# Appendices

## 1 Preliminaries

Statistical divergence, also called statistical distance, measures the similarity between two or more distributions. Mathematically, statistical divergence is a function which describes the "distance" of one probability distribution to the other on a statistical manifold. Let $\mathbb{S}$ be a space of probability distributions, then a divergence is a function from $\mathbb{S}$ to non-negative real numbers:

$$D(\cdot||\cdot) : \mathbb{S} \times \mathbb{S} \to \mathbb{R}^+ \cup \{0\} \tag{1}$$

Divergence between two distributions $P$ and $Q$, written as $D(P||Q)$, satisfies:

1. $D(P||Q) \geq 0, \forall P, Q \in \mathbb{S}$
2. $D(P||Q) = 0$, if and only if $P = Q$

For our purposes, we do not require the function $D$ to have the property: $D(P||Q) = D(Q||P)$. But we do need it to be true that if $Q$ is more similar with $P$ than $U$, then $D(Q||P) < D(U||P)$. There are ways to calculate divergence, several frequently used divergence metrics are as follows:

### 1.1 Kullback-Leibler Divergence

Let $P, Q$ be discrete probability distributions, $Q(x) = 0$ implies $P(x) = 0$ for $\forall x$, the *Kullback-Leibler Divergence* from $Q$ to $P$ is defined to be:

$$KLD(P||Q) = \sum_{Q(x)\neq 0} P(x) log\Big(\frac{P(x)}{Q(x)}\Big) \tag{2}$$

For $P, Q$ being continuous distributions:

$$KLD(P||Q) = \int_{q(x)\neq 0} p(x) log\frac{p(x)}{q(x)} dx \tag{3}$$

### 1.2 Jensen-Shannon Divergence

Let $P, Q$ be discrete probability distributions, *Jensen-Shannon Divergence* between $P$ and $Q$ is defined to be:

$$JSD(P||Q) = \frac{1}{2} KLD(P||M) + \frac{1}{2} KLD(Q||M) \tag{4}$$

where $M = \frac{1}{2}(P + Q)$.

A more generalized form is defined to be:

$$JSD(P_1, \ldots, P_n) = H\Big(\sum_{i=1}^{n} \pi_i P_i\Big) - \sum_{i=1}^{n} \pi_i H(P_i) \tag{5}$$

where $H$ is Shannon Entropy, $M = \sum_{i=1}^{n} \pi_i P_i$ and $\sum_{i=1}^{n} \pi_i = 1$.

Especially, if $\pi_i = \frac{1}{n}$, then:

$$JSD(P_1, \ldots, P_n) = \frac{1}{n} \sum_{i=1}^{n} KLD(P_i||M) \tag{6}$$

Jensen-Shannon divergence has some fine properties:

1. $JSD(P||Q) = JSD(Q||P), \forall P, Q \in \mathbb{S}$.
2. $0 \leq JSD(P_1, \ldots, P_n) \leq log_k(n)$. If a $k$ based algorithm is adopted.
3. To calculate $JSD(P||Q)$, it need not necessarily to be true that $Q(x) = 0$ implies $P(x) = 0$.

### 1.3 Bhattacharyya Distance

Let $P, Q$ be discrete probability distributions over same domain $X$, *Bhattacharyya Distance* between $P$ and $Q$ is defined to be:

$$BD(P||Q) = -ln\Big(\sum_{x \in X} \sqrt{P(x)Q(x)}\Big) \tag{7}$$

### 1.4 Hellinger Distance

Let $P, Q$ be discrete probability distributions, *Hellinger Distance* between $P$ and $Q$ is defined to be:

$$HD(P||Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_x \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2} \tag{8}$$

