arXiv:2106.04465v1 [cs.LG] 8 Jun 2021

Detecting Anomalous Event Sequences with Temporal Point Processes

Oleksandr Shchur*
Technical University of Munich
shchur@in.tum.de

Ali Caner Türkmen
Amazon Research
atturkm@amazon.com

Tim Januschowski
Amazon Research
tjnsch@amazon.com

Jan Gasthaus
Amazon Research
gasthaus@amazon.com

Stephan Günnemann
Technical University of Munich
guennemann@in.tum.de

Abstract

Automatically detecting anomalies in event data can provide substantial value in domains such as healthcare, DevOps, and information security. In this paper, we frame the problem of detecting anomalous continuous-time event sequences as out-of-distribution (OoD) detection for temporal point processes (TPPs). First, we show how this problem can be approached using goodness-of-fit (GoF) tests. We then demonstrate the limitations of popular GoF statistics for TPPs and propose a new test that addresses these shortcomings. The proposed method can be combined with various TPP models, such as neural TPPs, and is easy to implement. In our experiments, we show that the proposed statistic excels at both traditional GoF testing, as well as at detecting anomalies in simulated and real-world data.

1 Introduction

Event data is abundant in the real world and is encountered in various important applications. For example, transactions in financial systems, server logs, and user activity traces can all naturally be represented as discrete events in continuous time. Detecting anomalies in such data can provide immense industrial value. For example, abnormal entries in system logs may correspond to unnoticed server failures, atypical user activity in computer networks may correspond to intrusions, and irregular patterns in financial systems may correspond to fraud or shifts in the market structure.

Manual inspection of such event data is usually infeasible due to its sheer volume. At the same time, hand-crafted rules quickly become obsolete due to software updates or changing trends (He et al., 2016). Ideally, we would like to have an adaptive system that can learn the normal behavior from the data, and automatically detect abnormal event sequences. Importantly, such a system should detect anomalies in a completely unsupervised way, as high-quality labels are usually hard to obtain.

Assuming "normal" data is available, we can formulate the problem of detecting anomalous event sequences as an instance of out-of-distribution (OoD) detection. Multiple recent works consider OoD detection for image data based on deep generative models (Ren et al., 2019; Nalisnick et al., 2019; Wang et al., 2020). However, none of these papers consider continuous-time event data. Deep generative models for such variable-length event sequences are known as neural temporal point processes (TPPs) (Du et al., 2016). Still, the literature on neural TPPs mostly focuses on prediction tasks, and the problem of anomaly detection has not been adequately addressed by existing works (Shchur et al., 2021). We aim to fill this gap in our paper.

*Work done during an internship at Amazon Research.

Preprint. Under review.

Our main contributions are the following:

1. **Approach for anomaly detection for TPPs.** We draw connections between OoD detection and GoF testing for TPPs (Section 2). By combining this insight with neural TPPs, we propose an approach for anomaly detection that shows high accuracy on synthetic and real-world event data.
2. **A new test statistic for TPPs.** We highlight the limitations of popular GoF statistics for TPPs and propose the sum-of-squared-spacings statistic that addresses these shortcomings (Section 4).

2 Anomaly detection and goodness-of-fit testing

Background. A temporal point process (TPP) (Daley & Vere-Jones, 2003), denoted as $\mathbb{P}$, defines a probability distribution over variable-length event sequences in an interval $[0, T]$. A TPP realization X consists of strictly increasing arrival times $(t_1, \ldots, t_N)$, where N, the number of events, is itself a random variable. A TPP is characterized by its conditional intensity function $\lambda^*(t) := \lambda(t|\mathcal{H}_t)$ that is equal to the rate of arrival of new events given the history $\mathcal{H}_t = \{t_j : t_j < t\}$. Equivalently, a TPP can be specified with the integrated intensity function (a.k.a. the compensator) $\Lambda^*(t) = \int_0^t \lambda^*(u)du$.

Out-of-distribution (OoD) detection. We formulate the problem of detecting anomalous event sequences as an instance of OoD detection (Liang et al., 2018). Namely, we assume that we are given a large set of training sequences $\mathcal{D}_{\text{train}} = \{X_1, \ldots, X_M\}$ that were sampled i.i.d. from some *unknown* distribution $\mathbb{P}_{\text{data}}$ over a domain $\mathcal{X}$. At test time, we need to determine whether a new sequence X was also drawn from $\mathbb{P}_{\text{data}}$ (i.e., X is in-distribution or "normal") or from another distribution $\mathbb{Q} \neq \mathbb{P}_{\text{data}}$ (i.e., X is out-of-distribution or anomalous). We can phrase this problem as a null hypothesis test:

$$H_0 \colon X \sim \mathbb{P}_{\text{data}} \qquad H_1 \colon X \sim \mathbb{Q} \quad \text{for some} \quad \mathbb{Q} \neq \mathbb{P}_{\text{data}}. \tag{1}$$

To reiterate, here we consider the case where X is a variable-length event sequence and $\mathbb{P}_{\text{data}}$ is some unknown TPP. However, the rest of the discussion in Section 2 also applies to distributions over other data types, such as images.

Goodness-of-fit (GoF) testing. First, we observe that the problem of OoD detection is closely related to the problem of GoF testing (D'Agostino, 1986). We now outline the setup and approaches for GoF testing, and then describe how these can be applied to OoD detection. The goal of a GoF test to determine whether a random element X follows a *known* distribution $\mathbb{P}_{\text{model}}$[2]

$$H_0 \colon X \sim \mathbb{P}_{\text{model}} \quad H_1 \colon X \sim \mathbb{Q} \quad \text{for some} \quad \mathbb{Q} \neq \mathbb{P}_{\text{model}}. \tag{2}$$

We can perform such a test by defining a test statistic $s(X)$, where $s \colon \mathcal{X} \to \mathbb{R}$ (Fisher, 1936). For this, we compute the (two-sided) p-value for an observed realization x of X as[3]

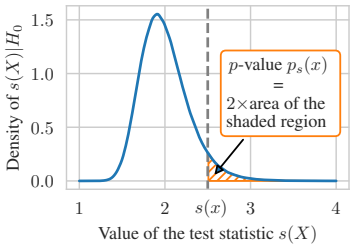

Figure 1: p-value is computed as the tail probability under the sampling distribution $s(X)|H_0$.

$$p_s(x) = 2 \times \min\{\Pr(s(X) \leq s(x)|H_0), 1 - \Pr(s(X) \leq s(x)|H_0)\}. \tag{3}$$

The factor 2 accounts for the fact that the test is two-sided. We reject the null hypothesis (i.e., conclude that X doesn't follow $\mathbb{P}_{\text{model}}$) if the p-value is below some predefined confidence level α. Note that computing the p-value requires evaluating the cumulative distribution function (CDF) of the sampling distribution, i.e., the distribution test statistic $s(X)$ under the null hypothesis H_0.

GoF testing vs. OoD detection. The two hypothesis tests (Equations 1 and 2) appear similar—in both cases the goal is to determine whether X follows a certain distribution $\mathbb{P}$ and no assumptions are made about the alternative $\mathbb{Q}$. This means that we can perform OoD detection using the procedure described above, that is, by defining a test statistic $s(X)$ and computing the respective p-value (Equation 3). However, in case of GoF testing (Equation 2), the distribution $\mathbb{P}_{\text{model}}$ is known. Therefore, we can analytically compute or approximate the CDF of $s(X)|X \sim \mathbb{P}_{\text{model}}$, and thus the p-value. In contrast, in an OoD detection hypothesis test (Equation 1), we make no assumptions about $\mathbb{P}_{\text{data}}$ and only

[2] We test a single realization X, as is common in TPP literature (Brown et al., 2002). Note that this differs from works on *univariate* GoF testing that consider multiple realizations, i.e., $H_0 \colon X_1, \ldots, X_M \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{model}}$.

[3] In the rest of the paper, the difference between the random element X and its realization x is unimportant, so we denote both as X, as is usually done in the literature.

have access to samples $\mathcal{D}_{\text{train}}$ that were drawn from this distribution. For this reason, we cannot compute the CDF of $s(X)|X \sim \mathbb{P}_{\text{data}}$ analytically. Instead, we can approximate the p-value using the empirical distribution function (EDF) of the test statistic $s(X)$ on $\mathcal{D}_{\text{train}}$.

The above procedure can be seen as a generalization of many existing methods for unsupervised OoD detection. These approaches usually define the test statistic based on the log-likelihood (LL) of a generative model fitted to $\mathcal{D}_{\text{train}}$ (Choi et al., 2018; Ren et al., 2019; Ruff et al., 2021). However, as follows from our discussion above, there is no need to limit ourselves to LL-based statistics. For instance, we can define a test statistic for event sequences based on the rich literature on GoF testing for TPPs. We show in Section 6 that this often leads to more accurate anomaly detection compared to LL. By drawing a clear distinction between OoD detection and GoF testing, we can also avoid some of the pitfalls encountered by other works (Nalisnick et al., 2019), as we elaborate in Appendix A.

To perform anomaly detection using the above-proposed approach, we need to find a good test statistic for continuous-time event sequences. In Section 3, we take a look at existing GoF statistics for TPPs and analyze their limitations. Then in Section 4), we propose a new test statistic that addresses these shortcomings and describe in more detail how it can be used for OoD detection.

3 Review of existing GoF test statistics for TPPs

Here, we consider a GoF test (Equation 2), where the goal is to determine whether an event sequence $X = (t_1, \dots, t_N)$ was generated by a known TPP $\mathbb{P}_{\text{model}}$ with compensator Λ^*. We will return to the problem of OoD detection, where the data-generating distribution $\mathbb{P}_{\text{data}}$ is unknown, in Section 4.2.

