Statistical Evaluation of Anomaly Detectors for Sequences

Erik Scharwächter
scharwaechter@bit.uni-bonn.de
University of Bonn, Germany

Emmanuel Müller
mueller@bit.uni-bonn.de
University of Bonn, Germany

ABSTRACT

Although precision and recall are standard performance measures for anomaly detection, their statistical properties in sequential detection settings are poorly understood. In this work, we formalize a notion of precision and recall with temporal tolerance for point-based anomaly detection in sequential data. These measures are based on time-tolerant confusion matrices that may be used to compute time-tolerant variants of many other standard measures. However, care has to be taken to preserve interpretability. We perform a statistical simulation study to demonstrate that precision and recall may overestimate the performance of a detector, when computed with temporal tolerance. To alleviate this problem, we show how to obtain null distributions for the two measures to assess the statistical significance of reported results.

KEYWORDS

anomaly detection, precision, recall, confusion matrix, simulations

1 INTRODUCTION

Anomaly detection in sequential data is a highly active research topic [3, 10, 13, 17, 18, 23]. Precision and recall are two measures routinely used to evaluate the performance of anomaly detectors, both for iid data and for sequential data. An important characteristic of sequential data is that the decisions of a detector can be imprecise [1, 2] without impairing its practical utility: if an anomaly at time step t is detected at time step $t + 1$, this is still a useful result. Recently, Tatbul et al. [19] pointed out that the classical precision and recall measures, when applied to sequential detection problems, may misrepresent the performance of the detector. They introduced novel precision and recall measures for *range-based* anomaly detection. However, the problem persists even for *point-based* anomalies, where the ground-truth anomaly label is a single time step. In this work, we study in detail time-tolerant notions of precision and recall for *point-based* anomaly detection in sequential data, with a special focus on the statistical properties of these measures. Our
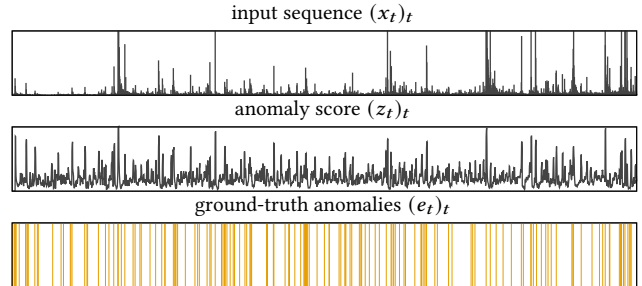
Figure 1: The running example for earthquake detection.

work is closely related to recent advances in the statistical association between event series and time series [4, 9, 14, 15, 20], and uses results from event coincidence analysis (ECA) [7, 12, 16].

1.1 Anomaly detection problem

We are given an input sequence $(x_t)_{t=1,...,T}$ over an arbitrary input domain. Furthermore, we are given an anomaly scoring function z to compute a numeric sequence of anomaly scores $(z_t)_{t=1,...,T}$ from the input sequence. If the observation at time step t is likely an anomaly, the anomaly score z_t should be high; if the observation at time step t appears normal, z_t should be low. An anomaly is predicted at time step t if the anomaly score is larger than some predefined threshold, $z_t \geq \tau$. The exact notion of what constitutes an anomaly is highly domain-specific and should be reflected in the choice of the anomaly scoring function. Anomaly detectors of this type are widely used across many disciplines. For example, Wiedermann et al. [21] use the clustering coefficient as an anomaly score for dynamic networks to detect El Niño events in climate data, and Earle et al. [8] use an energy transient score [22] to detect earthquakes from Twitter[1] time series.

