Anytime-Valid A/B Testing of Counting Processes

Michael Lindon, Nathan Kallus

Netflix, Cornell

Netflix Encountered an Error

Figure 1. Example of the Netflix UI when Device Encounters an Error

Monitoring Error Events in Software A/B Tests

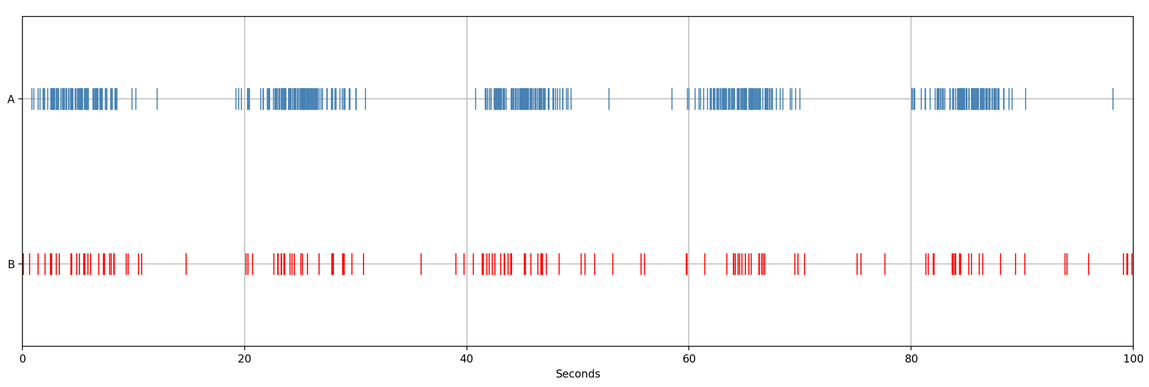

Figure 2. Timestamps of Errors from Devices Running Software Versions A and B

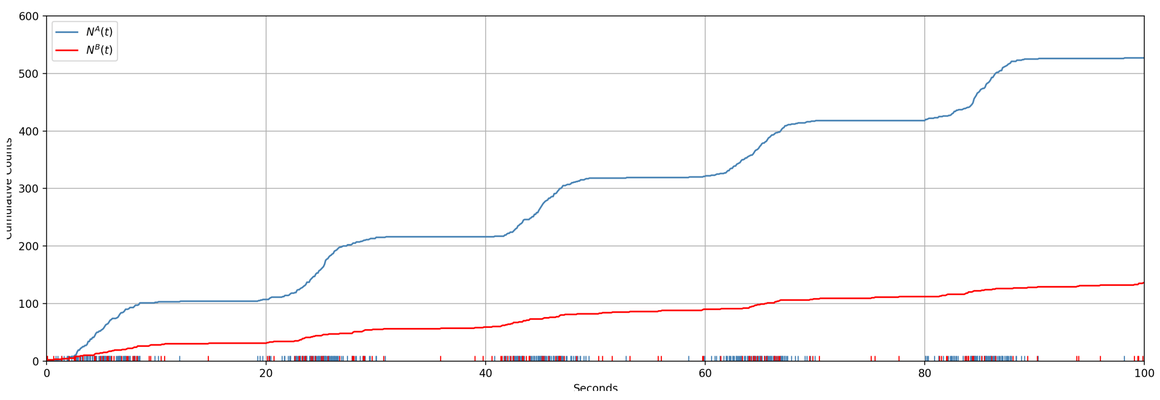

Figure 3. Counting Processes A and B (cumulative number of events by time t)

To de-risk software rollouts, Netflix conducts an A/B "Canary" test by releasing the new software version to a randomized subset of devices. Key performance metrics—such as successful stream starts, exceptions, and crashes—are continuously monitored and compared between control and treatment groups. If the new software significantly alters the incidence of such events, the rollout is halted promptly, preventing widespread user impact.

To rapidly detect issues while rigorously controlling false detections, we utilize anytime-valid sequential statistical tests based on e-processes.

General Counting Processes and Compensators

We model the point processes in Figure 3 as general counting processes. Formally, let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ be a filtered probability space. A counting process $N(t)$ is an $\mathcal{F}_t$-adapted, integer-valued, non-decreasing, càdlàg stochastic process with $N(0) = 0$.

