Online FDR Controlled Anomaly Detection for Streaming Time Series

Weinan Wang
University of Southern California
Los Angeles, California
weinanw@usc.edu

Zhengyi Liu
Snap Inc.
Venice, California
zliu@snapchat.com

Xiaolin Shi
Snap Inc.
Venice, California
xiaolin@snap.com

Lucas Pierce
Snap Inc.
Venice, California
lpierce@snap.com

ABSTRACT

As a large social network platform with more than 100 million daily active users, we have a large number of user engagement metrics stored in Google Cloud in the form of time series, detecting anomalies in such time series data in a robust fashion can give meaningful insights and enable proper subsequent actions. In this paper, we tackle this problem by transforming it into a multiple testing problem in the statistical domain. We first use STL (seasonal trend residual decomposition using Loess) to decompose the time-series data, then we propose a novel empirical Bayes procedure for online False Discovery Rate (FDR) control at any nominal level on the residual terms. Our main contribution is the novel online FDR control procedure that's robust and fits nicely with our streaming anomaly detection goal. Furthermore, our online FDR control procedure is a powerful statistical tool for many other anomaly detection algorithms since it can be directly applied on score functions or error terms to determine proper threshold, which are oftentimes empirically determined based on training data in the literature. R code for reproducing the results in the paper is provided in *links hidden for double blind review*.

KEYWORDS

Anomaly detection, time series decomposition, multiple testing, online false discovery rate, empirical Bayes, non-parametric.

ACM Reference Format:
Weinan Wang, Zhengyi Liu, Xiaolin Shi, and Lucas Pierce. 2018. Online FDR Controlled Anomaly Detection for Streaming Time Series. In *MileTS '19: 5th KDD Workshop on Mining and Learning from Time Series, August 5th, 2019, Anchorage, Alaska, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

MileTS '19, August 5th, 2019, Anchorage, Alaska, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In many modern Internet companies, there is an exponential increase in the availability of streaming, time-series data. Among these data-sets, the most important ones are the user engagement metrics (active users, session length, session frequency, etc.), which are key indicators for company performances, external events, and potential infrastructure outages. Therefore, fast and reliable detection of anomalies in such streaming time series data has significant implications and use cases.

There are numerous research in time-series anomaly detection, dating back to [16]. A lot of them has been done in various domains such as, statistics, signal processing, finance, econometric, etc [5, 21, 23, 40]. Many techniques are supervised methods, which are unsuitable for robust anomaly detection since they hardly detect new and unknown anomalies [19]. Other techniques, like smoothing methods, clustering, simple thresholding methods, are only capable of detecting spatial anomalies [3]. For detecting temporal anomalies, change point detection methods using scan statistics and likelihood ratio tests have been employed in the field of genomics, engineering, and material science [7, 20, 31, 35]. Such methods are often sensitive to the size of the windows and pre-specified thresholds [3], making it difficult to control the number of false positives while maintaining good detection power.

At our company, most of the streaming time series data demonstrate strong seasonality with an underlying trend. In [22], they propose a hybrid approach using STL (seasonal trend decomposition using Loess) and the Extreme Studentized Deviate test (ESD), which is robust to high percentage of anomalies and can elect arbitrary number of anomalies for rejection. However, their Seasonal ESD and Seasonal Hybrid ESD methods do not provide any control on the false discovery rate [1] or offer any power analysis, which is of utmost importance considering the vast amount of data being processed and potential augmented number of false positives due to the statistical artifact known as "multiple testing" [2].

Considering the aforementioned issues, we develop a novel online FDR procedure for real time anomaly detection for such user engagement metrics at our company. The proposed framework is fully data-driven, and can control FDR in an online fashion with superior power.

[1] https://en.wikipedia.org/wiki/False_discovery_rate
[2] https://en.wikipedia.org/wiki/Multiple_comparisons_problem

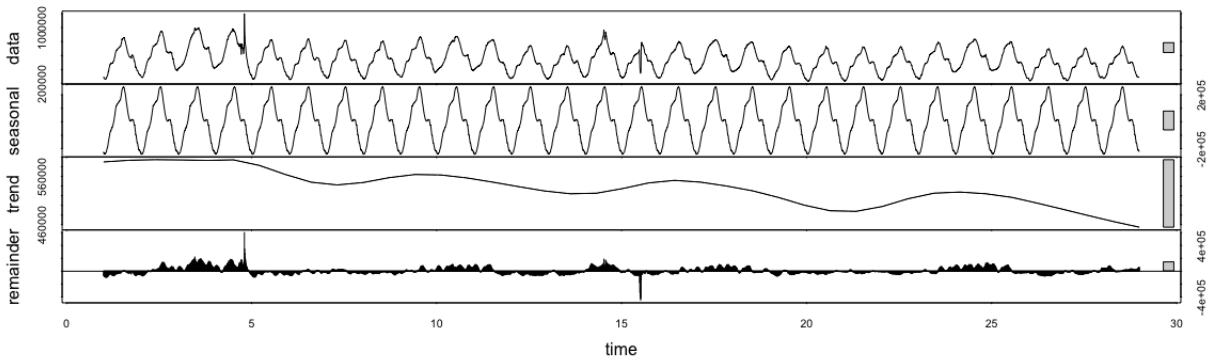

Figure 1: An example time series data after STL decomposition. Y-axis is rescaled in order not to show the absolute numbers.