Many popular GoF tests for TPPs are based on the following result (Ogata, 1988; Brown et al., 2002).

Theorem 1 (Random time change theorem (Brown et al., 2002))**.** *A sequence $X = (t_1, \dots, t_N)$ is distributed according to a TPP with compensator Λ^* on the interval $[0, V]$ if and only if the sequence $Z = (\Lambda^*(t_1), \dots, \Lambda^*(t_N))$ is distributed according to the standard Poisson process on $[0, \Lambda^*(V)]$.*

Intuitively, Theorem 1 can be viewed as a TPP analogue of how the CDF of an arbitrary random variable over $\mathbb{R}$ transforms its realizations into samples from $\text{Uniform}([0, 1])$. Similarly, the compensator Λ^* converts a random event sequence X into a realization Z of the standard Poisson process (SPP). Therefore, the problem of GoF testing for an arbitrary TPP reduces to testing whether the transformed sequence Z follows the SPP on $[0, \Lambda^*(T)]$. In other words, we can define a GoF statistic for a TPP with compensator Λ^* by (1) applying the compensator to X to obtain Z and (2) computing one of the existing GoF statistics for the SPP on the transformed sequence. This can also be generalized to marked TPPs (where events can belong to one of K classes) by simply concatentating the transformed sequences $Z^{(k)}$ for each event type $k \in \{1, \dots, K\}$ (see Appendix D for details).

SPP, i.e., the Poisson process with constant intensity $\lambda^*(t) = 1$, is the most basic TPP one can conceive. However, as we will shortly see, existing GoF statistics even for this simple model have considerable shortcomings and can only detect a limited class of deviations from the SPP. More importantly, test statistics for general TPPs defined using the above recipe (Theorem 1) inherit the limitations of the SPP statistics.

For brevity, we denote the transformed arrival times as $Z = (v_1, \dots, v_N) = (\Lambda^*(t_1), \dots, \Lambda^*(t_N))$ and the length of the transformed interval as $V = \Lambda^*(T)$. One way to describe the generative process of an SPP is as follows (Pasupathy, 2010)

$$N|V \sim \text{Poisson}(V) \qquad u_i|N, V \sim \text{Uniform}([0, V]) \quad \text{for } i = 1, \dots, N. \tag{4}$$

An SPP realization $Z = (v_1, \dots, v_N)$ is obtained by sorting the u_i's in increasing order. This is equivalent to defining the arrival time v_i as the i-th order statistic $u_{(i)}$. We can also represent Z by the inter-event times $(w_1, \dots, w_{N+1})$ where $w_i = v_i - v_{i-1}$, assuming $v_0 = 0$ and $v_{N+1} = V$.

Barnard (1953) proposed a GoF test for the SPP based on the above description (Equation 4) and the Kolmogorov–Smirnov (KS) statistic. The main idea of this approach is to check whether the arrival times $v_1, \dots, v_N$ are distributed uniformly in the $[0, V]$ interval. For this, we compare $\hat{F}_{\text{arr}}$, the empirical CDF of the arrival times, with $F_{\text{arr}}(u) = u/V$, the CDF of the $\text{Uniform}([0, V])$ distribution. This can be done using the *KS statistic on the arrival times (KS arrival)*, defined as

$$\kappa_{\text{arr}}(Z) = \sqrt{N} \cdot \sup_{u \in [0,V]} |\hat{F}_{\text{arr}}(u) - F_{\text{arr}}(u)| \qquad \text{where} \qquad \hat{F}_{\text{arr}}(u) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(v_i \le u). \tag{5}$$

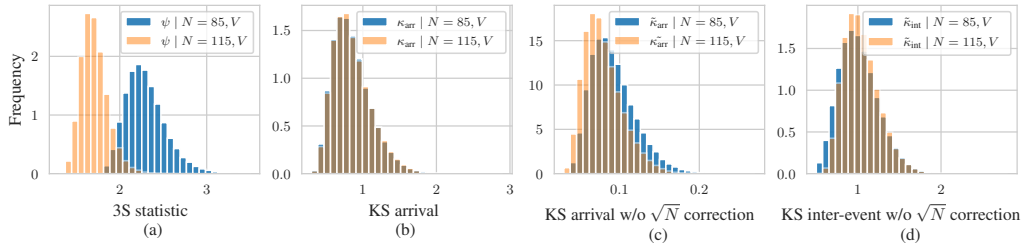

Figure 2: Distribution of different test statistics for the standard Poisson process on $[0, 100]$, conditioned on different event counts N. The 3S statistic allows us to differentiate between different values of N, while the KS statistics are not sensitive to the changes in N.

Another popular GoF test for the SPP is based on the fact that the inter-event times w_i are distributed according to the $\text{Exponential}(1)$ distribution (Cox, 1966). The test compares $\hat{F}_{\text{int}}$, the empirical CDF of the inter-event times, and $F_{\text{int}}(u) = 1 - \exp(-u)$, the CDF of the $\text{Exponential}(1)$ distribution. This leads to the *KS statistic for the inter-event times (KS inter-event)*

$$\kappa_{\text{int}}(Z) = \sqrt{N} \cdot \sup_{u \in [0,\infty)} |\hat{F}_{\text{int}}(u) - F_{\text{int}}(u)| \quad \text{where} \quad \hat{F}_{\text{int}}(u) = \frac{1}{N+1} \sum_{i=1}^{N+1} \mathbb{1}(w_i \leq u). \tag{6}$$

KS arrival and KS inter-event statistics are often presented as the go-to approach for testing the goodness-of-fit of the standard Poisson process (Daley & Vere-Jones, 2003). Combining them with Theorem 1 leads to simple GoF tests for arbitrary TPPs that are widely used to this day (Gerhard et al., 2011; Alizadeh et al., 2013; Kim & Whitt, 2014; Tao et al., 2018; Li et al., 2018).

Limitations of the KS statistics. The KS statistics $\kappa_{\text{arr}}(Z)$ and $\kappa_{\text{int}}(Z)$ are only able to differentiate the SPP from a narrow class of alternative processes. For example, KS arrival only checks if the arrival times v_i are distributed uniformly, conditioned on the event count N. But what if the observed N is itself extremely unlikely under the SPP (Equation 4)? KS inter-event can be similarly insensitive to the event count—removing all events $\frac{V}{2} < v_i \leq V$ from an SPP realization Z will only result in just a single atypically large inter-event time w_i, which changes the value of $\kappa_{\text{int}}(Z)$ at most by $\frac{1}{N+1}$. We demonstrate these limitations of $\kappa_{\text{arr}}(Z)$ and $\kappa_{\text{int}}(Z)$ in our experiments in Section 6.1. Other failure modes of the KS statistics were described by Pillow (2009). Note that ad-hoc fixes to the KS statistics do not address these problems. For example, combining multiple tests performed separately for the event count and arrival times using Fisher's method (Fisher, 1948; Cox, 1966) consistently decreases the accuracy, as we show in Appendix G. In the next section, we introduce a different test statistic that aims to address these shortcomings.

4 Sum-of-squared-spacings (3S) statistic for TPPs

4.1 Goodness-of-fit testing with the 3S statistic

A good test statistic should capture multiple properties of the SPP at once: it should detect deviations w.r.t. both the event count N and the distribution of the arrival or inter-event times. Here, we propose to approach GoF testing with a *sum-of-squared-spacings (3S) statistic* that satisfies these desiderata,

$$\psi(Z) = \frac{1}{V} \sum_{i=1}^{N+1} w_i^2 = \frac{1}{V} \sum_{i=1}^{N+1} (v_i - v_{i-1})^2. \tag{7}$$

This statistic extends the sum-of-squared-spacings statistic proposed as a test of uniformity for fixed-length samples by Greenwood (1946). The important difference between our definition (Equation 7) and prior works (D'Agostino, 1986) is that we, for the first time, consider the TPP setting, where the number of events N is random as well. For this reason, we use the normalizing constant $1/V$ instead of N/V^2 (see Appendix B for details). As we will see, this helps capture abnormalities in the event count and results in more favorable asymptotic properties for the case of SPP.

Intuitively, for a fixed N, the statistic ψ is maximized if the spacings are extremely imbalanced, i.e., if one inter-event time w_i is close to V and the rest are close to zero. Conversely, ψ attains its minimum when the spacings are all equal, that is $w_i = \frac{V}{N+1}$ for all i.

In Figure 2a we visualize the distribution of $\psi|N, V$ for two different values of N. We see that the distribution of ψ depends strongly on N, therefore a GoF test involving ψ will detect if the event count N is atypical for the given SPP. This is in contrast to κ_{arr} and κ_{int}, the distributions of which, by design, are (asymptotically) invariant under N (Figure 2b). Even if one accounts for this effect, e.g., by removing the correction factor $\sqrt{N}$ in Equations 5 and 6, their distributions change only slightly compared to the sum of squared spacings (see Figures 2c and 2d). To analyze other properties of the statistic, we consider its moments under the null hypothesis.

