Using the Delay in a Treatment Effect to Improve Sensitivity and Preserve Directionality of Engagement Metrics in A/B Experiments

Alexey Drutsa
Yandex, Moscow, Russia
adrutsa@yandex.ru

Gleb Gusev
Yandex, Moscow, Russia
gleb57@yandex-team.ru

Pavel Serdyukov
Yandex, Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

State-of-the-art user engagement metrics (such as session-per-user) are widely used by modern Internet companies to evaluate ongoing updates of their web services via A/B testing. These metrics are predictive of companies' long-term goals, but suffer from this property due to slow user learning of an evaluated treatment, which causes a delay in the treatment effect. That, in turn, causes low sensitivity of the metrics and requires to conduct A/B experiments with longer duration or larger set of users from a limited traffic. In this paper, we study how the delay property of user learning can be used to improve sensitivity of several popular metrics of user loyalty and activity. We consider both novel and previously known modifications of these metrics, including different methods of quantifying a trend in a metric's time series and delaying its calculation. These modifications are analyzed with respect to their sensitivity and directionality on a large set of A/B tests run on real users of Yandex. We discover that mostly loyalty metrics gain profit from the considered modifications. We find such modifications that both increase sensitivity of the source metric and are consistent with the sign of its average treatment effect as well.

Keywords: User engagement; online controlled experiment; A/B test; delay; DFT; trend; sensitivity; directionality; quality metric; time series

1. INTRODUCTION

A/B testing (i.e., online controlled experiments) is a well known and widely applicable technique by modern Internet companies (such as web search engines [26, 20, 14], social networks [2, 41], streaming media providers [40] etc.). This state-of-the-art approach is used to improve web services based on data-driven decisions [39, 28, 10] in permanent manner and on a large scale: leading companies reported the number of run experiments per day that grew exponentially over the years (200 by Bing in 2013 [27], while 400 by LinkedIn [42] and more than 1000 by Google [20] in 2015).

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052664

Usually, a controlled experiment compares two variants of a service at a time: its current version A (control) and a new one B (treatment), by exposing them to two groups of users. The goal of this experiment is to detect the causal effect of the service update on its performance in terms of an *Overall Evaluation Criterion (OEC)* [30], a user behavior metric that is assumed to correlate with the quality of the service. Leading industrial Internet companies permanently develop new metrics that surpass the existing ones [28, 35]. This goal is challenging since an appropriate OEC should satisfy two crucial qualities: *directionality* (*interpretability*) and *sensitivity* [26, 33, 35, 10].

On the one hand, the value of the OEC must have a clear interpretation and, more importantly, a clear directional interpretation [10]: the sign of the detected treatment effect should align with positive/negative impact of the treatment on user experience. A metric with acceptable directionality allows analysts to be confident in their conclusions about the change in the system's quality, particularly, about the sign and magnitude of that change [33]. Many even popular user behavior metrics may result in contradictory interpretations and their use in practice may be misleading [26, 28]. On the other hand, the OEC must be sensitive: it has to detect the difference between versions A and B at a high level of statistical significance in order to distinguish the existing treatment effect from the noise observed when the effect does not exist [30, 35, 10]. A more sensitive metric allows analysts to make decisions in a larger number of cases when a subtle change of the service is being tested or a small amount of traffic is affected by the system change [28]. Improvement of sensitivity is also important in the context of optimization of resources used by the experimentation platform [24, 23, 35], since a less sensitive metric consumes more user traffic to achieve a desired level of sensitivity.

In the current study, we focus on the improvement of sensitivity of user engagement metrics, since the ones that represent user loyalty (the state-of-the-art *number of user sessions* [26, 38] and the *absence time* [17]) are accepted by modern Internet companies as good predictors of their long-term success [36, 26, 27, 28]. Besides the loyalty metrics, we consider the ones that represent the activity aspect of user engagement: the *number of user queries*, the *number of user clicks*, the *number of clicks per query*, and the *presence time* [13, 14, 15]. The loyalty metrics often very slowly respond to an evaluated service change, this effect is referred to as user learning of the treatment [20]: behavior of a user may change in terms of these metrics much later than the first interaction of the user with the treatment version (e.g.,

after several days or weeks). Hence, A/B tests with these OECs usually run for one or more weeks [26, 38], in particularly, in order to detect this *delayed treatment effect*. On the contrary, the activity metrics react faster to a treatment[1], are more sensitive than the loyalty ones [14, 16], but have ambiguous directional interpretations [26, 28].