Specifically, we first use STL to decompose the time series data into three components: seasonal, trend, and residuals. Then we model the residuals as a two group mixture model with noise and signals, a novel data-driven procedure is subsequently applied for online adaptive FDR control. We summarise our main contributions as follows:

- Formulate anomaly detection on time series as a multiple testing problem.
- Propose a novel data-driven online FDR control procedure with empirically investigated strong performances.
- Propose novel and robust non-parametric estimation methods for the test statistics.
- Provides a practical paradigm for mitigating false discoveries of many anomaly detection algorithms that involve binary decision making on score functions.

2 THEORETICAL FRAMEWORK

In this section, we first introduce the STL decomposition for extracting the residual component, and subsequently formulate the anomaly detection problem on time series as a multiple testing problem. Then we propose a novel online FDR control procedure. Subsequently, we discuss data-driven estimation methods.

2.1 STL decomposition

STL is a non-parametric technique coined by [13] to deal with time series data of such type. It decomposes a time series into three additive components- *seasonal, trend and remainder*:

$$Y_t = T_t + S_t + R_t,\ t = 1, \cdots, N. \tag{1}$$

N here is the total number of measured data points. The main motivation for using STL is that we want to leave the potential anomaly points in the residual for further analysis, while providing useful insights on the trend and seasonality terms. For the algorithmic details, readers can refer to the outline delineated in [17].

The main quantity of interest is the residual term, which we would now consider a mixture of random noise and anomalies which cannot be accounted for in the seasonal or trend components.

There are many techniques in the literature for time series decomposition, we chose STL because of its versatility and robustness, especially towards outliers. Figure 1 is an example of time series data after STL decomposition.

2.2 Multiple testing formulation and online FDR control

After decomposing the original time series data, we want to further analyze the residual terms for anomaly points. The quantity of interest is the residual term R_t from the STL decomposition. With anomalies present, the residual terms R_t can be thought of as a mixture of two groups, the noise and potential signals (i.e. anomalies). Specifically, let $\theta_1, \cdots, \theta_N$ be independent Bernoulli(p_t) variables and let R_t be generated as

$$R_t|\theta_t \sim (1-\theta_t)F_0 + \theta_t F_{1t},\ t = 1, \cdots, N. \tag{2}$$

R_t are observed, while the variables θ_t are unobserved. Here F_0 is the cumulative distribution function (cdf.) for the noise, and F_{1t} the cdf. for the signals at time t. Then the marginal cdf. of R_t is the mixture distribution $F_t(r) = (1-p_t)F_0(r) + p_t F_{1t}(r)$, and the probability distribution function (pdf.) is $f_t(r) = (1-p_t)f_0(r) + p_t f_{1t}(r)$. Here p_t can also be thought of as the varying signal proportions. Note here in our formulation, we allow p_t and f_{1t} to vary with time, allowing more general cases.

Under such framework, the anomaly detection problem can be further cast as a multiple testing problem, when the end goal is to determine which time points are from the alternative distribution, i.e. anomalies:

$$H_{0t} : \theta_t = 0,\ H_{1t} : \theta_t \neq 0,\ t = 1, \cdots, N, \tag{3}$$

where the solution to which can be represented by a decision rule,

$$\boldsymbol{\delta} = (\delta_1, \cdots, \delta_N) \in \mathcal{I} = \{0,1\}^N. \tag{4}$$

For each time point t, define $\delta^t = (\delta_1, \delta_2, \cdots, \delta_t)$ be the collection of all decisions up until time t, we aim to control false discovery rate at time t as

$$\mathrm{FDR}_t(\delta^t) = \mathbb{E}\left(\frac{\sum_{i=1}^t (1-\theta_i)\delta_i}{\sum_{i=1}^t \delta_i \vee 1}\right) \leq \alpha.$$

By novelly formulating anomaly detection as an online FDR controlled multiple testing problem, we simultaneously deal with an array of issues where traditional methods fail to address.

Firstly, simple thresholding is the common procedure for many anomaly detection algorithms based on some sort of statistics or score functions. The thresholds are often determined based on development set for supervised methods or ad-hoc values for unsupervised methods, which do not provide any theoretical guarantee on the controlled FDR level. Our formulation can be applied on a series of anomaly detection algorithms where FDR control is appropriate, and thresholds would be chosen in a data-driven fashion based on the desired significance level α.

Secondly, online FDR control ensures the safety and stability of our decisions at any time. Most existing methods can only address the globally controlled number of false discoveries.