We use the problem of earthquake detection on Twitter as the running example in this work. Figure 1 (top row) shows the daily volume of tweets that were posted in Germany between 2010 and 2017 and contain the word "earthquake," translated to various languages. The plot also shows all severe earthquakes that occurred globally in the same time period (bottom row). We obtained the Twitter data using the ForSight platform from Crimson Hexagon/Brandwatch[2], and the earthquake data from the International Disaster Database EM-DAT, provided by the Centre for Research on the Epidemiology of Disasters[3]. Our goal is to evaluate whether an anomaly detector on the Twitter time series has the potential to detect earthquakes globally. For this purpose, we stick to Earle et al. [8] and use the
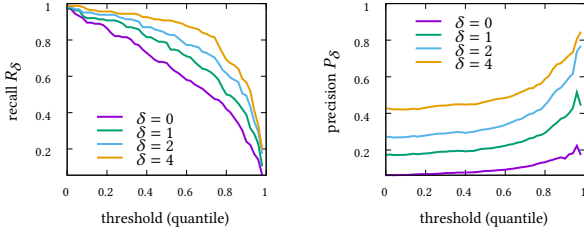
---

[1]https://www.twitter.com/
[2]https://www.brandwatch.com/
[3]https://www.emdat.be/

Figure 2: Precision and recall by threshold and tolerance δ.

energy transient score as the anomaly score (middle row), i.e., we set $z_t = STA_t/(LTA_t + 1)$, where STA is the short-term average of the input sequence over the past 3 days, while LTA is the long-term average over the past 14 days. The energy transient score reacts to drastic changes in the level of the time series.

1.2 Evaluation measures

While the anomaly score encodes the *feature* of interest that the anomaly detector should react to, the detection threshold τ controls the *precision and recall* of the anomaly detector. Let $(e_t)_{t=1,..,T}$ be a ground-truth sequence of actual anomalies, with value $e_t = 1$ if there is an actual anomaly at time step t, and $e_t = 0$ if there is no anomaly. In our example, actual anomalies correspond to reported earthquakes. We define precision P_0 and recall R_0 as

$$P_0 = \frac{\sum_t e_t \cdot \mathcal{I}(z_t \geq \tau)}{\sum_t \mathcal{I}(z_t \geq \tau)} \text{ and } R_0 = \frac{\sum_t e_t \cdot \mathcal{I}(z_t \geq \tau)}{\sum_t e_t}. \quad (1)$$

The function $\mathcal{I}(c)$ evaluates to 1 if and only if the condition c is true. The numerator is the number of true positives, while the denominator is either the number of predicted anomalies or the number of actual anomalies. The relationship between the threshold and precision and recall for the earthquake detection problem is visualized in Figure 2. The values for P_0 and R_0 from Equation 1 correspond to the lines labeled $\delta = 0$ in the plots. The results are not particularly good: we can obtain acceptable recall values at low thresholds, but the cost is an unacceptably low precision. We observe that recall is a monotonically decreasing function of the threshold τ: the number of true positives in the numerator decreases with increasing τ while the denominator stays constant. Precision, on the other hand, is a non-monotone function of the threshold, since both the numerator and the denominator change with τ.

2 TIME-TOLERANT MEASURES

2.1 Precision and recall

In sequential data, a predicted anomaly can often be considered a true positive if there is an actual anomaly *close to* the predicted time point. The higher the temporal tolerance, the higher the number of true positives, and the higher will be *both* precision and recall. We follow ECA [7, 14, 16] and define measures for precision P_δ and

recall R_δ with temporal tolerance δ as:

$$P_\delta = \frac{\sum_t \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right) \cdot \mathcal{I}(z_t \geq \tau)}{\sum_t \mathcal{I}(z_t \geq \tau)} \quad (2)$$

$$R_\delta = \frac{\sum_t e_t \cdot \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)}{\sum_t e_t}. \quad (3)$$

If $\delta = 0$, the tolerant measures are equivalent to the standard measures. If $\delta > 0$, the definition of a true positive in the numerator changes. In fact, there are now two different types of true positives: In the case of precision, a true positive is a predicted anomaly at time step t with an actual anomaly within the range $[t - \delta, t + \delta]$. In the case of recall, a true positive is an actual anomaly at time step t with a predicted anomaly within the range $[t - \delta, t + \delta]$.