Proposition 1. *Suppose the sequence Z is distributed according to the standard Poisson process on the interval $[0, V]$. Then the first two moments of the statistic $\psi := \psi(Z)$ are*

$$\mathbb{E}[\psi|V] = \frac{2}{V}(V + e^{-V} - 1) \quad \textit{and} \quad \mathrm{Var}[\psi|V] = \frac{4}{V^2}(2V - 7 + e^{-V}(2V^2 + 4V + 8 - e^{-V})).$$

The proof of Proposition 1 can be found in Appendix C. From Proposition 1 it follows that

$$\lim_{V\to\infty} \mathbb{E}[\psi|V] = 2 \qquad\qquad \lim_{V\to\infty} \mathrm{Var}[\psi|V] = 0. \tag{8}$$

This leads to a natural notion of *typicality* in the sense of Nalisnick et al. (2019) and Wang et al. (2020) for the standard Poisson process. We can define the typical set of the SPP as the set of variable-length sequences Z on the interval $[0, V]$ that satisfy $|\psi(Z) - 2| \leq \epsilon$ for some small $\epsilon > 0$. It follows from Equation 8 and Chebyshev's inequality that for large enough V, the SPP realizations will fall into the typical set with high probability. Therefore, at least for large V, we should be able to detect sequences that are not distributed according the SPP based on the statistic ψ.

Summary. To test the GoF of a TPP with a known compensator Λ^* for an event sequence $X = (t_1, \ldots, t_N)$, we first obtain the transformed sequence $Z = (\Lambda^*(t_1), \ldots, \Lambda^*(t_N))$ and compute the statistic $\psi(Z)$ according to Equation 7. Since the CDF of the statistic under H_0 cannot be computed analytically, we approximate it using samples drawn from $\mathbb{P}_{\text{model}}$. That is, we draw realizations $\mathcal{D}_{\text{model}} = \{X_1, \ldots, X_M\}$ from the TPP (e.g., using the inversion method (Rasmussen, 2018)) and compute the p-value for X (Equation 3) using the EDF of the statistic on $\mathcal{D}_{\text{model}}$ (North et al., 2002).

4.2 Out-of-distribution detection with the 3S statistic

We now return to the original problem of OoD detection in TPPs, where we have access to a set of in-distribution sequences $\mathcal{D}_{\text{train}}$ and do not know the data-generating process $\mathbb{P}_{\text{data}}$.

Our idea is to perform the OoD detection hypothesis test (Equation 1) using the sum-of-squared-spacings test statistic that we introduced in the previous section. However, since the data-generating TPP $\mathbb{P}_{\text{data}}$ is unknown, we do not know the corresponding compensator that is necessary to compute the statistic. Instead, we can fit a neural TPP model $\mathbb{P}_{\text{model}}$ (Du et al., 2016) to the sequences in $\mathcal{D}_{\text{train}}$ and use the compensator Λ^* of the learned model to compute the statistic $s(X)$.[4] High flexibility of neural TPPs allows these models to more accurately approximate the true compensator. Having defined the statistic, we can approximate its distribution under H_0 (i.e., assuming $X \sim \mathbb{P}_{\text{data}}$) by the EDF of the statistic on $\mathcal{D}_{\text{train}}$. We use this EDF to compute the p-values for our OoD detection hypothesis test and thus detect anomalous sequences. We provide the pseudocode description of our OoD detection method in Appendix D.

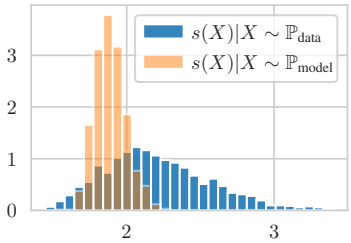

Figure 3: Sampling distribution for the OoD test (blue) and the GoF test (orange). While the same statistic $s(X)$ is used in both cases, the p-values are computed differently depending on which test we perform.

We highlight that an OoD detection procedure like the one above is *not* equivalent to a GoF test for the learned generative model $\mathbb{P}_{\text{model}}$, as suggested by earlier works (Nalisnick et al., 2019). While we use the compensator of the learned model to define the test statistic $s(X)$, we compute the p-value for the OoD detection test based on $s(X)|X \sim \mathbb{P}_{\text{data}}$. This is different from the distribution $s(X)|X \sim \mathbb{P}_{\text{model}}$ used in a GoF test, since in general $\mathbb{P}_{\text{model}} \neq \mathbb{P}_{\text{data}}$. Therefore, even if the distribution of a test statistic under the GoF test can be approximated analytically (as, e.g., for the KS statistic (Marsaglia et al., 2003)), we have to use the EDF of the statistic on $\mathcal{D}_{\text{train}}$ for the OoD detection test. Figure 3 visualizes

[4] We can replace the 3S statistic on the transformed sequence Z with any other statistic for the SPP, such as KS arrival. In Sections 6.2 and 6.3, we compare different statistics constructed this way.

this difference. Here, we fit a TPP model on the in-distribution sequences from the STEAD dataset (Section 6.3) and plot the empirical distribution of the respective statistic $s(X)$ on $\mathcal{D}_{\text{train}}$ (corresponds to $s(X)|X \sim \mathbb{P}_{\text{data}}$) and on model samples $\mathcal{D}_{\text{model}}$ (corresponds to $s(X)|X \sim \mathbb{P}_{\text{model}}$).

5 Related work

Unsupervised OoD detection. OoD detection approaches based on deep generative models (similar to our approach in Section 4.2) have received a lot of attention in the literature. However, there are several important differences between our method and prior works. First, most existing approaches perform OoD detection based on the log-likelihood (LL) of the model or some derived statistic (Choi et al., 2018; Ren et al., 2019; Nalisnick et al., 2019; Morningstar et al., 2021; Ruff et al., 2021). We observe that LL can be replaced by any other test statistic, e.g., taken from the GoF testing literature, which often leads to more accurate anomaly detection (Section 6). Second, unlike prior works, we draw a clear distinction between OoD detection and GoF testing. While this difference may seem obvious in hindsight, it is not acknowledged by the existing works, which may lead to complications (see Appendix A). Also, our formulation of the OoD detection problem in Section 2 provides an intuitive explanation to the phenomenon of "typicality" (Nalisnick et al., 2019; Wang et al., 2020). The $(\epsilon, 1)$-typical set of a distribution $\mathbb{P}$ simply corresponds to the acceptance region of the respective hypothesis test with confidence level ϵ (Equation 1). Finally, most existing papers study OoD detection for image data and none consider variable-length event sequences, which is the focus of our work.

Our OoD detection procedure is also related to the *rarity* anomaly score (Ferragut et al., 2012; Janzing et al., 2019). The rarity score can be interpreted as the negative logarithm of a one-sided p-value (Equation 3) of a GoF test that uses the log-likelihood of some *known* model as the test statistic. In contrast, we consider a broader class of statistics and learn the model from the data.

Anomaly detection for TPPs. OoD detection, as described in Section 2, is not the only way to formalize anomaly detection for TPPs. For example, Ojeda et al. (2019) developed a distance-based approach for Poisson processes. Recently, Zhu et al. (2020) proposed to detect anomalous event sequences with an adversarially-trained model. Unlike these two methods, our approach can be combined with any TPP model without altering the training procedure. Liu & Hauskrecht (2019) studied anomalous *event* detection with TPPs, while we are concerned with entire event sequences.

GoF tests for TPPs. Existing GoF tests for the SPP usually check if the arrival times are distributed uniformly, using, e.g., the KS (Lewis, 1965) or chi-squared statistic (Cox, 1955). Our 3S statistic favorably compares to these approaches thanks to its dependence on the event count N, as we explain in Section 4 and show experimentally in Section 6.1. Methods combining the random time change theorem with a GoF test for the SPP (usually, the KS test) have been used at least since Ogata (1988), and are especially popular in neuroscience (Brown et al., 2002; Gerhard et al., 2011; Tao et al., 2018). However, these approaches inherit the limitations of the underlying KS statistic. Replacing the KS score with the 3S statistic consistently leads to a better separation between different TPP distributions (Section 6).

Gerhard & Gerstner (2010) discussed several GoF tests for discrete-time TPPs, while we deal with continuous time. Yang et al. (2019) proposed a GoF test for point processes based on Stein's identity, which is related to a more general class of kernel-based GoF tests (Chwialkowski et al., 2016; Liu et al., 2016). Their approach isn't suitable for neural TPPs, where the Papangelou intensity cannot be computed analytically. A recent work by Wei et al. (2021) designed a GoF test for self-exciting processes under model misspecification. In contrast to these approaches, our proposed GoF test from Section 4.1 can be applied to any TPP with a known compensator.

Sum-of-squared-spacings statistic. A similar statistic was first used by Greenwood (1946) for testing whether a fixed number of points are distributed uniformly in an interval. Several follow-up works studied the limiting distribution of the statistic (conditioned on N) as $N \to \infty$ (Hill, 1979; Stephens, 1981; Rao & Kuo, 1984). Our proposed statistic (Equation 7) is not invariant w.r.t. N and, therefore, is better suited for testing TPPs. We discuss other related statistics in Appendix B.

6 Experiments

Our experimental evaluation covers two main topics. In Section 6.1, we compare the proposed 3S statistic with existing GoF statistics for the SPP. Then in Sections 6.2 and 6.3, we evaluate our

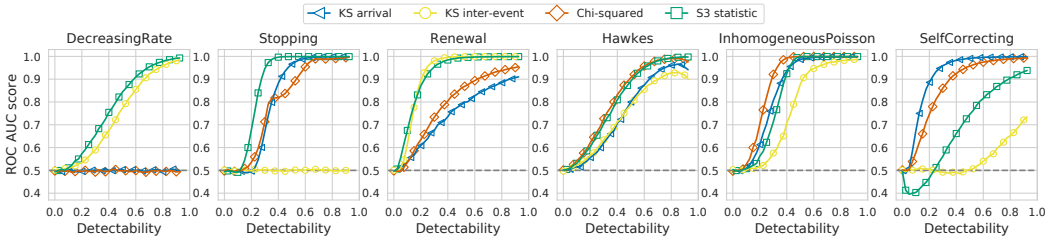

Figure 4: GoF testing for the standard Poisson process using different test statistics, measured with ROC AUC (higher is better). See Section 6.1 for the description of the experimental setup.