### 1.5 Kolmogorov-Smirnov Statistic

Let $P, Q$ be discrete one-dimensional probability distributions, $CDF_P$ and $CDF_Q$ are their cumulative probability functions respectively, *Kolmogorov-Smirnov Statistic* between $P$ and $Q$ is defined to be:

$$KSS(P||Q) = \sup_x |CDF_P(x) - CDF_Q(x)| \tag{9}$$

## 2 Adaptive Threshold Calculation

$$\begin{aligned} T &= \arg\min_T \int_0^T PDF_a(x)dx + \int_T^{\sup(D)} PDF_n(x)dx \\ &\approx \arg\min_T \int_{-\infty}^T \frac{e^{-\frac{(x-\mu_a)^2}{2\sigma_a^2}}}{\sqrt{2\pi}\sigma_a}dx + \int_T^{+\infty} \frac{e^{-\frac{(x-\mu_n)^2}{2\sigma_n^2}}}{\sqrt{2\pi}\sigma_n}dx \\ &= \begin{cases} \frac{1}{\sigma_a^2 - \sigma_n^2}\left[(\sigma_a^2\mu_n - \sigma_n^2\mu_a) \pm \sigma_a\sigma_n\sqrt{(\mu_a - \mu_n)^2 + 2(\sigma_a^2 - \sigma_n^2)ln\frac{\sigma_a}{\sigma_n}}\right], & \sigma_a \neq \sigma_n \\ \frac{\mu_n + \mu_a}{2}, & \sigma_a = \sigma_n \end{cases} \end{aligned} \tag{10}$$

Note: when $\sigma_a \neq \sigma_n$, keep the root s.t. $\frac{T-\mu_a}{\sigma_a^3}e^{-\frac{(T-\mu_a)^2}{2\sigma_a^2}} < \frac{T-\mu_n}{\sigma_n^3}e^{-\frac{(T-\mu_n)^2}{2\sigma_n^2}}$

$$\begin{aligned} T &= \arg\min_T \alpha\int_0^T PDF_a(x)dx + (1-\alpha)\int_T^{\sup(D)} PDF_n(x)dx \\ &\approx \arg\min_T \alpha\int_{-\infty}^T \frac{e^{-\frac{(x-\mu_a)^2}{2\sigma_a^2}}}{\sqrt{2\pi}\sigma_a}dx + (1-\alpha)\int_T^{+\infty} \frac{e^{-\frac{(x-\mu_n)^2}{2\sigma_n^2}}}{\sqrt{2\pi}\sigma_n}dx \\ &= \begin{cases} \frac{1}{\sigma_a^2 - \sigma_n^2}\left[(\sigma_a^2\mu_n - \sigma_n^2\mu_a) \pm \sigma_a\sigma_n\sqrt{(\mu_a - \mu_n)^2 + 2(\sigma_a^2 - \sigma_n^2)ln\frac{(1-\alpha)\sigma_a}{\alpha\sigma_n}}\right], & \sigma_a \neq \sigma_n \\ \frac{\mu_n + \mu_a}{2} + \frac{k^2 ln\frac{1-\alpha}{\alpha}}{\mu_a - \mu_n}, & \sigma_a = \sigma_n = k \end{cases} \end{aligned} \tag{11}$$

Note: when $\sigma_a \neq \sigma_n$, keep the root s.t. $\frac{\alpha(T-\mu_a)}{\sigma_a^3}e^{-\frac{(T-\mu_a)^2}{2\sigma_a^2}} < \frac{(1-\alpha)(T-\mu_n)}{\sigma_n^3}e^{-\frac{(T-\mu_n)^2}{2\sigma_n^2}}$

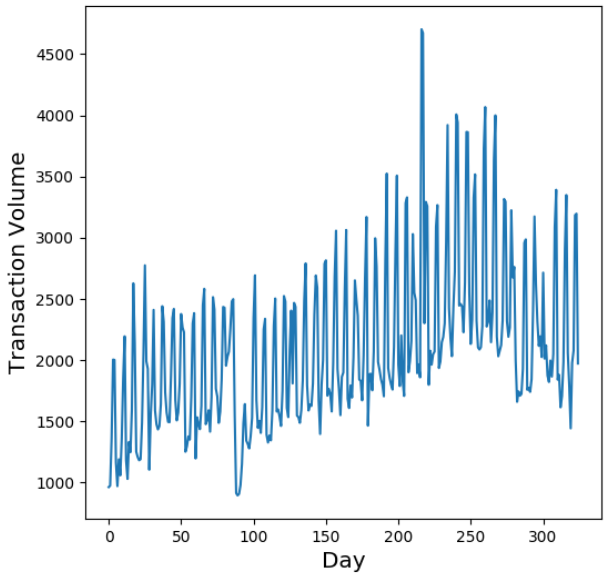


Figure 1: Changing of the daily sales volume shows that environment of online sales had been changing all the time.