Figure 2 shows the impact of various choices for the temporal tolerance δ on the measured values for precision and recall. Depending on the choice of the threshold and the temporal tolerance, the reported values for precision and recall vary drastically. The adoption of temporal tolerance in the evaluation is undoubtedly valid in many applications. However, the example shows that the use of moderate temporal tolerances may already lead to overstated performance measures that do not necessarily reflect the actual utility of the results. In Section 3, we perform a simulation study to further investigate this issue.

2.2 Confusion matrices

The extension of precision and recall to time-tolerant measures via relaxed notions of *true positives* is intuitive, but has some subtleties not discussed before. In fact, these two measures are computed from two distinct *confusion matrices*, where temporal tolerance is allowed *either* in the ground-truth time steps *or* in the predicted time steps. The general structure of a confusion matrix is:

	AA	AnA	
PA	TP	FP	Σ
PnA	FN	TN	Σ
	Σ	Σ	T

It contains the numbers of observations that fall into the four categories true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), along with marginal sums. The row and column headings define the marginal conditions: actual anomaly (AA), actually no anomaly (AnA), predicted anomaly (PA), predicted no anomaly (PnA). The confusion matrix partitions the observations so that every observations falls in exactly one category. Many performance measures can be computed from confusion matrices [11], typically by normalizing individual entries by marginal sums. The measures are interpretable because all entries and marginals have straightforward interpretations.

When introducing temporal tolerance in the confusion matrix, we have to make sure that the result is still a partition with interpretable entries and marginals. Tables 1 and 2 show the confusion matrices obtained when introducing temporal tolerance either into the ground-truth time steps or the predicted time steps, using

Table 1: Relaxed confusion matrix for sequential data, with tolerance in ground-truth

	AAδ	AnAδ	
PA	$\sum_t \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right) \mathcal{I}(z_t \geq \tau)$	$\sum_t \left(1 - \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right)\right) \mathcal{I}(z_t \geq \tau)$	$\sum_t \mathcal{I}(z_t \geq \tau)$
PnA	$\sum_t \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right)(1 - \mathcal{I}(z_t \geq \tau))$	$\sum_t \left(1 - \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right)\right)(1 - \mathcal{I}(z_t \geq \tau))$	$\sum_t (1 - \mathcal{I}(z_t \geq \tau))$
	$\sum_t \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right)$	$\sum_t \left(1 - \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} e_{t'} > 0\right)\right)$	T

AAδ: actual anomaly with tolerance δ, AnAδ: actually no anomaly with tolerance δ, PA: predicted anomaly, PnA: predicted no anomaly; we use zero-padding at the boundaries

Table 2: Relaxed confusion matrix for sequential data, with tolerance in predictions

	AA	AnA	
PAδ	$\sum_t e_t \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)$	$\sum_t (1 - e_t) \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)$	$\sum_t \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)$
PnAδ	$\sum_t e_t \left(1 - \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)\right)$	$\sum_t (1 - e_t) \left(1 - \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)\right)$	$\sum_t \left(1 - \mathcal{I}\left(\sum_{t'=t-\delta}^{t+\delta} \mathcal{I}(z_{t'} \geq \tau) > 0\right)\right)$
	$\sum_t e_t$	$\sum_t (1 - e_t)$	T

AA: actual anomaly, AnA: actually no anomaly, PAδ: predicted anomaly with tolerance δ, PnAδ: predicted no anomaly with tolerance δ; we use zero-padding at the boundaries

the formal notation introduced above. Both confusion matrices partition the observations, but not all entries and marginals have straightforward interpretations. Some of the measures usually computed from confusion matrices are therefore uninformative. The tolerant precision from Equation 2 is the TP entry from Table 1 (PA-AAδ) normalized by the marginal row sum (PA), whereas the tolerant recall from Equation 3 is given by the TP entry from Table 2 (PAδ-AA) normalized the marginal column sum (AA). In both cases, the TP entries and normalization terms are interpretable and yield informative evaluation measures. We thus restrict our analysis to these two cases and defer other measures to future work.