OoD detection approach on simulated and real-world data, respectively. The experiments were run on a machine with a 1080Ti GPU. Details on the setup and datasets construction are provided in Appendix E & F.

6.1 Standard Poisson process

In Section 3 we mentioned several failure modes of existing GoF statistics for the SPP. Then, in Section 4.1 we introduced the 3S statistic that was supposed to address these limitations. Hence, the goal of this section is to compare the proposed statistic with the existing ones in the task of GoF testing for the SPP. We consider four test statistics: (1) KS statistic on arrival times (Equation 5), (2) KS statistic on inter-event times (Equation 6), (3) chi-squared statistic on the arrival times (Cox, 1955; Tao et al., 2018), and (4) the proposed 3S statistic (Equation 7).

To quantitatively compare the discriminative power of different statistics, we adopt an evaluation strategy similar to Gerhard & Gerstner (2010); Yang et al. (2019). First, we generate a set $\mathcal{D}_{\text{model}}$ consisting of 1000 SPP realizations. We use $\mathcal{D}_{\text{model}}$ to compute the empirical distribution function of each statistic $s(Z)$ under H_0. Then, we define two test sets: $\mathcal{D}_{\text{test}}^{\text{ID}}$ (consisting of samples from $\mathbb{P}_{\text{model}}$, the SPP) and $\mathcal{D}_{\text{test}}^{\text{OOD}}$ (consisting of samples from $\mathbb{Q}$, another TPP), each with 1000 sequences. Importantly, in this and following experiments, the training and test sets are always disjoint.

We follow the GoF testing procedure described at the end of Section 4.1, which corresponds to the hypothesis test in Equation 2. That is, we compute the p-value (Equation 3) for each sequence in the test sets using the EDF of $s(Z)$ on $\mathcal{D}_{\text{model}}$. A good test statistic $s(Z)$ should assign lower p-values to the OoD sequences from $\mathcal{D}_{\text{test}}^{\text{OOD}}$ than to ID sequences from $\mathcal{D}_{\text{test}}^{\text{ID}}$, allowing us to discriminate between samples from $\mathbb{Q}$ and $\mathbb{P}_{\text{model}}$. We quantify how well a given statistic separates the two distributions by computing the area under the ROC curve (ROC AUC). This effectively averages the performance of a statistic for the GoF hypothesis test over different significance levels α.

Datasets. We consider six choices for the distribution $\mathbb{Q}$: (i) RATE, a homogeneous Poisson process with intensity $\mu < 1$, (ii) STOPPING, where events stop after some time $t_{\text{stop}} \in [0, V]$, (iii) RENEWAL, where inter-event times are drawn i.i.d. from the Gamma distribution, (iv) HAWKES, where events are more clustered compared to the SPP, (v) INHOMOGENEOUS, a Poisson process with non-constant intensity $\lambda(t) = \beta \sin(\omega t)$, and (vi) SELFCORRECTING, where events are more evenly spaced compared to the SPP. For cases (iii)–(vi), the expected number of events is the same as for the SPP.

For each choice of $\mathbb{Q}$ we define a *detectability* parameter $\delta \in [0, 1]$, where higher δ corresponds to TPPs that are increasingly dissimilar to the SPP. That is, setting $\delta = 0$ corresponds to a distribution $\mathbb{Q}$ that is exactly equal to the SPP, and $\delta = 1$ corresponds to a distribution that deviates significantly from the SPP. For example, for a Hawkes with conditional intensity $\lambda^*(t) = \mu + \beta \sum_{t_j < t} \exp(-(t - t_j))$, the detectability value of $\delta = 0$ corresponds to $\mu = 1$ and $\beta = 0$ (i.e., $\lambda^*(t) = 1$) making $\mathbb{Q}$ indistinguishable from $\mathbb{P}$. The value of $\delta = 0.5$ corresponds to $\mu = 0.5$ and $\beta = 0.5$, which preserves the expected number of events N but makes the arrival times t_i "burstier." We describe how the parameters of each distribution $\mathbb{Q}$ are defined based on δ in Appendix E. Note that, in general, the ROC AUC scores are not guaranteed to monotonically increase as the detectability δ is increased.

Results. In Figure 4, we present AUC scores for different statistics as δ is varied. As expected, KS arrival accurately identifies sequences that come from $\mathbb{Q}$ where the absolute time of events are non-uniform (as in INHOMOGENEOUS). Similarly, KS inter-event is good at detecting deviations in the distribution of inter-event times, as in RENEWAL. The performance of the chi-squared statistic is similar to that of KS arrival. Nevertheless, the above statistics fail when the expected number of

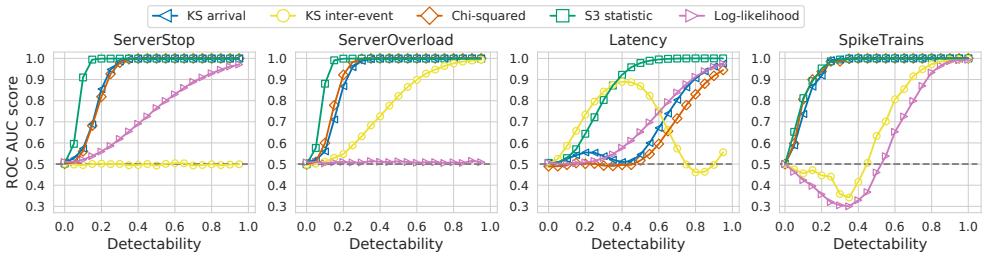

Figure 5: OoD detection on simulated data using different test statistics, measured with with ROC AUC (higher is better). See Section 6.2 for the description of the experimental setup.

events, N, changes substantially—as in KS arrival and chi-squared on RATE, and KS inter-event on STOPPING. These failure modes match our discussion from Section 3.

In contrast, the 3S statistic stands out as the most consistent test (best or close-to-best performance in 5 out of 6 cases) and does not completely fail in any of the scenarios. The relatively weaker performance on SELFCORRECTING implies that the 3S statistic is less sensitive to superuniform spacings (D'Agostino, 1986) than to imbalanced spacings. The results show that the 3S statistic is able to detect deviations w.r.t. both the event count N (RATE and STOPPING), as well as the distributions of the inter-event times w_i (RENEWAL) or the arrival times v_i (HAWKES and INHOMOGENEOUS)—something that other GoF statistics for the SPP cannot provide.

6.2 Detecting anomalies in simulated data

In this section, we test the OoD detection approach discussed in Section 4.2, i.e., we perform anomaly detection for a TPP with an *unknown* compensator. This corresponds to the hypothesis test in Equation 1. We use the training set $\mathcal{D}_{\text{train}}$ to fit an RNN-based neural TPP model (Shchur et al., 2020) via maximum likelihood estimation. Then, we define test statistics for the general TPP as follows. We apply the compensator Λ^* of the learned model to each event sequence X and compute the four statistics for the SPP from Section 6.1 on the transformed sequence $Z = \Lambda^*(X)$. We highlight that these methods are not "baselines" in the usual sense—the idea of combining a GoF statistic with a learned TPP model to detect anomalous event sequences is itself novel and hasn't been explored by earlier works. The rest of the setup is similar to Section 6.1. We use $\mathcal{D}_{\text{train}}$ to compute the EDF of each statistic under H_0, and then compute the ROC AUC scores on the p-values. In addition to the four statistics discussed before, we consider a two-sided test on the log-likelihood $\log q(X)$ of the learned generative model, which corresponds to the approach by Nalisnick et al. (2019).

Datasets. Like before, we define a detectability parameter δ for each scenario that determines how dissimilar ID and OoD sequences are. SERVER-STOP, SERVER-OVERLOAD and LATENCY are inspired by applications in DevOps, such as detecting anomalies in server logs. SERVER-OVERLOAD and SERVER-STOP contain data generated by a multivariate Hawkes process with 3 marks, e.g., modeling network traffic among 3 hosts. In OoD sequences, we change the influence matrix to simulate scenarios where a host goes offline (SERVER-STOP), and where a host goes down and the traffic is routed to a different host (SERVER-OVERLOAD). Higher δ implies that the change in the influence matrix happens earlier in time. LATENCY contains events of two types, sampled as follows. The first mark, the "trigger," is sampled from a homogeneous Poisson process with rate $\mu = 3$. The arrival times of the second mark, the "response," are obtained by shifting the times of the first mark by an offset sampled i.i.d. from $\text{Normal}(\mu = 1, \sigma = 0.1)$. In OoD sequences, the delay is increased by an amount proportional to δ. OoD examples, in this case, emulate an increased latency in the system. SPIKETRAINS (Stetter et al., 2012) contains sequences of firing times of 50 neurons, each represented by a distinct mark. We generate OoD sequences by shuffling the indices of k neurons (e.g., switching marks 1 and 2), where higher detectability δ implies more switches k. In this experiment we study how different statistics behave for TPPs with a large number of marks.