2.3 Streaming online FDR Control procedure

In large-scale statistical inference, FDR control has become the standard practice [4, 8, 11, 34]. They can be further categorized into two groups of methodologies, p-value based and empirical Bayes[3] type methods. In general, p-value based techniques are inferior compared with empirical Bayes type methods because it fails to take into consideration the compound structure of the problem [11, 34, 39], whereas the multiplicity can be used to improve power over traditional tests, especially given the high dimensionality of most anomaly detection problems in real time streaming data.

The seminal work by [34] propose the adaptive z-procedure, an empirical Bayes type method that demonstrates stability and robustness under the offline setting. It involves calculating the Lfdr (local false discovery rate) statistics and choose a cutoff along the ranked Lfdrs to control FDR.

However, the adaptive z-procedure assumes a fixed proportion of anomalies p and alternative distribution f_1, and it involves ranking all the observations which is not feasible in a streaming setting at our company. Therefore, controlling FDR_t at any time t is the natural solution for real time anomaly detection.

This online streaming setting imposes two new constraints. Firstly, we cannot revise decisions that's already made on past data points. Secondly, it prohibits ranking test statistics which is how typical multiple testing procedures exploit the global structure. In order to address these issues, we propose the following novel online FDR control procedure.

In order to accommodate varying anomaly proportion p_t and f_t, consider the following conditional Lfdr statistic:

$$\text{CLfdr}_t = \mathbb{P}(\theta_t = 0 | R_t = r_t) = \frac{(1-p_t) f_0(r_t)}{f_t(r_t)},\ t = 1, \cdots, N.$$

Let $\mathcal{A}_t = \{i : i \le t, \delta_i = 1\}$ be the collection of locations we rejected up until time t, α be the nominal FDR level, then the following oracle procedure denoted as $\boldsymbol{d}^{\text{OR}}_{\text{on}}$ would guarantee control of the FDR_t for all t:

(1) **Initialization**: $\mathcal{A}_0 = \emptyset$.
(2) **Decision**: $\delta_t = \mathbb{I}\left(\frac{\sum_{i \in \mathcal{A}_{t-1}} \text{CLfdr}_i + \text{CLfdr}_t}{|\mathcal{A}_{t-1}|+1} \le \alpha\right)$.

The proof of oracle procedure's validity is in the appendix.

[3] https://en.wikipedia.org/wiki/Empirical_Bayes_method

3 IMPLEMENTATION AND ESTIMATION METHODS

In order to practically implement the STL and data-driven online FDR procedure for streaming time series, we need to specify the parameters for STL and the estimation methods for key quantities in the CLfdr test statistics, together with how to enable real time decision making. We devote this section to practical implementation setup and bring forth novel estimation methods.

3.1 Parameters for STL decomposition

For STL decomposition, There are six primary parameters involved:

- $\boldsymbol{n_{(p)}}$: the periodicity of the seasonality, e.g., if we were to model daily data with 10 minutes intervals, $n_{(p)} = 60/10 \times 24 \times 7$.
- $\boldsymbol{n_{(i)}}$, $\boldsymbol{n_{(o)}}$: number of cycles through the inner and outer loop.
- $\boldsymbol{n_{(l)}}$: the span in lags for the LPF. It's recommended to take the next odd integer greater than $n_{(p)}$.
- $\boldsymbol{n_{(s)}}$, $\boldsymbol{n_{(t)}}$: smoothing parameter for the seasonal filter and the trend behavior.

In practice, the most important two parameters are $n_{(p)}$ and $n_{(s)}$, we recommend choosing $n_{(p)}$ based on total number of data points in a week while let $n_{(s)}$ be relatively large, say 35, therefore believing the changes are resulted from aberrant behaviors (in the residuals) instead of seasonal behaviors.

Using simulated data, we further noticed that by letting the number of inner and outer iterations both equal to 2 yields more stable results than default values.

For real time prediction, we can use the method **predict** in R on our fitted STL model, then use the difference of the observed value and the fitted value as the residuals.

3.2 Estimations in the online FDR procedure

The oracle statistics in our online FDR procedure is the conditional local false discovery rate CLfdr:

$$\text{CLfdr}_t = \frac{(1-p_t) f_0(z_t)}{f_t(z_t)},\ t = 1, \cdots, N. \tag{5}$$

p_t here can be interpreted as the proportion of anomalies among observations until time t, which intrinsically should be small, while $f_0(z_t)$ is the null distribution of the noises. After standardizing the observations R_t into z-scores z_t, we can either use the theoretical null distribution $N(0,1)$ as f_0, or we can use the method proposed in [27] to obtain consistent estimators for $\hat{f}_0 = N(\hat{\mu}_0, \hat{\sigma}_0^2)$ (*the empirical null distribution*) using empirical characteristic function and Fourier analysis. We suggest the usage of the empirical null distribution for online anomaly detection as we can update the parameters periodically.

In this section, we propose a novel estimator for the test statistics $\widehat{\text{CLfdr}}_t$.