Such processes are characterized through their compensators. Specifically, a non-decreasing, predictable process $\Lambda(t)$ is called the compensator of $N(t)$ if

$$M(t) = N(t) - \Lambda(t)$$

is an $\mathcal{F}_t$-martingale. When comparing two counting processes A and B, we compare their compensators $\Lambda^A(t)$ and $\Lambda^B(t)$, using simultaneous confidence processes for estimation and e-process for testing the hypothesis $H_0 : \Lambda^A(t) = \Lambda^B(t)\ \forall t > 0$.

Simultaneous Confidence Processes

For any fixed $\phi > 0$, the sets defined by

$$C^\alpha(t) = \left\{ (L^A, L^B) \in \mathbb{R}^2_{\geq 0} : \prod_{i\in\{A,B\}} \frac{\phi^\phi}{(\phi + L^i)^{\phi+N^i(t)}} \frac{\Gamma(\phi + N^i(t))}{\Gamma(\phi)} e^{L^i} \leq \alpha^{-1} \right\} \quad (1)$$

form a $1-\alpha$ confidence process for $(\Lambda^A(t), \Lambda^B(t))$, that is,

$$\mathbb{P}[(\Lambda^A(t), \Lambda^B(t)) \in C^\alpha(t)\ \forall t > 0] = 1 - \alpha.$$

Consider the following example where $N^A(t)$ and $N^B(t)$ are inhomogeneous Poisson counting processes with compensators $\Lambda^A(t) = \int_0^t e^{3\sin(2\pi s/20)} ds$ and $\Lambda^B(t) = \int_0^t e^{2\sin(2\pi s/20)} ds$ respectively.

Figure 4 shows the simultaneous $1-\alpha$ confidence processes on $\Lambda^A(t)$ and $\Lambda^B(t)$ from equation 1. Figure 5 visualizes the confidence process for $\Lambda^B(t)/t - \Lambda^A(t)/t$.

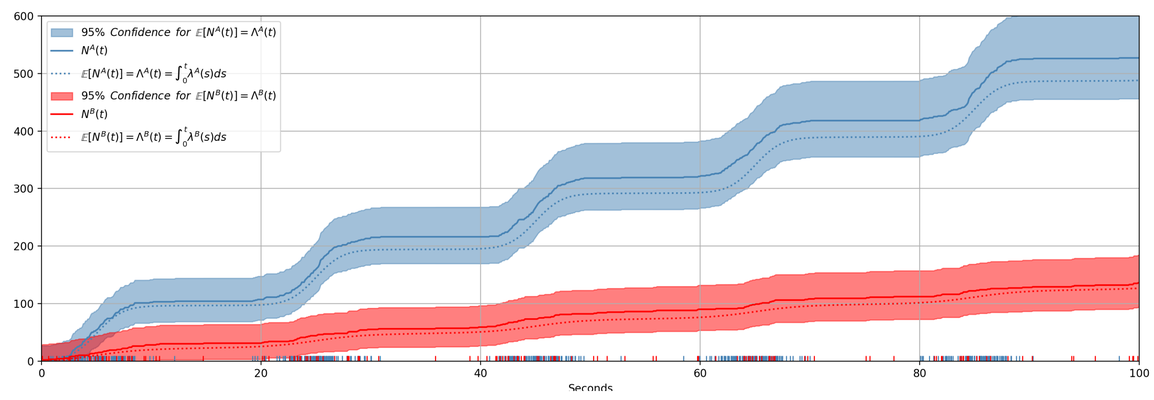

Figure 4. Simultaneous Confidence Processes for the Intensity Measures of A and B