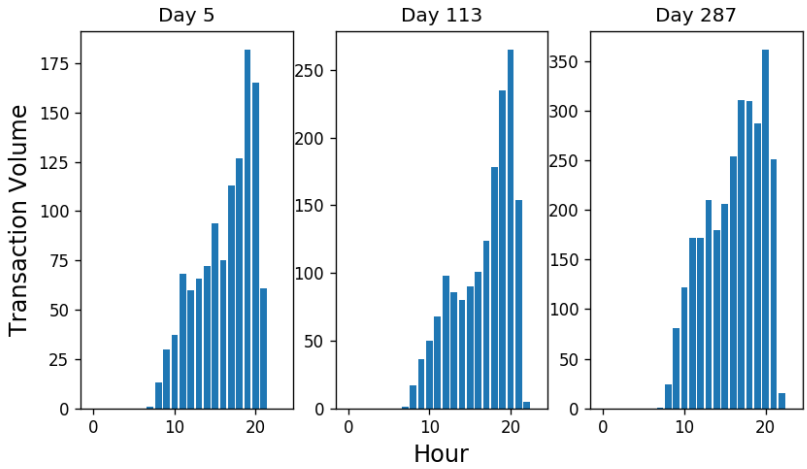


Figure 2: We selected 3 days randomly and drew sales distribution by counting hourly volume. Although sales volume has changed from day to day, the shape of the distribution remain almost alike.

# 3 Extra Insights on Koubei Data Set

# 4 Results on Synthetic Data Set

In this experiment, we tested all seven algorithms on totally synthetic data sets. Results are shown in TABLE 1. It shows that our technique can be applied towards any kind of distributions. And these techniques worked better under irregular distributions since difference were clearer among these. Comparison between SDD-R and static SDD-E shows that adaptive thresholds provided more flexible classifiers. Results under random-shape drifting proves the efficiency of sliding windows toward drifting context.

The last experiment was carried out on random-shaped distribution data set, with $alpha = 0.1$ and rest parameters the same. Under different divergence metrics mentioned in section 1, F1 scores were calculated among all SDD algorithms and ROC curves were recorded on SDD-R and static SDD-E. Given that MGoF was defined specifically on Kullback-Leibler divergence, it cannot be tested in the same way. Results are shown in Fig. 6 and Fig. 7.

Table 1: Performance Comparison on Synthetic Data Sets

| | Uniform | | | | Gaussian | | | | Random-shape | | | | Random-shape with Drift | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Pre(%)** | **Rec(%)** | **F1(%)** | **T(ms)** | **Pre(%)** | **Rec(%)** | **F1(%)** | **T(ms)** | **Pre(%)** | **Rec(%)** | **F1(%)** | **T(ms)** | **Pre(%)** | **Rec(%)** | **F1(%)** | **T(ms)** |
| **SDD-R** | 10.00 | **100.00** | 18.18 | 349.42 | 24.78 | 69.20 | 36.49 | 401.82 | 10.00 | **100.00** | 18.18 | 351.65 | 9.97 | **99.00** | 18.11 | 353.65 |
| **SDD-R+** | 22.40 | 22.40 | 22.40 | 348.49 | 34.40 | 34.40 | 34.40 | **398.02** | 58.40 | 58.40 | 58.40 | **350.87** | 1.20 | 1.20 | 1.20 | **346.67** |
| **SDD-E Static** | 43.71 | 82.60 | 57.17 | 372.74 | 33.34 | 66.60 | 44.44 | 410.72 | 69.24 | 93.20 | 79.45 | 372.77 | 12.34 | 97.20 | 21.91 | 369.63 |
| **SDD-E Static+** | **75.10** | 60.20 | **66.83** | 371.72 | **45.93** | 45.80 | **45.86** | 412.95 | **89.52** | 85.40 | **87.41** | 368.97 | 12.60 | 94.20 | 22.23 | 370.12 |
| **SDD-E Dynamic** | 18.53 | 81.80 | 30.21 | 6808.58 | 20.28 | **71.40** | 31.58 | 8985.92 | 25.90 | 97.40 | 40.92 | 5881.37 | 13.44 | 97.20 | 23.61 | 5747.19 |
| **SDD-E Dynamic+** | 27.85 | 25.60 | 26.68 | 8019.23 | 26.19 | 56.40 | 35.77 | 9197.94 | 77.60 | 79.40 | 78.49 | 6408.12 | **50.55** | 84.00 | **63.11** | 6176.57 |
| **MGoF** | 10.94 | 4.40 | 6.28 | **347.97** | 10.08 | 51.60 | 16.87 | 571.93 | 6.40 | 13.60 | 8.70 | 440.26 | 2.75 | 11.60 | 4.45 | 509.58 |