Proposition 3.1 ($\widehat{\text{CLfdr}}_t$ estimation). *At time t,*

- *calculate the p-values for R_i, $i = 1, \cdots, t$ as P_{1i}, obtain sample $\mathcal{T}_t(\tau) = \{i : P_{1i} > \tau\}$, a rule of thumb for τ here is 0.8.*

- *use standard bivariate density estimator to estimate $f(z_t, t)$, record bandwidth for z_t terms as h_1, bandwidth for t terms as h_2, bandwidths are chosen based on [33].*
- *apply kernel smoothing to $\mathcal{T}_t(\tau)$:* $\hat{q}_t^\tau := \frac{\sum_{i\in\mathcal{T}_t(\tau)} K_{h_2}(t-i)}{t(1-\tau)}$,
- *estimate $\widehat{CLfdr}_t$:* $\widehat{CLfdr}_t = \frac{\hat{q}_t^\tau \hat{f}_0(z_t)}{\hat{f}(z_t,t)} \wedge 1$.

By using time t as covariate, we can improve the power of the CLfdr test statistics. We are assuming p_t as a continuous function of time t, which is natural.

One caveat of the above estimation is the initial "burn-in" period when we simply do not have enough data to construct a large enough $\mathcal{T}_t(\tau)$ for kernel smoothing. Alternatively, during such period, one can omit the estimation of $\hat{p}_t$ for simplicity, since

$$\text{CLfdr}_t = \frac{(1-p_t)f_0(z_t)}{f_t(z_t)} < \frac{f_0(z_t)}{f_t(z_t)}, \ t = 1, \cdots, N,$$

providing conservative control of the FDR_t under nominal level α.

Note here the estimations involved in the online FDR procedure are all non-parametric and data-driven, which means minimum user-specifications and theoretical guarantee of consistency.

3.3 Real time anomaly detection algorithm

Now we summarize the implementation details into the following algorithm.

online FDR procedure

(1) Specify $n_{(p)}$ and $n_{(s)}$ for STL decomposition, get the trend T_i, seasonal S_i and residual R_i.
(2) Standardize the residuals R_i into z-scores z_i, then estimate the null distribution $N(\hat{\mu}_0, \hat{\sigma}_0^2)$.
(3) Use method delineated in Proposition 3.1 to estimate $\widehat{\text{CLfdr}}_t$.
(4) Let

$$\delta_t = \mathbb{I}\left(\frac{\sum_{i\in\mathcal{A}_{t-1}} \widehat{\text{CLfdr}}_i + \widehat{\text{CLfdr}}_t}{|\mathcal{A}_{t-1}| + 1} \le \alpha\right), \ t = 1, \cdots,$$

4 EXPERIMENTAL RESULTS

In this section, we apply our method on real user engagement metrics at our company. We demonstrate two cases to showcase the power of our real-time anomaly detection algorithm, nominal FDR level is $\alpha = 0.01$. The X-axis and Y-axis in our plots are hidden intentionally for privacy concerns.

We run TW as well for comparison (significance level 0.01), ● are the anomaly points labeled by our data-driven procedure, while △ are the points labeled by TW.

In the first use case, we are able to detect the spikes in March, 2018 due to a student walk out, while TW failed to.

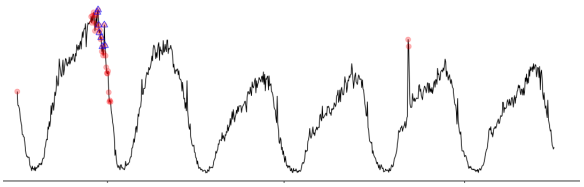

Figure 2: Detected anomalies on user engagement metrics in March, 2018.

In the second use case, our algorithm also picked up an anomaly triggered by a soccer game in Paris in February, 2018.

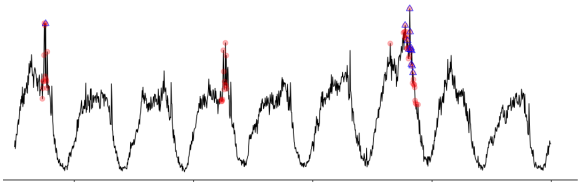

Figure 3: Detected anomalies on user engagement metrics in February and March, 2018.

5 LITERATURE REVIEW

In this section we briefly review relevant work done in anomaly detection, specifically for time series data together with an review of FDR control in the statistics literature and the idea of online FDR control.

5.1 Anomaly Detection in Time Series

Anomaly detection can be categorized into two types, supervised or semi-supervised methods where some labels of anomalies are known, and unsupervised techniques where only the internal structures of the data are used for modeling normal baselines.

While labeled data can be helpful in predicting known type of anomalies, they are typically unsuitable for predicting new types of anomalies [19]. Some of the most notable supervised methods include clustering analysis [6, 25, 30], isolation forests [14, 28], classifiers using artificial neural networks [10, 29]. Most of these techniques often are most effective when there are many additional features.

Other unsupervised technique includes simple thresholds, change point detection, time series analysis, principal component analysis, etc. For instances, Netflix employs the Robust PCA approach to decompose multivariate time series into a sparse and low-rank component [12], whereas the sparse part can be deemed as anomalies. The Skyline project also provides an ensemble technique for streaming data anomaly detection at Etsy. Numenta uses the Hierarchical Temporal Memory (HTM) for anomaly detection which can adapt to changing statistics [2, 3]. [36] focuses on detecting efficiency regression in performance metrics. Most recently, [24] proposed a LSTM based approach on detecting spacecraft anomalies.