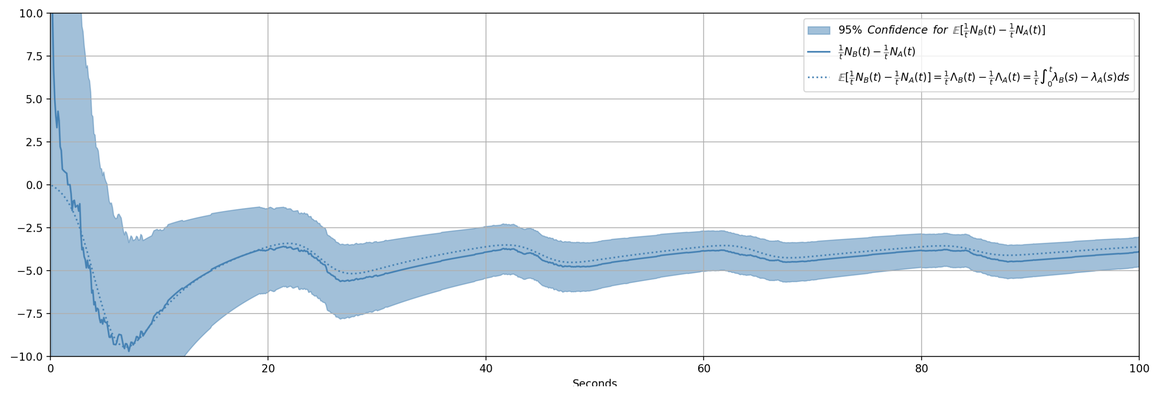

Figure 5. $1-\alpha$ simultaneous confidence processes on $\Lambda^B(t)/t - \Lambda^A(t)/t$

E-Processes for Testing $H_0 : \Lambda^A(t) = \Lambda^B(t)\ \forall t > 0$

Let $N^A(t)$ and $N^B(t)$ be two independent counting processes with $\Lambda^A(t) = \Lambda^B(t)\ \forall t > 0$. For any fixed $\phi > 0$, the process

$$E(t) := \prod_{i\in\{A,B\}} \frac{\phi^\phi}{(\phi + \hat{\Lambda}(t))^{\phi+N^i(t)}} \frac{\Gamma(\phi + N^i(t))}{\Gamma(\phi)} e^{\hat{\Lambda}(t)}, \quad (2)$$

where $\hat{\Lambda}(t) = \frac{1}{2}\left(N^A(t) + N^B(t)\right)$, is an e-process, which implies $\mathbb{P}[\exists t > 0 : E(t) > \alpha^{-1}] \leq \alpha$. The null hypothesis can be rejected at the α level as soon as $E(t) > \alpha^{-1}$.

E-Power

Let $N^A(t)$ and $N^B(t)$ be independent inhomogeneous Poisson processes with $\frac{\Lambda^A(t)}{t} \to \bar{\lambda}^A$ and $\frac{\Lambda^B(t)}{t} \to \bar{\lambda}^B$ as $t \to \infty$ respectively. Let $\bar{\lambda}^M = \frac{1}{2}\left(\bar{\lambda}^B + \bar{\lambda}^A\right)$, then for any $\phi > 0$

$$\begin{aligned} \frac{\log E(t)}{t} &\overset{a.s.}{\to} D_{KL}(\bar{\lambda}^B || \bar{\lambda}^M) + D_{KL}(\bar{\lambda}^A || \bar{\lambda}^M) \\ &= \bar{\lambda}^A \log \frac{2\bar{\lambda}^A}{\bar{\lambda}^A + \bar{\lambda}^B} + \bar{\lambda}^B \log \frac{2\bar{\lambda}^B}{\bar{\lambda}^A + \bar{\lambda}^B} \end{aligned} \quad (3)$$

where $D_{KL}(\bar{\lambda}^B || \bar{\lambda}^M)$ is the Kullback-Leibler divergence of a Poisson$(\bar{\lambda}^B)$ distribution from Poisson$(\bar{\lambda}^M)$

As long as the average rates differ in the limit, then our test will eventually reject the equality hypothesis $\lambda^A = \lambda^B$ with probability one:

$$\mathbb{P}[\exists t > 0 : E(t) \geq \alpha^{-1}] = 1.$$

Under the alternative, the e-process grows exponentially quickly with a rate defined by equation (3).

Conclusion

Every single action taken by the user on the Netflix application triggers an event to be logged. Examples of such events are - login, stream start, thumbnail preview start, ad break, rebuffer, app crash, app error etc. This observability data provides deep visibility into the performance and health of the Netflix service.

Data is received in two streams, events from devices running app version A, and events from devices running app version B.

This methodology enables the incidence of such events to be monitored in real-time, rapidly alerting when differences are detected while guaranteeing strict configurable false detection probabilities across time.

Read more about this methodology by following the QR code

Figure 6. Scan for Paper