2.3 Statistical significance

They key question when analyzing evaluation measures from a statistical point of view is whether the reported values are statistically significant. To assess statistical significance, we have to treat the quantities in the confusion matrix as *random variables* that follow some probability distribution. Only if the reported number of true positives (or any other entry of the confusion matrix) is *much larger (or smaller) than expected* due to random coincidences, the result should be considered statistically significant. Donges et al. [7] have derived the probability distribution of the two types of true positives (PA-AAδ and PAδ-AA) from Tables 1 and 2, under the assumption that the ground-truth anomalies and the predicted anomalies follow independent Bernoulli processes. In this case, both quantities follow simple binomial distributions. Scharwächter and Müller [14] have generalized the formal analysis for a larger class of problems, where the anomaly score is a strictly stationary process. They show that in this case PAδ-AA from Table 2 also follows a

binomial distribution, where the success probability can be approximated using a result from Extreme Value Theory [5]. Unfortunately, there is no analogous derivation for PA-AAδ from Table 1 under the strict stationarity assumption. In this work, we do not use the existing analytical results, but perform Monte Carlo simulations to estimate the required probability distributions without potentially limiting assumptions on the data generating processes.

3 SIMULATION STUDY

We now use the anomaly score $(z_t)_{t=1,\dots,T}$ from the earthquake detection example in Section 1 and compute time-tolerant confusion matrices, as well as the time-tolerant precision and recall measures, for randomized ground-truth sequences of anomalies. We evaluate the anomaly score against 10,000 random permutations of the ground-truth sequence of anomalies $(e_t)_{t=1,\dots,T}$ from the example. In doing so, we keep the number of ground-truth anomalies constant and assume that they follow a Bernoulli process. We believe that this assumption is reasonable for ground-truth anomalies, which typically occur rarely and are not clustered.

3.1 Monte Carlo precision and recall

First, we visualize the precision and recall values obtained from a subset of 100 random permutations for various temporal tolerances and thresholds in Figure 3. The visualization also shows the performance measures observed on the non-permuted ground-truth sequence of anomalies.

The observed precision and recall values on the non-permuted sequence are generally higher than the values from the randomly permuted sequences, especially at larger thresholds. This confirms
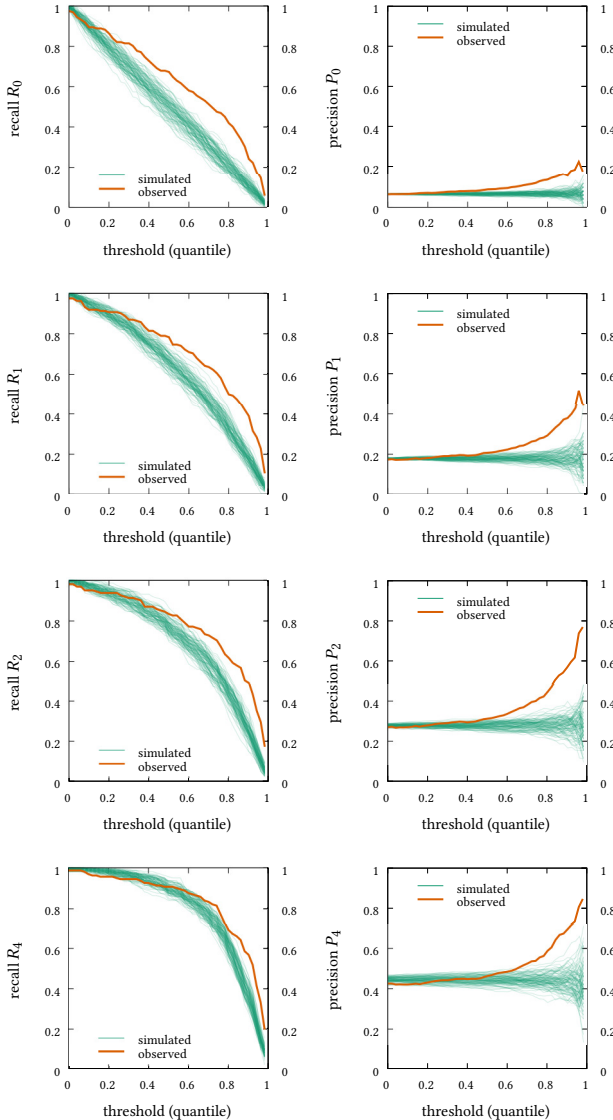
Figure 3: Simulated and observed values for the precision P_δ and recall R_δ, for $\delta \in \{0, 1, 2, 4\}$ and various thresholds.