5.2 FDR Control

In large-scale multiple comparisons problems, the goal is to effectively separate the non-null cases (in our case, anomalies) from the null cases. The well-known step-up procedure of Benjamini and Hochberg [8] aims to maximize the number of true positives while controlling the proportion of false positives among all rejections, i.e. FDR. Other notable methods include the adaptive procedure that takes the proportion of signals into consideration [9], the plug-in procedure [18], and the augmentation procedure [38].

However, such p-value based methods are typically inefficient under the compound decision theoretic setting [34]. Sun and Cai in [34] further propose an empirical Bayes method named adaptive-z procedure, which ranks the Lfdr (local false discovery rate) statistics and choose the cutoff. Recent advancement in the FDR literature

also includes the knockoff procedure [4], which is robust to many dependent cases.

Empirical Bayes type procedures use the posterior probability as the primary test statistic. Under the assumption that all tests are independent, the adaptive z-procedure in [34] takes into consideration of the global signal proportion and the mixture distribution, providing additional power. Along the line of the adaptive-z procedure, [11] propose the CARS procedure which utilizes an auxiliary statistics constructed from the original data for power gain.

5.3 Online FDR Control

Under the online constraint, multiple testing problems deal with hypotheses that arrive in a stream, whereas decisions must be made immediately after they arrive. Online FDR control deals with the problem on how to control FDR at any given time under the nominal level while maintaining consistency for all decisions made up until the current time.

[15] designed the first online α-investing procedures that can control mFDR (marginal false discovery rate) dynamically. [1] extended the α-investing idea to a more generalized class (GAI) which controls the mFDR as well. [26] propose the LORD and LOND algorithms, which are two special cases of GAI methods. [32] further proposed the GAI++ methods, which uniformly improve the power of GAI methods and can deal with more general cases.

REFERENCES

[1] Ehud Aharoni and Saharon Rosset. 2014. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 4 (2014), 771–794.

[2] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* 262 (2017), 134–147.

[3] Subutai Ahmad and Scott Purdy. 2016. Real-time anomaly detection for streaming analytics. *arXiv preprint arXiv:1607.02480* (2016).

[4] Rina Foygel Barber, Emmanuel J Candès, et al. 2015. Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 43, 5 (2015), 2055–2085.

[5] Vic Barnett and Toby Lewis. 1974. *Outliers in statistical data*. Wiley.

[6] Guilherme A Barreto and Leonardo Aguayo. 2009. Time series clustering for anomaly detection using competitive neural networks. In *International Workshop on Self-Organizing Maps*. Springer, 28–36.

[7] Michèle Basseville, Igor V Nikiforov, et al. 1993. *Detection of abrupt changes: theory and application*. Vol. 104. Prentice Hall Englewood Cliffs.

[8] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[9] Yoav Benjamini and Yosef Hochberg. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics* 25, 1 (2000), 60–83.

[10] Loïc Bontemps, James McDermott, Nhien-An Le-Khac, et al. 2016. Collective anomaly detection based on long short-term memory recurrent neural networks. In *International Conference on Future Data and Security Engineering*. Springer, 141–152.

[11] T Tony Cai, Wenguang Sun, and Weinan Wang. 2018. CARS: Covariate Assisted Ranking and Screening for Large-Scale Two-Sample Inference. *Journal of the Royal Statistical Society, Series B, for discussion, to appear* (2018).

[12] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58, 3 (2011), 11.

[13] Robert B Cleveland, William S Cleveland, and Irma Terpenning. 1990. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 6, 1 (1990), 3.

[14] Zhiguo Ding and Minrui Fei. 2013. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes* 46, 20 (2013), 12–17.

[15] Dean P Foster and Robert A Stine. 2008. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 2 (2008), 429–444.

[16] Anthony J Fox. 1972. Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)* (1972), 350–363.

[17] Dillon R. Gardner. [n. d.]. STL Algorithm Explained: STL Part II. http://www.gardner.fyi/blog/STL-Part-II/. Accessed: 2018-06-05.

[18] Christopher Genovese, Larry Wasserman, et al. 2004. A stochastic process approach to false discovery control. *The Annals of Statistics* 32, 3 (2004), 1035–1061.

[19] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. 2013. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research* 46 (2013), 235–262.

[20] Ning Hao, Yue Selena Niu, and Heping Zhang. 2013. Multiple change-point detection via a screening and ranking algorithm. *Statistica Sinica* 23, 4 (2013), 1553.

[21] Douglas M Hawkins. 1980. *Identification of outliers*. Vol. 11. Springer.

[22] Jordan Hochenbaum, Owen S Vallis, and Arun Kejariwal. 2017. Automatic Anomaly Detection in the Cloud Via Statistical Learning. *arXiv preprint arXiv:1704.07706* (2017).

[23] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (2004), 85–126.