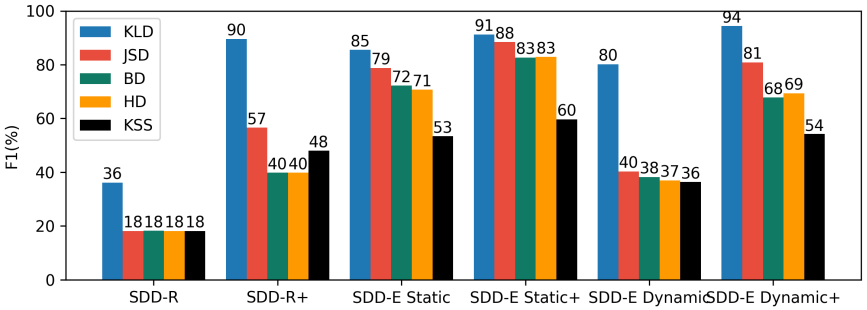


Figure 6: F1 under Different Divergence Metric

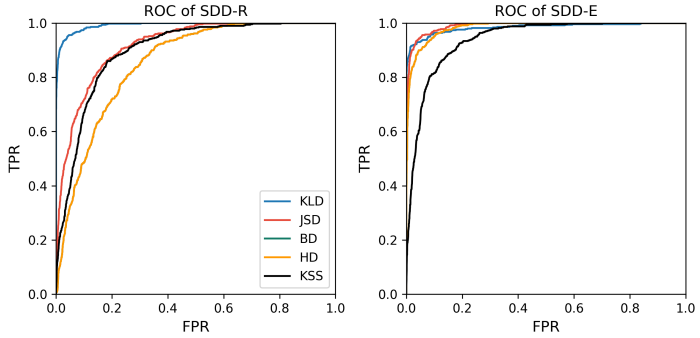


Figure 7: ROC under Different Divergence Metric

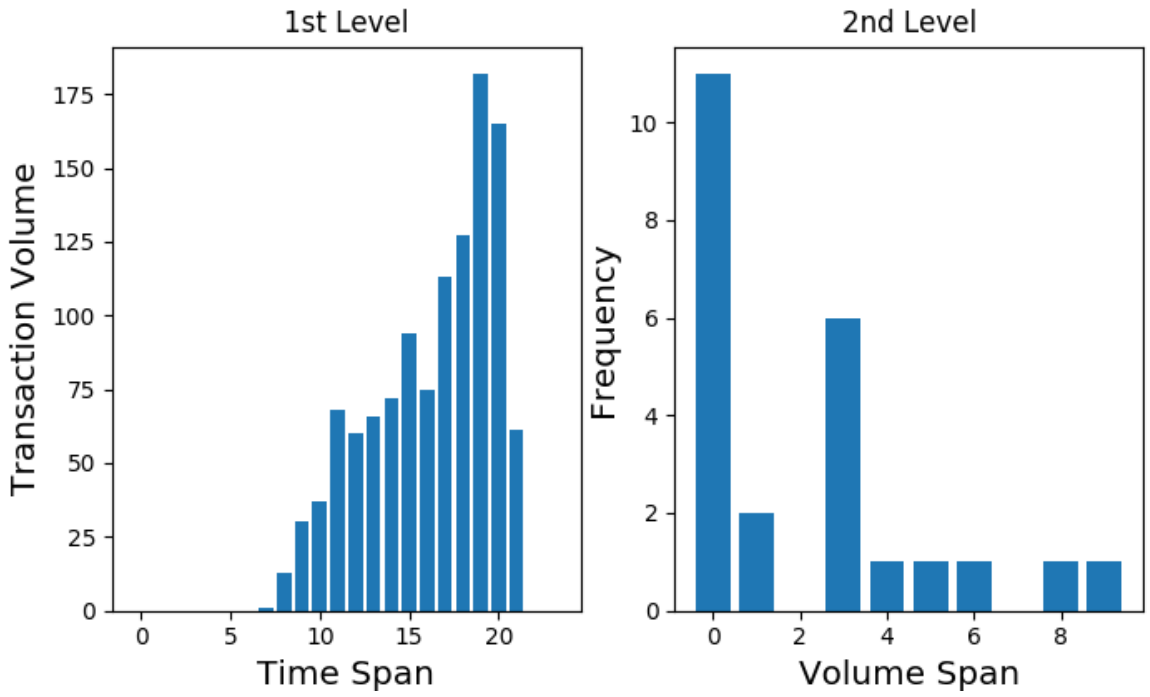


Figure 3: Examples of 1st and 2nd Level Histogram

It is indicated that Jensen-Shannon divergence is suited to all techniques due to its symmetry. Kullback-Leibler divergence provides more evident differences when references were given. Bhattacharyya distance and Hellinger distance turned out almost as good as Jensen-Shannon divergence, but they consumed less time. Kolmogorov-Smirnov Statistic performed relatively poor since it considers only the largest gap between two distributions, which provides little information.