that the anomaly score contains useful information on earthquake occurrences. However, when the temporal tolerance is increased, the gap between the simulated and the observed performance measures tends to shrink: the performance measures on the simulated sequences increase to a stronger degree than the performance measures on the observed sequence. The consequence is that reported performance measures, in particular when computed with a high temporal tolerance, may not reflect the actual performance of the anomaly detector. In the worst case, one might conclude that the anomaly score allows detection of anomalies that are *statistically independent* of the anomaly score.
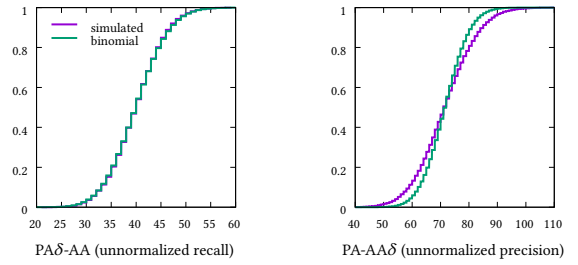

Figure 4: Cumulative distribution functions of the two types of true positives required for precision and recall.

3.2 Null distributions

The simulations clearly show that assessment of the statistical significance of the observed performance measures is imperative. For this purpose, we fix the temporal tolerance to $\delta = 2$ and set the threshold τ to the .9-quantile of the anomaly score. We observe PAδ-AA = 80 (recall R_δ = .49) and PA-AAδ = 145 (precision P_δ = .56) on the non-permuted ground-truth anomaly sequence from our example. To assess the statistical significance of the reported numbers, we now have a closer look at the null distributions for the performance measures obtained in the Monte Carlo simulations.

Figure 4 (simulated) shows the cumulative distribution functions for the numbers of true positives obtained from 10,000 simulations for the specific choice of δ and τ mentioned above. Given the simulated distributions, we can easily compute Monte Carlo p-values [6] for the numbers of true positives: The p-value is the probability that we obtain a value for the true positive at least as high as the observed one. Since the performance measures were smaller than the reported values in *all* of the simulated runs, we have $p < .0001$ for both precision and recall, which is highly significant.

The analytical null distributions derived in the literature [7, 14] are all binomial. To complete our analysis, we now check whether our simulations also yield binomial distributions. Figure 4 (binomial) shows the cumulative distribution functions of binomial random variables when the binomial success probabilities are estimated from our Monte Carlo simulations. The plots suggest that the true positive PAδ-AA for the recall follows a binomial distribution, whereas the true positive PA-AAδ for the precision seems to be overdispersed with respect to the binomial distribution (it has a larger variance). We have repeated the experiment with different thresholds and temporal tolerances and observed the same behavior across all experiments. The exact form of the overdispersed distribution should be investigated more deeply in future work.