Results are shown in Figure 5. The 3S statistic demonstrates excellent performance in all four scenarios, followed by KS arrival and chi-squared. KS inter-event and log-likelihood statistics completely fail on SERVER-STOP and SERVER-OVERLOAD, respectively. These two statistics also struggle to discriminate OoD sequences in LATENCY and SPIKETRAINS scenarios. The non-

Table 1: ROC AUC scores for OoD detection on real-world datasets (mean & standard error are computed over 5 runs). Best result in **bold**, results within 2 pp. of the best underlined.

	KS arrival	KS inter-event	Chi-squared	Log-likelihood	3S statistic
LOGS — Packet corruption (1%)	57.4 ± 1.7	62.1 ± 0.9	66.6 ± 1.8	75.9 ± 0.1	$\mathbf{95.5} \pm 0.3$
LOGS — Packet corruption (10%)	59.2 ± 2.3	97.8 ± 0.6	59.1 ± 2.3	99.0 ± 0.0	$\mathbf{99.4} \pm 0.1$
LOGS — Packet duplication (1%)	81.1 ± 5.2	82.8 ± 5.0	74.6 ± 6.5	88.1 ± 0.1	$\mathbf{90.9} \pm 0.3$
LOGS — Packet delay (frontend)	95.6 ± 1.2	98.9 ± 0.4	$\mathbf{99.3} \pm 0.1$	90.9 ± 0.0	97.6 ± 0.1
LOGS — Packet delay (all services)	$\mathbf{99.8} \pm 0.0$	94.7 ± 1.1	$\mathbf{99.8} \pm 0.0$	96.1 ± 0.0	99.6 ± 0.1
STEAD — Anchorage, AK	59.6 ± 0.2	79.7 ± 0.1	67.4 ± 0.2	88.0 ± 0.1	$\mathbf{88.3} \pm 0.6$
STEAD — Aleutian Islands, AK	53.8 ± 0.5	88.8 ± 0.3	62.2 ± 0.9	97.0 ± 0.0	$\mathbf{99.8} \pm 0.0$
STEAD — Helmet, CA	59.1 ± 0.9	$\mathbf{98.7} \pm 0.0$	70.0 ± 0.6	96.9 ± 0.0	92.6 ± 0.3

monotone behavior of the ROC AUC scores for some statistics (as the δ increases) indicates that these statistics are poorly suited for the respective scenarios.

6.3 Detecting anomalies in real-world data

Finally, we apply our methods to detect anomalies in two real-world event sequence datasets. LOGS: We generate server logs using Sock Shop microservices (Weave, 2017) and represent them as marked event sequences. Sock Shop is a standard testbed for research in microservice applications (Aderaldo et al., 2017) and contains a web application that runs on several containerized services. We generate OoD sequences by injecting various failures (e.g., packet corruption, increased latency) among these microservices using a chaos testing tool Pumba (Ledenev et al., 2016). We split one large server log into 30-second subintervals, that are then partitioned into train and test sets.

STEAD (Stanford Earthquake Dataset) (Mousavi et al., 2019) includes detailed seismic measurements on over 1 million earthquakes. We construct four subsets, each containing 72-hour subintervals in a period of five years within a 350km radius of a fixed geographical location. We treat sequences corresponding the San Mateo, CA region as in-distribution data, and the remaining 3 regions (Anchorage, AK, Aleutian Islands, AK and Helmet, CA) as OoD data.

Results. Table 1 shows the ROC AUC scores for all scenarios. KS arrival and chi-squared achieve surprisingly low scores in 6 out of 8 scenarios, even though these two methods showed strong results on simulated data in Sections 6.1 and 6.2. In contrast, KS inter-event and log-likelihood perform better here than in previous experiments, but still produce poor results on Packet corruption. The 3S statistic is the only method that consistently shows high ROC AUC scores across all scenarios.

7 Discussion

Limitations. Our approach assumes that the sequences in $\mathcal{D}_{\text{train}}$ were drawn i.i.d. from the true data-generating distribution $\mathbb{P}_{\text{data}}$ (Section 2). This assumption can be violated in two ways: some of the training sequences might be anomalous or there might exist dependencies between them. We have considered the latter case in our experiments on SPIKETRAINS and LOGS datasets, where despite the non-i.i.d. nature of the data our method was able to accurately detect anomalies. However, there might exist scenarios where the violation of the assumptions significantly degrades the performance.

No single test statistic can be "optimal" for either OoD detection or GoF testing, since we make no assumptions about the alternative distribution $\mathbb{Q}$ (Section 2). We empirically showed that the proposed 3S statistic compares favorably to other choices over a range of datasets and applications domains. Still, for any *fixed* pair of distributions $\mathbb{P}$ and $\mathbb{Q}$, one can always find a statistic that will have equal or higher power s.t. the same false positive rate (Neyman & Pearson, 1933). Hence, it won't be surprising to find cases where our (or any other chosen a priori) statistic is inferior.

Broader impact. Continuous-time variable-length event sequences provide a natural representation for data such as electronic health records (Enguehard et al., 2020), server logs (He et al., 2016) and user activity traces (Zhu et al., 2020). The ability to perform unsupervised anomaly detection in such data can enable practitioners to find at-risk patients, reduce DevOps costs, and automatically detect security breaches—all of which are important tasks in the respective fields. One of the risks when applying an anomaly detection method in practice is that the statistical anomalies found by the method will not be relevant for the use case. For example, when looking for health insurance

fraud, the method might instead flag legitimate patients who underwent atypically many procedures as "suspicious" and freeze their accounts. To avoid such situations, automated decisions systems should be deployed with care, especially in sensitive domains like healthcare.

Conclusion. We have presented an approach for OoD detection for temporal point processes based on goodness-of-fit testing. At the core of our approach lies a new GoF test for standard Poisson processes based on the 3S statistic. Our method applies to a wide class of TPPs and is extremely easy to implement. We empirically showed that the proposed approach leads to better OoD detection accuracy compared to both popular GoF statistics for TPPs (Kolmogorov–Smirnov, chi-squared) and approaches commonly used in OoD detection literature (model log-likelihood). While our analysis focuses on TPPs, we believe our discussion on similarities and distinctions between GoF testing and OoD detection offers insights to the broader machine learning community.

References

Aderaldo, C. M., Mendonça, N. C., Pahl, C., and Jamshidi, P. Benchmark requirements for microservices architecture research. In *International Workshop on Establishing the Community-Wide Infrastructure for Architecture-Based Software Engineering*, 2017.

Alizadeh, M., Scaglione, A., Davies, J., and Kurani, K. S. A scalable stochastic model for the electricity demand of electric and plug-in hybrid vehicles. *IEEE Transactions on Smart Grid*, 5(2), 2013.

Barnard, G. Time intervals between accidents—a note on Maguire, Pearson and Wynn's paper. *Biometrika*, 40(1-2):212–213, 1953.

Blum, A., Hopcroft, J., and Kannan, R. *Foundations of data science*. 2016.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, 14(2), 2002.

Choi, H., Jang, E., and Alemi, A. A. WAIC, but why? Generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2016.

Cox, D. R. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B*, 17(2), 1955.

Cox, D. R. The statistical analysis of series of events. *Monographs on Applied Probability and Statistics*, 1966.

D'Agostino, R. B. *Goodness-of-fit-techniques*. 1986.

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes, volume 1: Elementary theory and methods*. 2003.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. Recurrent marked temporal point processes: Embedding event history to vector. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

Enguehard, J., Busbridge, D., Bozson, A., Woodcock, C., and Hammerla, N. Neural temporal point processes for modelling electronic health records. In *Machine Learning for Health*, 2020.

Ferragut, E. M., Laska, J., and Bridges, R. A. A new, principled approach to anomaly detection. In *International Conference on Machine Learning and Applications*, 2012.

Fisher, R. A. Design of experiments. *British Medical Journal*, 1(3923):554–554, 1936.

Fisher, R. A. Answer to question 14 on combining independent tests of significance. 1948.

Gerhard, F. and Gerstner, W. Rescaling, thinning or complementing? On goodness-of-fit procedures for point process models and generalized linear models. In *Advances in Neural Information Processing Systems*, 2010.

Gerhard, F., Haslinger, R., and Pipa, G. Applying the multivariate time-rescaling theorem to neural population models. *Neural computation*, 23(6), 2011.

Greenwood, M. The statistical study of infectious diseases. *Journal of the Royal Statistical Society: Series A*, 109:85–110, 1946.

He, S., Zhu, J., He, P., and Lyu, M. R. Experience report: System log analysis for anomaly detection. In *International Symposium on Software Reliability Engineering*, 2016.

Hill, I. Approximating the distribution of Greenwood's statistic with Johnson distributions. *Journal of the Royal Statistical Society: Series A*, 142(3):378–380, 1979.

Janzing, D., Budhathoki, K., Minorics, L., and Blöbaum, P. Causal structure based root cause analysis of outliers. *arXiv:1912.02724*, 2019.

Kim, S.-H. and Whitt, W. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management*, 16(3), 2014.

Ledenev, A. et al. Pumba: Chaos testing tool for Docker. https://github.com/alexei-led/pumba, 2016.

Lewis, P. A. W. Some results on tests for Poisson processes. *Biometrika*, 52(1/2), 1965.

Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., and Song, L. Learning temporal point processes via reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2018.

Liu, Q., Lee, J., and Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, 2016.

Liu, S. and Hauskrecht, M. Detection of outlier events in continuous-time event sequences. *arXiv preprint arXiv:1912.09522*, 2019.

Marsaglia, G., Tsang, W. W., Wang, J., and Others. Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8(18), 2003.

Moran, P. The random division of an interval. *Supplement to the Journal of the Royal Statistical Society*, 9(1), 1947.