[24] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. *arXiv preprint arXiv:1802.04431* (2018).

[25] Hesam Izakian and Witold Pedrycz. 2013. Anomaly detection in time series data using a fuzzy c-means clustering. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint*. IEEE, 1513–1518.

[26] Adel Javanmard, Andrea Montanari, et al. 2018. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics* 46, 2 (2018), 526–554.

[27] Jiashun Jin and T Tony Cai. 2007. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* 102, 478 (2007), 495–506.

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422.

[29] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. 2015. Long short term memory networks for anomaly detection in time series. In *Proceedings*. Presses universitaires de Louvain, 89.

[30] Erick Giovani Sperandio Nascimento, Orivaldo de Lira Tavares, and Alberto Ferreira De Souza. 2015. A Cluster-based Algorithm for Anomaly Detection in Time Series Using Mahalanobis Distance. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 622.

[31] Yue S. Niu and Heping Zhang. 2012. The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* 6, 3 (09 2012), 1306–1326. https://doi.org/10.1214/12-AOAS539

[32] Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. 2017. Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*. 5650–5659.

[33] Bernard W Silverman. 2018. *Density estimation for statistics and data analysis*. Routledge.

[34] Wenguang Sun and T Tony Cai. 2007. Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* 102, 479 (2007), 901–912.

[35] Maciej Szmit and Anna Szmit. 2012. Usage of modified Holt-Winters method in the anomaly detection of network traffic: Case studies. *Journal of Computer Networks and Communications* 2012 (2012).

[36] Martin Valdez-Vivas, Caner Gocmen, Andrii Korotkov, Ethan Fang, Kapil Goenka, and Sherry Chen. 2018. A Real-time Framework for Detecting Efficiency Regressions in a Globally Distributed Codebase. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 821–829.

[37] Owen Vallis, Jordan Hochenbaum, and Arun Kejariwal. [n. d.]. A Novel Technique for Long-Term Anomaly Detection in the Cloud.

[38] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. [n. d.]. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical applications in genetics and molecular biology* 3, 1 ([n. d.]), 1–25.

[39] Weinan Wang and Wenguang Sun. 2017. Multistage Adaptive Testing of Sparse Signals. *arXiv preprint arXiv:1707.07215* (2017).

[40] Keith Worden, Graeme Manson, and Nick RJ Fieller. 2000. Damage detection using outlier analysis. *Journal of Sound and Vibration* 229, 3 (2000), 647–667.

A PROOFS

Theorem A.1 (Validity). *The oracle online FDR procedure is valid for FDR_t, $t = 1, \cdots, N$ control, i.e.*

$$FDR_t(\boldsymbol{d}_{on}^{OR}) \leq \alpha, \ t = 1, \cdots, N.$$

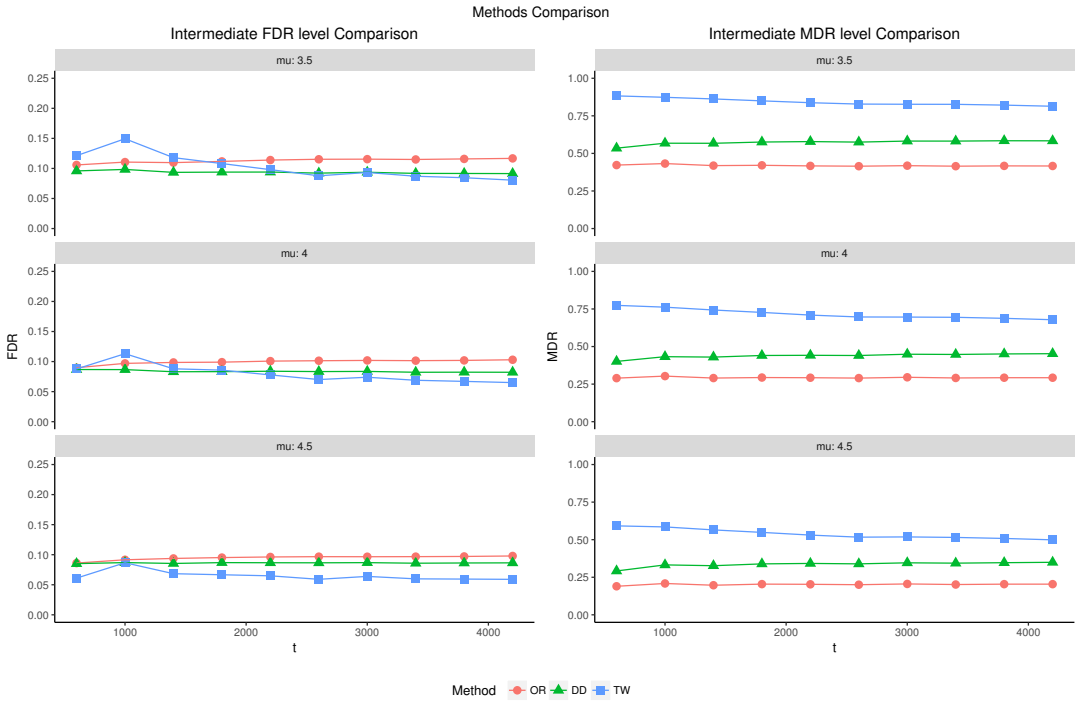

Figure 4: Simulation Setting 1, varying effect size with $p = 0.05$, i.i.d. random noise.

