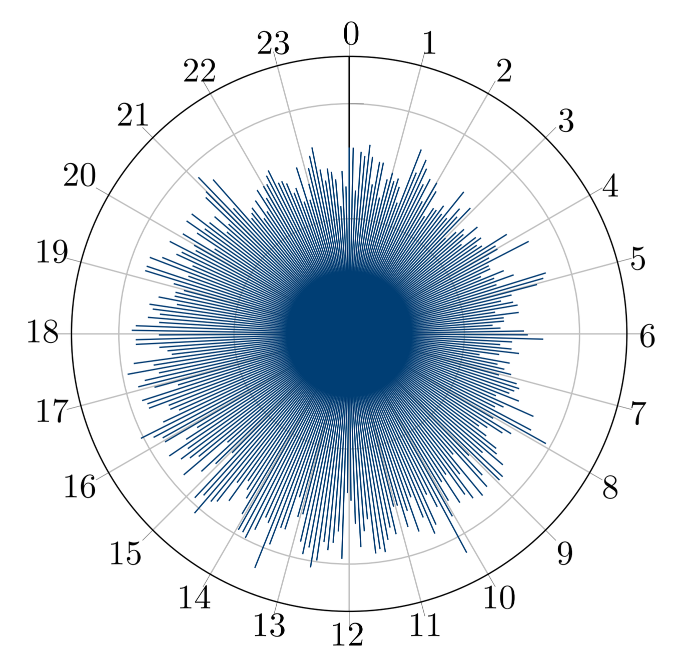

Figure 4: *Daily distribution of the data, slight evidence of increased activity during working hours.*

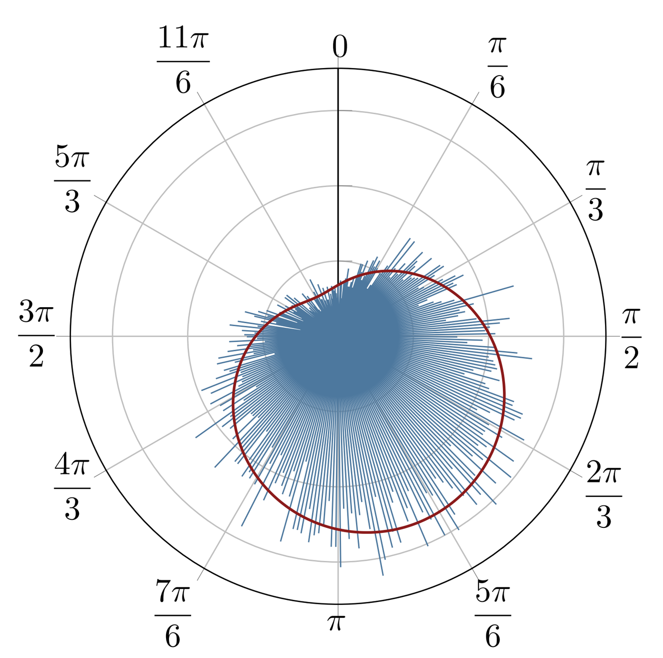

Figure 5: *Distribution of the wrapped data, $p = 4089.86$s and model fit (MAP estimates of μ and σ^2).*

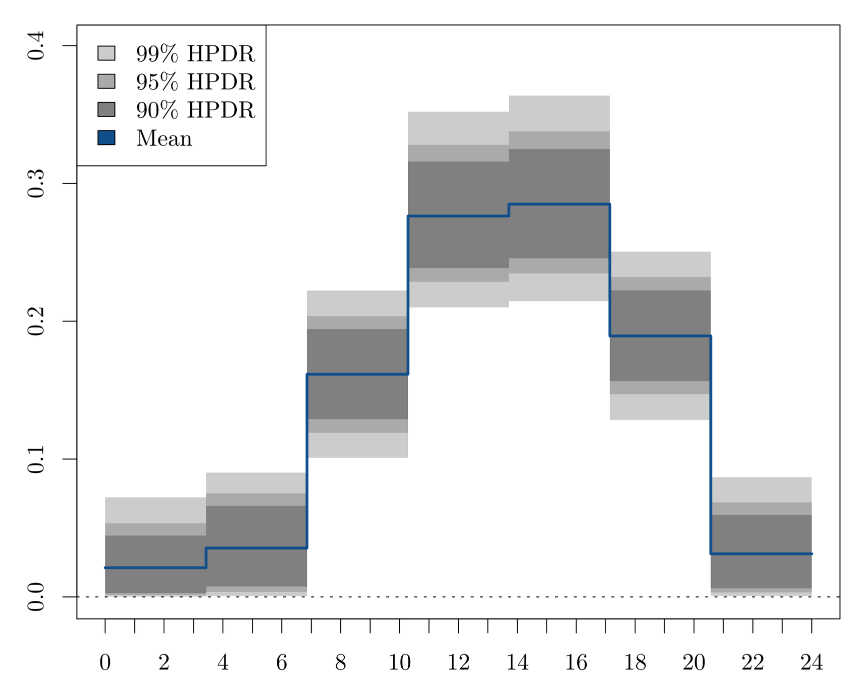

Figure 6: *Estimated optimal histogram of human events, $\hat{B}_{\text{opt}} = 7$. Clear diurnal pattern, activity concentrated in working hours only.*

6. Comments

- Simple algorithm to separate human and automated activity on a single edge in a computer network.
- Gibbs sampler with conjugate priors $\to$ scalable to multiple edges and nodes across the entire network.
- Results on multiple real and simulated dataset show good performance of the model.

References

- Heard, N.A, P.T.G. Rubin–Delanchy, and D.J. Lawson (2014), *"Filtering automated polling traffic in computer network flow data"*. In: *Proceedings of the IEEE Joint Intelligence & Security Informatics Conference (JISIC 2014)*, pp. 268-271 (2014).