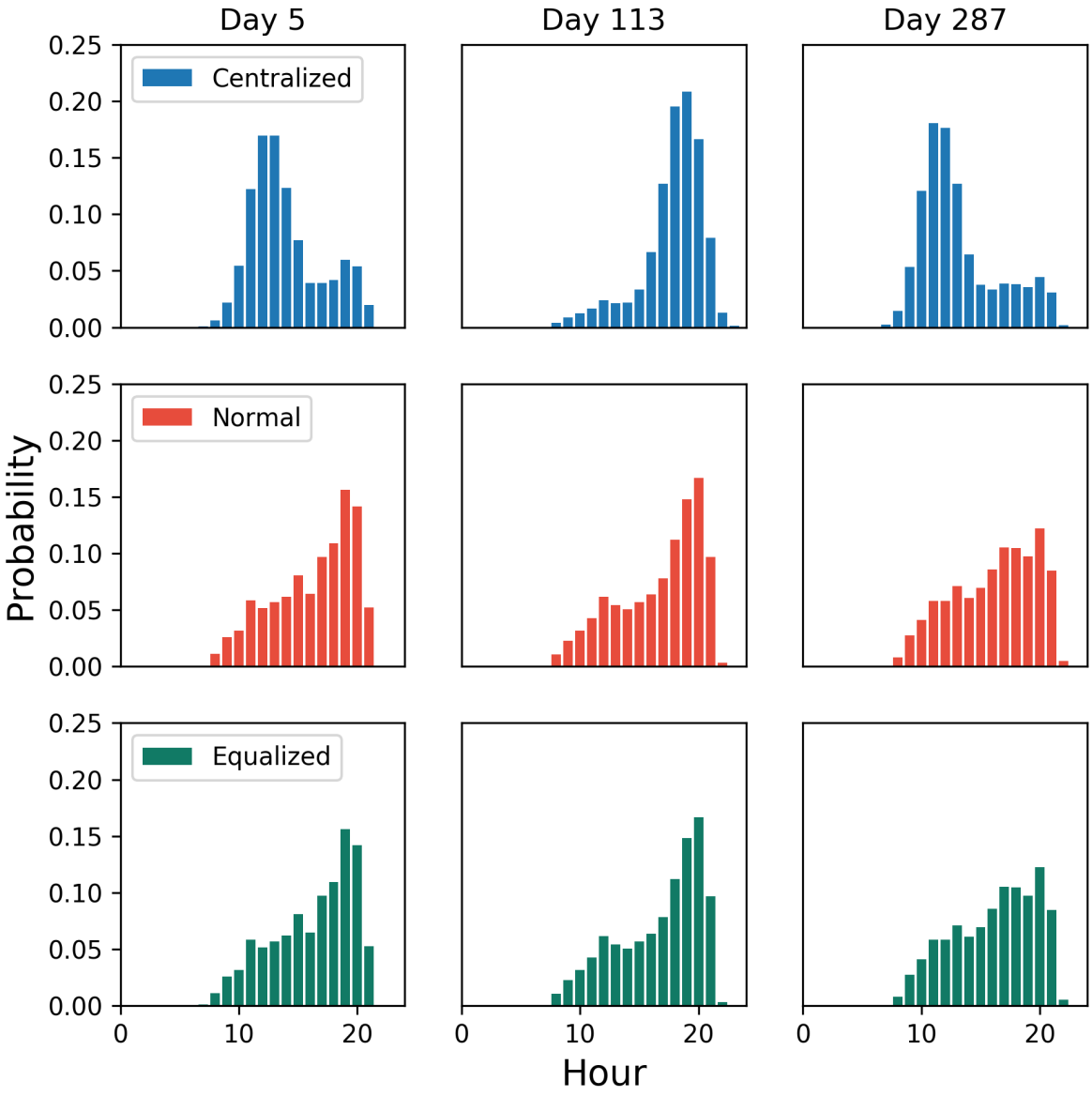


Figure 4: 1st level histogram of day 5, 113 and 287, each day in a column. Distributions after centralized and equalized click farming are in 1st and 3rd rows correspondingly. And