Proof. $\forall t$,

$$\mathrm{FDR}_t = \mathbb{E}\left(\frac{\sum_{i=1}^t (1-\theta_i)\delta_i}{\sum_{i=1}^t \delta_i \vee 1}\right) = \frac{\mathbb{E}\left\{\mathbb{E}\left(\sum_{i=1}^t (1-\theta_i)\,\delta_i | R_i\right)\right\}}{|\mathcal{A}_t|}$$
$$= \frac{\mathbb{E}\left\{\sum_{i\in\mathcal{A}_t} \mathrm{CLfdr}_i\right\}}{|\mathcal{A}_t|} \le \alpha.$$

□

B SIMULATIONS

B.1 Comparison with Twitter's method

In this section, we simulate data with known anomaly data points that has seasonal patterns to demonstrate the performance of our proposed real-time anomaly detection algorithm. We compare the oracle and data-driven version of our algorithm (denoted by **OR** and **DD** respectively), together with anomaly detection algorithm in [37] (denoted by **TW**). Note that TW uses simple thresholding on test statistics, which cannot guarantee FDR control even globally, let alone anytime FDR control. We consider setting with anomalies demonstrating different signal strengths, auto-correlated errors, and different proportion of anomalies.

We use real user engagement metric data at Snap Inc. to extract seasonal and trend data, and manually add in anomalies with varying structures and effect sizes. We add in random noises with distribution $N(0, \sigma_0^2)$ where $\sigma_0 = 144$ based on our sample data.

In general, we consider total number of observations $m = 4458$, we continuously monitor FDR_t at $t = 600, 1000, \cdots, 4200$ with step-size 400. FDR_t and MDR_t (missed discovery rate at time t) are plotted for comparison. In order to make comparisons more meaningful, we choose $\alpha = 0.1$ for universal nominal FDR level for our DD and OR procedures, and choose significance level 0.001 for TW. Note MDR_t is defined as:

$$\mathrm{MDR}_t = \mathbb{E}\left(\frac{\sum_{i=1}^t \theta_i(1-\delta_i)}{\sum_{i=1}^t \theta_i \vee 1}\right),\ t = 1, \cdots,$$

Power is further defined as $1 - \mathrm{MDR}$. Under the same realized FDR level, lower MDR level means more powerful procedure.

Specifically ,we consider the following settings, the corresponding results are summarized in Figure 4 to 6:

- **Setting 1**: Vary signal's effect size (fixed) μ from $3.5\sigma_0$ to $4.5\sigma_0$, proportion of signals p_t linearly vary from 0.01 to $p = 0.05$, error terms are i.i.d. $N(0, \sigma_0^2)$ where $\sigma_0 = 144$, signal locations are uniformly sampled based on binom(p_t). [4]
- **Setting 2**: Signal's effect size are uniformly sampled from $\pm 3.5\sigma_0$ to $\pm 5\sigma_0$, linearly vary proportion of signals p_t from 0.01 to $p = 0.02, 0.03, 0.05$, error terms are i.i.d. $N(0, \sigma_0^2)$ where $\sigma_0 = 144$, signal locations are uniformly sampled based on binom(p_t).
- **Setting 3**: Signal's effect size are uniformly sampled from $\pm 3.5\sigma_0$ to $\pm 5\sigma_0$, linearly vary proportion of signals p_t from 0.01 to $p = 0.02, 0.03, 0.05$, error terms are generated from ARIMA model of order $(2, 0, 1)$ with sd$\approx$ 144 (the ARIMA model is based on estimation from real data at Snap Inc.), signal locations are uniformly sampled based on binom(p_t).

Our real-time anomaly detection algorithm can always control the FDR level below the nominal level $\alpha = 0.1$ at any time under

[4] the size of σ_0 is based on empirical null distribution estimation from the real data at Snap Inc.

all three settings, and it adheres closely to the oracle case where all parameters in the model are assumed to be known. The anomaly detection algorithm in [37] performs reasonably well too, however, there's no direct relationship between the significance level (0.001 here) specified with the realized FDR level. Furthermore, TW fails to control FDR in an online fashion in setting 1 and seems to be overly restrictive in other settings. Our data-driven procedure uniformly achieves better power (lower MDR level) under various anomaly effect sizes and proportions, even when errors are correlated. Furthermore, our procedure is fully data-driven and can achieve more precise error control as more data becomes available over time.