Morningstar, W. R., Ham, C., Gallagher, A. G., Lakshminarayanan, B., Alemi, A. A., and Dillon, J. V. Density of states estimation for out-of-distribution detection. *International Conference on Artificial Intelligence and Statistics*, 2021.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, 2019.

Nalisnick, E., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Neyman, J. and Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London: Series A*, 231(694-706):289–337, 1933.

North, B. V., Curtis, D., and Sham, P. C. A note on the calculation of empirical p-values from Monte Carlo procedures. *The American Journal of Human Genetics*, 71(2), 2002.

Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401), 1988.

Ojeda, C. A. M., Cvejoski, K., Sifa, R., Schuecker, J., and Bauckhage, C. Patterns and outliers in temporal point processes. In *SAI Intelligent Systems Conference*, 2019.

Pasupathy, R. Generating homogeneous Poisson processes. *Wiley encyclopedia of operations research and management science*, 2010.

Pillow, J. Time-rescaling methods for the estimation and assessment of non-Poisson neural encoding models. *Advances in Neural Information Processing Systems*, 22, 2009.

Rao, J. and Kuo, M. Asymptotic results on the Greenwood statistic and some of its generalizations. *Journal of the Royal Statistical Society: Series B*, 46(2):228–237, 1984.

Rasmussen, J. G. Lecture notes: Temporal point processes and the conditional intensity function. *arXiv preprint arXiv:1806.00221*, 2018.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2019.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Shchur, O., Biloš, M., and Günnemann, S. Intensity-free learning of temporal point processes. In *International Conference on Learning Represenations*, 2020.

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. Neural temporal point processes: A review. *International Joint Conference on Artificial Intelligence*, 2021.

Stephens, M. A. Further percentage points for Greenwood's statistic. *Journal of the Royal Statistical Society: Series A*, 144(3):364–366, 1981.

Stetter, O., Battaglia, D., Soriano, J., and Geisel, T. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Computational Biology*, 8(8), 2012.

Tao, L., Weber, K. E., Arai, K., and Eden, U. T. A common goodness-of-fit framework for neural population models using marked point process time-rescaling. *Journal of Computational Neuroscience*, 45(2), 2018.

Wang, Z., Dai, B., Wipf, D., and Zhu, J. Further analysis of outlier detection with deep generative models. *Advances in Neural Information Processing Systems*, 2020.

Weave. Sock shop : A microservice demo application. https://github.com/microservices-demo/microservices-demo, 2017.

Wei, S., Zhu, S., Zhang, M., and Xie, Y. Goodness-of-fit test for mismatched self-exciting processes. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Yang, J., Rao, V., and Neville, J. A Stein–Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Zhu, S., Yuchi, H. S., and Xie, Y. Adversarial anomaly detection for marked spatio-temporal streaming data. In *International Conference on Acoustics, Speech and Signal Processing*, 2020.

A Difference between GoF testing and OoD detection

The connection between OoD detection and GoF testing was first pointed out by Nalisnick et al. (2019). They proposed to perform a GoF test for a deep generative model to detect OoD instances. However, as we explained in Section 2, these two problems are in fact *not* equivalent. We now demonstrate how this insight allows us to explain and improve upon some results obtained by Nalisnick et al. (2019).

First, we consider the **Gaussian annulus test** for normalizing flow models that was also used by Choi et al. (2018). A normalizing flow model $\mathbb{P}_{\text{model}}$ defines the distribution of a D-dimensional random vector X by specifying a diffeomorphism $f\colon \mathbb{R}^D \to \mathbb{R}^D$, such that $Z = f(X)$ is distributed according to $\mathcal{N}(\mathbf{0}_D, \boldsymbol{I}_D)$, the standard normal distribution. In other words, $f(X)|X \sim \mathbb{P}_{\text{model}}$ follows the standard normal distribution, so any test for the normal distribution can be used to test the GoF of a normalizing flow model. Based on this, Nalisnick et al. (2019) define the following test statistic

$$\phi(X) = \Big| \|f(X)\|_2 - \mathbb{E}_{X\sim\mathbb{P}_{\text{model}}}[\|f(X)\|_2] \Big| = \Big| \|f(X)\|_2 - \sqrt{D} \Big|. \tag{9}$$

The idea here is to replace a two-sided test on the statistic $\|f(X)\|_2$ with a one-sided test on the statistic $\phi(X)$ defined above. Since $f(X)|X \sim \mathbb{P}_{\text{model}}$ follows the standard normal distribution, the statistic $\phi(X)|H_0$ will concentrate near 0 (Blum et al., 2016, Theorem 2.9). Therefore, checking if $\phi(X)$ is below a certain threshold ϵ is equivalent to performing the GoF null hypothesis test (Equation 2).

However, the above approach will not work for an OoD detection hypothesis test (Equation 1). If we learn a model $\mathbb{P}_{\text{model}}$ on training instances $\mathcal{D}_{\text{train}}$ that were generated by some distribution $\mathbb{P}_{\text{data}}$, we will in general have $\mathbb{P}_{\text{model}} \neq \mathbb{P}_{\text{data}}$. This implies that $f(X)|X \sim \mathbb{P}_{\text{data}}$ will not follow the standard normal distribution. Therefore, $\mathbb{E}_{X\sim\mathbb{P}_{\text{data}}}[\|f(X)\|_2] \neq \sqrt{D}$ and the distribution of $\|f(X)\|_2$ might not even be symmetric around its mean. This means we cannot replace a two-sided test on $\|f(X)\|_2$ with a one-sided test on $\phi(X)$ when doing OoD detection. A better idea is to directly compute the two-sided p-value for the OoD detection test using the statistic $\|f(X)\|_2$, following our approach in Section 2.

Similarly, for the (single-instance) **typicality test**, the test statistic is defined as

$$\gamma(X) = \Big| \log q(X) - \mathbb{E}_{X\sim\mathbb{P}_{\text{model}}}[\log q(X)] \Big|, \tag{10}$$

where $\log q(X)$ is the log-likelihood of a generative model trained on $\mathcal{D}_{\text{train}}$. This leads to the same problems when trying to apply this statistic for OoD detection as we encountered with the Gaussian annulus test above—the expected value $\mathbb{E}_{X\sim\mathbb{P}_{\text{model}}}[\log q(X)]$ is only suitable for a GoF test. However, in this case Nalisnick et al. (2019) report that they found $\mathbb{E}_{X\sim\mathbb{P}_{\text{data}}}[\log q(X)]$ to work better in practice. By drawing a clear distinction between the OoD detection test and the GoF test we can explain this empirical result. An even better idea is to use the two-sided p-value (Equation 3) instead of Equation 10, since the distribution of the statistic $\log q(X)|X \sim \mathbb{P}_{\text{data}}$ is not guaranteed to be symmetric.

B Other statistics based on squared spacings

The following discussion is based on Moran (1947) and D'Agostino (1986).

Sum-of-squared spacings (3S) statistic for $\text{Uniform}([0,1])$**.** Suppose that $\{u_1, \dots, u_N\}$ are sampled i.i.d. from the $\text{Uniform}([0,1])$ distribution. Additionally, assume w.l.o.g. that the u_i's are sorted in an increasing order, i.e., $u_1 \leq \cdots \leq u_N$. The 3S statistic for $\text{Uniform}([0,1])$ is defined as

$$\psi_N^{\text{Unif}([0,1])} = N \sum_{i=1}^{N+1} (u_i - u_{i-1})^2, \tag{11}$$

where $u_0 = 0$ and $u_{N+1} = 1$. The factor N ensures that $\psi_N^{\text{Unif}([0,1])}$ approaches the standard normal distribution as $N \to \infty$. However, the convergence of $\psi_N^{\text{Unif}([0,1])}$ to its limiting distribution is rather slow.

3S statistic for $\mathrm{Uniform}([0, V])$**.** The statistic above can be generalized to the uniform distribution on an arbitrary interval $[0, V]$. Suppose $\{v_1, \ldots, v_N\}$ are drawn i.i.d. from the $\mathrm{Uniform}([0, V])$ distribution, and again are sorted in an increasing order. The 3S statistic for $\mathrm{Uniform}([0, V])$ is defined by simply dividing the v_i's by the interval length V.

$$\begin{aligned} \psi_N^{\mathrm{Unif}([0,V])} &= N \sum_{i=1}^{N+1} \left(\frac{v_i}{V} - \frac{v_{i-1}}{V} \right)^2 \\ &= \frac{N}{V^2} \sum_{i=1}^{N+1} (v_i - v_{i-1})^2, \end{aligned} \tag{12}$$

where $v_0 = 0$ and $v_{N+1} = V$.

3S statistic for the SPP on $[0, V]$**.** Remember that the N factor makes the distribution of $\psi_N^{\mathrm{Unif}([0,V])}$ (asymptotically) invariant for different values of N. However, we don't want our test statistic for the SPP to be invariant w.r.t. the event count N. Therefore, we define the *3S statistic for the standard Poisson process* by replacing N with its expectation $\mathbb{E}[N|V] = V$.

$$\begin{aligned} \psi^{\mathrm{SPP}([0,V])} &= \frac{\mathbb{E}[N|V]}{V^2} \sum_{i=1}^{N+1} (v_i - v_{i-1})^2 \\ &= \frac{1}{V} \sum_{i=1}^{N+1} (v_i - v_{i-1})^2 \end{aligned} \tag{13}$$

This is the definition that we introduced in Equation 7. As we show in Sections 4 and 6, the 3S statistic for the SPP allows us to detect a broad class of anomalies (i.e., deviations from the SPP) that differ both in the distribution of the event count N as well as the arrival times v_i.