the original distributions are shown in the 2nd row.

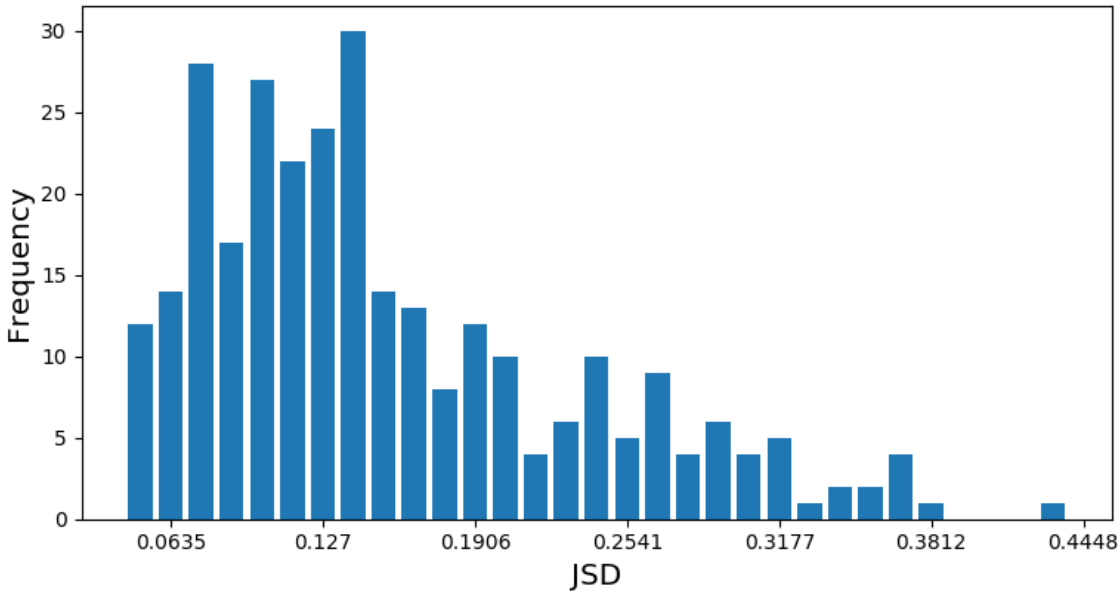


Figure 5: Distribution of Jensen-Shannon divergence on Taobao data set(without click farming) used in the experiments.