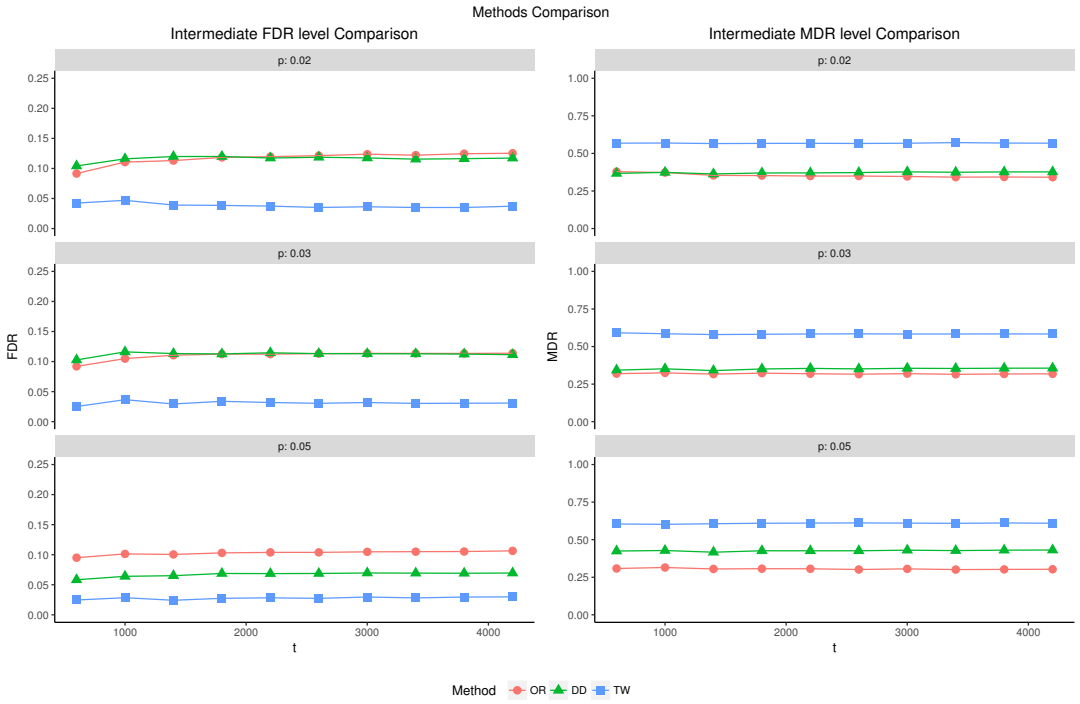

Figure 5: Simulation Setting 2, varying proportion of ignals with uniformly distributed signal strengths, i.i.d. random oise.

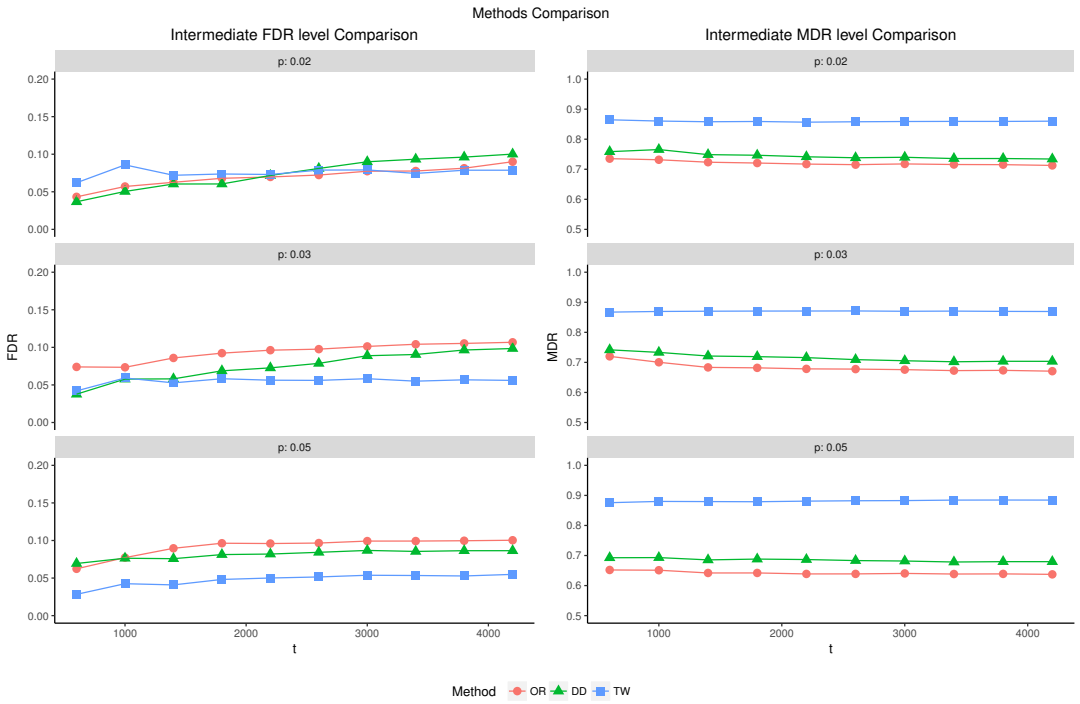

Figure 6: Simulation Setting 3, varying proportion of signals with uniformly distributed signal strengths, noise generated from ARIMA$(2, 0, 1)$.