C Proof of Proposition 1

To compute the moments of the 3S statistic for the standard Poisson process (Equation 7) we need to marginalize out the event count N, which is equivalent to applying the law of iterated expectation

$$\begin{aligned} \mathbb{E}[f(\psi)|T] &= \sum_{n=0}^{\infty} \mathbb{E}[f(\psi)|N = n, V] \Pr(N = n|V) \\ &= \sum_{n=0}^{\infty} \mathbb{E}[f(\psi)|N = n, V] \frac{V^n e^{-V}}{n!} \end{aligned} \tag{14}$$

where we used the fact that $N|V \sim \mathrm{Poisson}(V)$.

We obtain the expectations of ψ and ψ^2 conditioned on N and V using the result by Moran (1947) on the moments of $\psi_N^{\mathrm{Unif}([0,1])}$ (Equation 11), the 3S statistic for the $\mathrm{Uniform}([0, 1])$ distribution.

$$\begin{aligned} \mathbb{E}[\psi|N = n, V] &= \frac{2V}{(n+2)} \\ \mathbb{E}[\psi^2|N = n, V] &= \frac{4V^2(n+6)}{(n+2)(n+3)(n+4)} \end{aligned} \tag{15}$$

These can also be easily derived from the moments of the Dirichlet distribution, by using the fact that the scaled inter-event times $(w_1/V, \ldots, w_{N+1}/V)$ are distributed uniformly on the standard N-simplex (i.e., according to Dirichlet distribution with parameter $\boldsymbol{\alpha} = \mathbf{1}_{N+1}$).

By plugging in Equation 15 into Equation 14, we obtain

$$\begin{aligned} \mathbb{E}[\psi|V] &= 2Ve^{-V} \sum_{n=0}^{\infty} \frac{V^n}{n!(n+2)} \\ &= 2Ve^{-V} \frac{1}{V^2} \left(e^V (V-1) + 1 \right) \\ &= \frac{2}{V} (V + e^{-V} - 1). \end{aligned} \tag{16}$$

Similarly, we compute the non-centered second moment as

$$\begin{aligned}\mathbb{E}[\psi^2|V] &= 4V^2e^{-V}\sum_{n=0}^{\infty}\frac{V^n(n+6)}{n!(n+2)(n+3)(n+4)} \\ &= 4V^2e^{-V}\frac{1}{V^4}\left(e^V(V^2-6)+2(V^2+3V+3)\right) \\ &= \frac{4}{V^2}\left(V^2-6+2e^{-V}(V^2+3V+3)\right).\end{aligned} \tag{17}$$

Finally, we obtain the variance as

$$\begin{aligned}\mathrm{Var}[\psi|V] &= \mathbb{E}[\psi^2|V]-\mathbb{E}[\psi|V]^2 \\ &= \frac{4}{V^2}\left(2V-7+e^{-V}(2V^2+4V+8-e^{-V})\right).\end{aligned}$$

Higher-order moments of $\psi|V$ can be computed similarly using Equation 14.

D Implementation details

The following code describes the procedure for computing the p-values for both hypothesis tests discussed in Section 2—namely, the GoF test (Equation 2) and the OoD detection test (Equation 1). The code below is for demonstration purposes only, the actual implementation used in our experiments is better optimized.

```
def compute_p_value(x_test, samples, score_fn):
    scores_id = [score_fn(x) for x in samples]
    score_x = score_fn(x_test)
    num_train = len(samples)
    num_above = 0
    for s in scores_id:
        if s > score_x:
            num_above += 1
    num_below = num_train - num_above
    return min(
        (num_below + 1) / (num_train + 1),
        (num_above + 1) / (num_train + 1)
    )
```

The `+1` correction in the numerator and denominator for the p-value computation is done as described by North et al. (2002). If we define `samples` as the set of in-distribution sequences $\mathcal{D}_{\text{train}}$ that were generated from $\mathbb{P}_{\text{data}}$, we recover the OoD detection test (Equation 1 and Section 4.2). If we define `samples` as the set of sequences $\mathcal{D}_{\text{model}}$ that were generated from $\mathbb{P}_{\text{model}}$, we recover the GoF test (Equation 2 and Section 4.1).

In the snippet above, `score_fn` corresponds to a test statistic $s\colon \mathcal{X} \to \mathbb{R}$. In our experiments, we consider the following choices for s:

1. KS arrival (Equation 5).
2. KS inter-event (Equation 6).
3. Chi-squared: we partition the interval $[0, V]$ into $B = 10$ disjoint buckets of equal length, and compare the observed event count N_b in each bucket with the expected amount $L = V/B$

$$\chi^2(Z) = \sum_{b=1}^{B}\frac{(N_b-L)^2}{L}. \tag{18}$$

4. Sum-of-squared spacings (Equation 7).
5. Log-likelihood

$$\log q(X) = \sum_{i=1}^{N}\log\frac{\partial\Lambda^*(t_i)}{\partial t_i} - \Lambda^*(T). \tag{19}$$

All these statistics are computed based on some TPP model with compensator Λ^*. For statistic 1–4, we compute $s(X)$ by first obtaining the transformed sequence $Z = (\Lambda^*(t_1), \ldots, \Lambda^*(T))$ and then evaluating the respective SPP statistic on Z. The log-likelihood is directly evaluated based on the model's conditional intensity.

Marked sequences. In a marked sequence $X = \{(t_1, m_1), \ldots, (t_N, m_N)\}$ each event is represented by a categorical mark $m_i \in \{1, \ldots, K\}$ in addition to the arrival time t_i. A marked TPP model is specified by K compensators $\{\Lambda_1^*, \ldots, \Lambda_K^*\}$.

We obtain the transformed sequence Z necessary for statistics 1–4 as follows. Let $\left(t_1^{(k)}, \ldots, t_{N_k}^{(k)}\right)$ denote the events of mark k in a given sequence X. For each mark $k \in \{1, \ldots, K\}$, we obtain a transformed sequence $Z^{(k)} = \left(\Lambda_k^*(t_1^{(k)}), \ldots, \Lambda_k^*(t_{N_k}^{(k)}), \Lambda_k^*(T)\right)$. Then we concatenate the transformed sequences for each mark, thus obtaining a single SPP realization on the interval $[0, \sum_{k=1}^K \Lambda_k^*(T)]$. For example, suppose the transformed sequence for the first mark is $Z^{(1)} = (1.0, 2.5, 4.0)$ and for the second mark $Z^{(2)} = (0.5, 3.0)$. Then the concatenated sequence will be $Z = (0.0, 1.0, 2.5, 4.0 + 0.5, 4.0 + 3.0) = (0.0, 1.0, 2.5, 4.5, 7.0)$. Our approach based on concatenating the $Z^(k)$'s is simpler than other methods for combining multiple sequences by Gerhard et al. (2011) & Tao et al. (2018), and we found ours to work well in practice.

The log-likelihood for a marked sequence is computed as

$$\log q(X) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{1}(m_i = k) \log \frac{\partial \Lambda_k^*(t_i)}{\partial t_i} - \sum_{k=1}^{K} \Lambda_k^*(T). \tag{20}$$

E Datasets

E.1 Standard Poisson process

In-distribution sequences (corresponding to $\mathbb{P}_{\text{model}}$) are all generated from an SPP (i.e., a homogeneous Poisson process with rate $\mu = 1$) on the interval $[0, 100]$. The OoD sequences (corresponding to $\mathbb{Q}$) for each of the scenarios are generated as follows, where $\delta \in [0, 1]$ is the detectability parameter.

(i) RATE: homogeneous Poisson process with rate $\mu = 1 - 0.5\delta$.

(ii) STOPPING: We generate a sequence $X = (t_1, \ldots, t_N)$ from an SPP and then remove all the events $t_i \in [t_{\text{stop}}, T]$, where we compute $t_{\text{stop}} = T(1 - 0.3\delta)$.

(iii) RENEWAL: A renewal process, where the inter-event times τ_i are sampled i.i.d. from a Gamma distribution with shape $k = 1 - \delta$ and scale $\theta = \frac{1}{1-\delta}$. Thus, the expected inter-event time stays the same, but the variance of inter-event times increases for higher δ.

(iv) HAWKES: Hawkes process with conditional intensity $\lambda^*(t) = \mu + \alpha \sum_{t_j < t} \exp(-(t - t_j))$. The parameters are chosen as $\mu = 1 - \delta$ and $\alpha = \delta$.

(v) INHOMOGENEOUS: inhomogeneous Poisson process with intensity $\lambda(t) = 1 + \beta \sin(\omega t)$, where $\omega = 2\pi/50$ and $\beta = 2\delta$.

(vi) SELFCORRECTING: self-correcting process with intensity $\lambda^*(t) = \exp\left(\mu t - \sum_{t_j < t} \alpha\right)$, where we set $\mu = \delta + 10^{-5}$ and $\alpha = \delta$.

In all above scenarios setting $\delta = 0$ recovers the standard Poisson process, thus making $\mathbb{P}_{\text{data}}$ and $\mathbb{Q}$ indistinguishable. Note that the parameters in scenarios (iii)–(vi) are chosen such that the expected number of events N is always equal to T, like in the SPP. For all scenarios, $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}^{\text{ID}}$ and $\mathcal{D}_{\text{test}}^{\text{OOD}}$ consist of 1000 sequences each.

Additional experiments. For completeness, we consider two more scenarios.

(vii) INCREASINGRATE: Similar to scenario (i), but now the rate is increasing instead as $\mu = 1 + 0.5\delta$.