4 CONCLUSION

We have presented time-tolerant variants of the precision and recall measures routinely used to evaluate anomaly detectors for sequential data. We have shown that these measures are computed from two distinct time-tolerant confusion matrices. Time-tolerant confusion matrices can, in principle, be used to derive time-tolerant variants of other well-known measures. However, care has to be taken to preserve interpretability. We applied the time-tolerant

precision and recall measures on an example anomaly detection problem, and analyzed their statistical behaviors in a simulation study. Our experiments suggest that reported values for precision and recall can overestimate the performance of an anomaly detector even with moderate temporal tolerances. We have demonstrated how to obtain Monte Carlo p-values to assess the statistical significance of reported performance measures, using randomly permuted ground-truth sequences. We believe that establishing the statistical significance of reported precision and recall values should become a community standard. Future work should improve the analytical understanding of the null distributions required for this task.

REFERENCES

[1] Roy J. Adams and Benjamin M. Marlin. 2017. Learning time series detection models from temporally imprecise labels. In *AISTATS*.
[2] Roy J. Adams and Benjamin M. Marlin. 2018. Learning time series segmentation models from temporally imprecise labels. In *UAI*.
[3] Robert A Bridges, Jessie D Jamieson, and Joel W Reed. 2017. Setting the threshold for high throughput detectors. In *IEEE Big Data*.
[4] Lianhua Chi, Bo Han, and Yun Wang. 2016. Open Problem: Accurately Measuring Event Impacts on Time Series. In *KDD MiLeTS Workshop*.
[5] Stuart Coles. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London, Ltd, London, UK.
[6] A. C. Davison and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, UK.
[7] J. F. Donges, C.-F. Schleusner, J. F. Siegmund, and R. V. Donner. 2016. Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks. *European Physics Journal Special Topics* 487 (2016), 471–487.
[8] Paul S. Earle, Daniel C. Bowden, and Michelle Guy. 2011. Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of Geophysics* 54, 6 (2011), 708–715.
[9] Chen Luo, Jian-Guang Lou, Qingwei Lin, Qiang Fu, Rui Ding, Dongmei Zhang, and Zhe Wang. 2014. Correlating events with time series for incident diagnosis. In *KDD*.
[10] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. In *ESANN*.
[11] David M. W. Powers. 2007. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Technical Report. Flinders University of South Australia, Adelaide, Australia.
[12] R. Quian Quiroga, T. Kreuz, and P. Grassberger. 2002. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Physical Review E* 66, 4 (2002).
[13] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. 2019. Time-Series Anomaly Detection Service at Microsoft. In *KDD*.
[14] Erik Scharwächter and Emmanuel Müller. 2020. Does Terrorism Trigger Online Hate Speech? On the Association of Events and Time Series. *Annals of Applied Statistics* (2020).
[15] Erik Scharwächter and Emmanuel Müller. 2020. Two-Sample Testing for Event Impacts in Time Series. In *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*.
[16] Jonatan F. Siegmund, Marc Wiedermann, Jonathan F. Donges, and Reik V. Donner. 2016. Impact of temperature and precipitation extremes on the flowering dates of four German wildlife shrub species. *Biogeosciences* (2016).
[17] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet. 2017. Anomaly detection in streams with extreme value theory. In *KDD*.
[18] Ya Su, Rong Liu, Youjian Zhao, Wei Sun, Chenhao Niu, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *KDD*.
[19] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and recall for time series. In *NeurIPS*.
[20] M. A.M.M. van Dortmont, S. van den Elzen, and J. J. van Wijk. 2019. ChronoCorrelator: Enriching events with time series. In *EuroVis*.
[21] Marc Wiedermann, Alexander Radebach, Jonathan F. Donges, Jürgen Kurths, and Reik V. Donner. 2016. A climate network-based index to discriminate different types of El Niño and La Niña. *Geophysical Research Letters* 43, 13 (2016), 7176–7185.
[22] Mitchell Withers, Richard Aster, Christopher Young, Judy Beiriger, Mark Harris, Susan Moore, and Julian Trujillo. 1998. A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America* 88, 1 (1998), 95–106.
[23] Haowen Xu, Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, and Dan Pei. 2018. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. In *WWW*.