Supplementary materials for Paper "Attributing Hacks"

Ziqi Liu
Xi'an Jiaotong Univ.
Ant Financial Group

Alexander J. Smola
Amazon

Kyle Soska
CMU

Yu-Xiang Wang
CMU

Qinghua Zheng
Xi'an Jiaotong Univ.

A Proof of Theorem 1

When the feature x is constant, the hazard function for user j

$$\lambda(t) = \sum_{i=1}^{d} x_{ji}(t) w_i(t)$$

and the cumulative hazard function

$$\begin{aligned}\Lambda(t) &= \int_{-\infty}^{t} \sum_{i=1}^{d} x_{ji}(t) w_i(t) dt = \sum_{i=1}^{d} \int_{-\infty}^{t} x_{ji}(t) w_i(t) dt \\ &= \sum_{i=1}^{d} \left(\sum_{\tau \in \mathcal{T}_i, \tau \le t} \alpha_{i,\tau} W_i(\tau) + \alpha_{i,t} W_i(t) \right). \quad (1)\end{aligned}$$

where $W_i(t) = \int_{-\infty}^{t} w_j(t) dt$, $\mathcal{T}_i$ denotes all break points of th piecewise constant $x_{ji}(t)$ and $\alpha_{i,\tau}$ are coefficients that depends only on x_{ji}. When there are no uncensored observations, we can re-parameterize the above variational optimization problems using the $\Lambda(t)$ hence $W_j(t)$ alone:

$$\begin{aligned}\min_{(W_0, W_1, \ldots, W_d) \in \mathcal{F}^d} & \mathcal{L}(\{\boldsymbol{\tau}, \boldsymbol{\Psi}, \boldsymbol{Z}\}, \boldsymbol{W}) + \gamma \sum_{j=0}^{d} \mathrm{TV}(W_j) \\ \text{s.t. } & W_i(t) \ge 0, W_i(t+\delta) - W_i(t) \ge 0 \\ \text{for any } & i \in [p], t \in \mathbb{R}, \delta \in \mathbb{R}_+. \\ & W_i \text{ is convex.}\end{aligned}$$

Let $\mathcal{T}$ be the set of observed time points (including $0, T$ and all censored interval boundaries). For each i, let W_i^* be the optimal solution. By Proposition 7 of Mammen et al. (1997), we know that for each i, there is a spline $\tilde{W}_i$ of order 1 such that

$$\begin{cases} \text{All knots of the spline are contained in } \mathcal{T} \backslash \{0, T\} \\ \tilde{W}_i(\tau) = W_i^*(\tau) \text{ for all } \tau \in \mathcal{T} \\ \mathrm{TV}(\tilde{W}_i) \le \mathrm{TV}(W_i^*) \end{cases} \quad (2)$$

We will now show that $\tilde{W}_i$ also defines a set of optimal solution using these properties.

Note that the loss function $\mathcal{L}(\{\boldsymbol{\tau}, \boldsymbol{\Psi}, \boldsymbol{Z}\}, \boldsymbol{W})$ can be decomposed into the sum of negative log-probability of form as described in (8), and when there are no uncensored data, the value of the loss function is completely determined by the survival function $S(t)$ evaluated at $t \in \mathcal{T}$. There is a one-to-one mapping between survival functions and the cumulative hazard functions through $S(t) = \exp(-\Lambda(t))$. It follows from (1) that $\mathcal{L}(\{\boldsymbol{\tau}, \boldsymbol{\Psi}, \boldsymbol{Z}\}, \boldsymbol{W})$ is a function of $\boldsymbol{W}$ only through its evaluations at $\boldsymbol{W}(\mathcal{T})$, therefore

$$\mathcal{L}(\{\boldsymbol{\tau}, \boldsymbol{\Psi}, \boldsymbol{Z}\}, \tilde{\boldsymbol{W}}) = \mathcal{L}(\{\boldsymbol{\tau}, \boldsymbol{\Psi}, \boldsymbol{Z}\}, \boldsymbol{W}^*).$$

By $\mathrm{TV}(\tilde{W}_i) \le \mathrm{TV}(W_i^*)$, we know that $\tilde{\boldsymbol{W}}$ has a smaller overall objective function than the optimal solution.

It remains to show that $\tilde{\boldsymbol{W}}$ is feasible. First note that the only spline of order 1 that satisfy the first and second condition is the piecewise linear interpolation of the knots in T. For each i, the constraints require that W_i^* obeys that W_i^* is non-negative, non-decreasing and convex. This ensures that the piecewise linear interpolation of any subset of points in the domain of W_i^* to be also nonnegative, monotonically nondecreasing and convex, which ensures the feasibility of $\tilde{W}_i$.

Finally, $\tilde{W}_i$ can be represented by a nonnegative linear combination of truncated power basis functions defined on $\mathcal{T}$ and the corresponding hazard function w_i can be represented by the same nonnegative combination of step functions defined at $\mathcal{T}$. This completes the proof. □

B Additional experiments

B.1 Real-World Case Study

Another spot check for the model is the ability to corroborate existing literature on malicious web deteciton. Figure 1 demonstrates the change-points in $\lambda_i(t)$ for specific versions of Wordpress. The model assigns Wordpress 2.9.2, 3.2.1, 3.3.1 and 3.5.1 change-points around July 2011, August 2011, December 2011, and February 2013 respectively. The work of Soska et

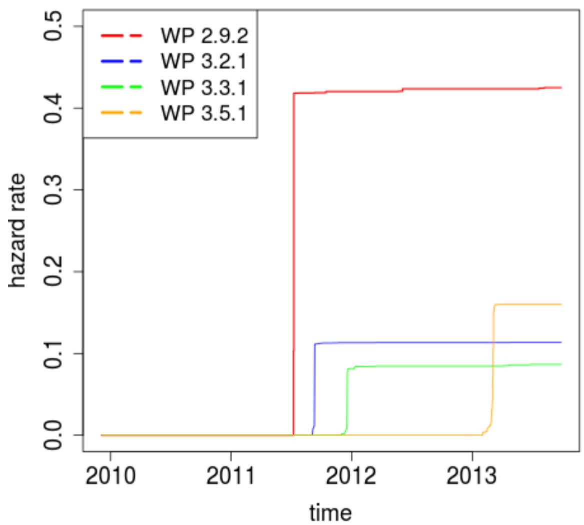

Figure 1: $\lambda_t(i)$ of features known to correspond directly to particular versions of Wordpress.

al. Soska & Christin (2014) found nearly identical attack campaigns for Wordpress 2.9.2, 3.2.1 and 3.3.1 but failed to produce a meaningful result for 3.5.1.

References

Mammen, Enno, van de Geer, Sara, et al. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

Soska, Kyle and Christin, Nicolas. Automatically detecting vulnerable websites before they turn malicious. *USENIX Security Symposium*, 2014.