(viii) RENEWALB: Similar to scenario (iii), but the variance now decreases for higher δ. For this we define the parameters of the Gamma distribution as $k = \frac{1}{1-\delta}$ and $\theta = 1 - \delta$.

The results are shown in Figure 6. As we can see, the same qualitative conclusions apply here as for the experiments in Section 6.1.

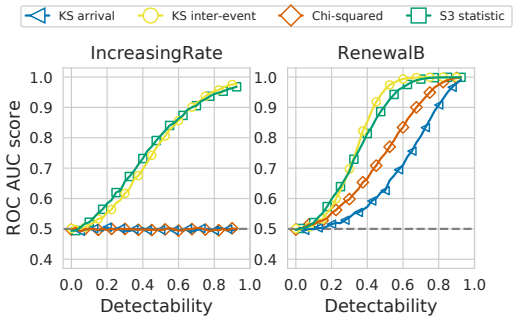

Figure 6: GoF testing for the SPP using additional scenarios.

E.2 Simulated data

SERVER-STOP and SERVER-OVERLOAD: In-distribution sequences for both scenarios are generated by a multivariate Hawkes process with $K = 3$ marks on the interval $[0, 100]$ with following base rates $\boldsymbol{\mu}$ and influence matrix $\boldsymbol{A}$:

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

This scenario represents communication between a server (mark 1) and two worker machines (marks 2 and 3)—events for the workers can only be triggered by incoming requests from the server.

In OoD sequences, the structure of the influence matrix is changed at time $t_{\text{stop}} = T(1 - 0.5\delta)$, which represents the time of a failure in the system. For SERVER-STOP, the influence matrix is changed to $\boldsymbol{A}^{\text{stop}}$, and for SERVER-OVERLOAD the influence matrix is changed to $\boldsymbol{A}^{\text{overload}}$.

$$\boldsymbol{A}^{\text{stop}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad \boldsymbol{A}^{\text{overload}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}$$

The sets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}^{\text{ID}}$ and $\mathcal{D}_{\text{test}}^{\text{OOD}}$ consist of 1000 sequences each.

LATENCY: Event sequences consist of two marks. ID sequences are generated as follows. Events of the first mark ("the trigger") are generated by a homogeneous Poisson process with rate $\mu = 3$. Events of the the second mark ("the response") are obtained by shifting the arrival times of the first mark by offsets that are sampled i.i.d. from $\text{Normal}(\mu = 1, \sigma = 0.1)$. In OoD sequences, the offsets are instead sampled from $\text{Normal}(\mu = 1 + 0.5\delta, \sigma = 0.1)$. That is, OoD sequences correspond to increased latency between the "trigger" and "response" events. The sets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}^{\text{ID}}$ and $\mathcal{D}_{\text{test}}^{\text{OOD}}$ consist of 1000 sequences each.

SPIKETRAINS: The original fluorescence data is provided at `www.kaggle.com/c/connectomics`. We extracted the spike times from the fluorescence recordings using the code by `https://github.com/slinderman/pyhawkes/tree/master/data/chalearn`. We dequantized the discrete spike times by adding $\text{Uniform}(-0.5, 0.5)$ noise and selected the first 50 marks.

The original data consists of a single sequence that is 3590 seconds long. We split the long sequence into overlapping windows that are 20 seconds long. We select the first 500 sequences for training (as $\mathcal{D}_{\text{train}}$), and 96 remaining sequences for testing (as $\mathcal{D}_{\text{test}}^{\text{ID}}$). OoD sequences (i.e., $\mathcal{D}_{\text{test}}^{\text{OOD}}$) are obtained by switching $k = \lfloor \delta K \rfloor$ marks. For example, if marks 5 and 10 are switched, all events that correspond to mark 5 in $\mathcal{D}_{\text{test}}^{\text{ID}}$ will be labeled as mark 10 in $\mathcal{D}_{\text{test}}^{\text{OOD}}$, and vice versa.

E.3 Real-world data

LOGS: We ran the Sock Shop microservices testbed (Weave, 2017) on our in-house server. We consider the logs corresponding to the `user` service. There are 4 types of log entries that we model

as 4 categorical marks. We use the timestamps of log entries as arrival times of a TPP. We slice the logs into 30-second-long non-overlapping windows, each corresponding to a single TPP realization.

We run the service for ≈14 hours to generate training data, and then for additional ≈5 hours to generate test data. The test data contains 5 types of injected anomalies produced by Pumba (Ledenev et al., 2016). See Table 1 for the list of anomalies. Each anomaly injection lasts 10 minutes. We mark a test sequence as OoD if the system was "attacked" by Pumba during the respective time window. In total, we use 1668 sequences as $\mathcal{D}_{\text{train}}$, 502 sequences as $\mathcal{D}_{\text{test}}^{\text{ID}}$, and 22 sequences as $\mathcal{D}_{\text{test}}^{\text{OOD}}$ for each of the attack scenarios (i.e., 110 OoD sequences in total).

STEAD: The original dataset by Mousavi et al. (2019) contains over 1 million earthquake recordings. We sample 72-hour sub-windows and treat times of earthquake as arrival times of a TPP, as usually done in seismological applications. We treat the sequences as unmarked. We select 4 geographic locations: (1) San Mateo, CA, (2) Anchorage, AK, (3) Aleutian Islands, AK, and (4) Hemet, CA. We group the earthquakes that happen within a 350 km radius (geodesic) around each of the locations, thus obtaining 4 sets of sequences (5000 sequences for each location). We use the sequences corresponding to (1) San Mateo, CA, as in-distribution data, and the remaining 3 locations as OoD data. We use 4000 ID sequences as $\mathcal{D}_{\text{train}}$, 1000 ID sequences and $\mathcal{D}_{\text{test}}^{\text{ID}}$, and 1000 sequences per each remaining location as $\mathcal{D}_{\text{test}}^{\text{OOD}}$.

F Experimental setup

F.1 GoF for the SPP (Section 6.1)

We compute the p-values for the GoF test using the procedure described in Appendix D. For the GoF test, we use 1000 event sequences generated by an SPP as `samples` in the algorithm. The test statistics are computed using the compensator of the SPP $\Lambda^*(t) = t$. We compute the p-value for each event sequence in $\mathcal{D}_{\text{test}}^{\text{ID}}$ and $\mathcal{D}_{\text{test}}^{\text{OOD}}$, and then compute the ROC AUC score based on these p-values. The results are averaged over 10 random seeds.

F.2 OoD detection (Sections 6.2 & 6.3)

We train a neural TPP model similar to Shchur et al. (2020). We parametrize the inter-event time distribution with a mixture of 8 Weibull distributions. The marks are conditionally independent of the inter-event times given the context embedding, as in the original model. Mark embedding size is set to 32, and the context embedding (i.e., RNN hidden size) is set to 64 for all experiments.

We optimize the model parameters by maximizing the log-likelihood of the sequences in $\mathcal{D}_{\text{train}}$ (batch size 64) using Adam with learning rate 10^{-3} and clipping the L_2-norm of the gradients to 5. We run the optimization procedure for up to 200 epochs, and perform early stopping if the training loss stops improving for 10 epochs. The p-values are computed according to the procedure described in Appendix D. The results reported in Section 6.2 are averaged over 10 random seeds. In Section 6.3, we train the neural TPP model with 5 different random initializations to compute the average and standard error in Table 1.

G Fisher's method for KS statistics

Here we show that ad-hoc fixes to the KS statistics that make them sensitive to the variations in the event count N lead to worse discriminative power in other scenarios. For this, we replicate the experimental setup from Section 6.1 with two additional statistics.

Fisher arrival. We compute the two-sided p-value p_N for the event count N using the CDF of the $\text{Poisson}(V)$ distribution. Then, we compute the two-sided p-value $p_{\kappa_{\text{arr}}}$ for KS arrival statistic (Equation 5) using Kolmogorov distribution. We combine the two p-values using Fisher's method (Fisher, 1948) as $-2(\log p_N + \log p_{\kappa_{\text{arr}}})$. **Fisher inter-event** is defined similarly using the two-sided p-value $p_{\kappa_{\text{int}}}$ for the KS inter-event statistic (Equation 6).

Results. Results are shown in Figure 7. We see that the Fisher versions of the statistics indeed fix the failure modes of the two KS scores on RATE and STOPPING, where the event count N changes in OoD sequences. However, the Fisher versions of the statistics perform worse than the respective

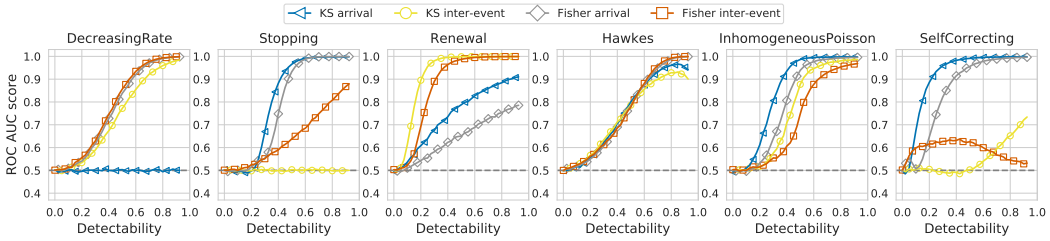

Figure 7: Comparing KS statistics with the respective Fisher versions that are sensitive to the event count N.

KS statistics in 3 out of 4 remaining scenarios. In contrast, the 3S statistic performs well both in scenarios where N changes, as well as when the distribution of the arrival/inter-event times is varied.