The primary research goal of our work is to improve sensitivity of a loyalty metric, while preserving its directionality, by means of constructing its modification which exploits the possible presence of a delay in user learning of the treatment in terms of this metric. First, a delayed treatment effect of a metric could be revealed through the daily time series of the metric's measurements over the days of an A/B test. Thus, we study 5 metrics that *quantify the trend* in such time series. Second, we consider an alternative approach in which we eliminate an initial time period of a user's interactions with a web service from a metric's calculation procedure. We hypothesize that the information on user behavior from this initial period, on the one hand, does not contribute a lot to the treatment effect (due to the delay), but, on the other hand, may carry an additional noise that reduces the metric's sensitivity. We study 2 novel metric modifications (with several variants of their parameters) that are based on this *delay-aware* approach. In our experimental analysis of the studied metrics, we use 164 large-scale A/B tests run on hundreds of thousands of real users of Yandex (www.yandex.com), one of the most popular global search engines.

To sum up, our paper focuses on the problem, which is recognized as fundamental for the *present and emerging Internet companies' needs*: to develop more sensitive A/B test metrics with a clear directional interpretation consistent with long-term goals of a web service. Specifically, the major contributions of our study include:

- Trend- and delay-based metrics as novel engagement OECs for online controlled experiments.
- Validation of these metrics w.r.t. sensitivity and directionality on the basis of 164 large-scale real A/B experiment run at Yandex, showing that delay-based modifications can improve sensitivity of baseline loyalty metrics while preserving their directionality.

The rest of the paper is organized as follows. In Sec. 2, the related work on A/B experiments and user engagement is discussed. In Sec. 3, we remind the key points of A/B testing and introduce the engagement measures. The studied trend- and delay-based modifications are presented in Sec. 4. In Sec. 5 we make a brief analysis of them. We present our experimentation and lesson learned in Sec. 6. In Sec. 7, the conclusions and our plans for the future work are provided.

2. RELATED WORK

Early studies [34, 29, 30] on A/B testing were devoted to the theoretical aspects of the methodology. Subsequent work included studies of various aspects of the application of A/B testing in Internet companies: evaluation of changes in various components of web services (e.g., the user interface [25, 13, 33, 15], ranking algorithms [38, 13, 33, 15], ad auctions [4], and mobile apps [41]); large-scale experimental infrastructure [39, 27, 42]; different parameters of user interaction with a web service (speed [31, 28], absence [3], abandonment [28], periodicity [13, 12, 15], engagement [13, 14, 12, 16], and switching to an alternative service [1]); optimal scheduling of the experimentation pipeline [23]. The trustworthiness of A/B test results was studied through several "rules of thumb", pitfalls, and puzzling outcomes [5, 26, 28, 10]. The authors of [20] proposed the adaptation "cookie-cookie-day" of the classical treatment assignment in A/B test design in order to study long-term user learning of a treatment via the ad CTR metric. In our work, we address in turn the problem of the sensitivity improvement (over A/B tests with a classical design) of the state-of-the-art metrics that align with a service's long-term goals by modifying them to exploit a delay in user learning of a treatment.

Studies focused on the problem of sensitivity improvement constitute a substantial part of online A/B testing literature. Some of them are devoted to the alterations of the user groups involved in an A/B test (e.g., expanding of user sample [30], elimination of users who were not affected by the service change in the treatment group [38, 7]), as well as of the experiment duration (either real increasing [30], or virtual one through the prediction of the future [14]). Some other studies address the problem by the variance reduction techniques: the stratification, linear [11, 40] and gradient boosted decision tree regression adjustment [35]. Finally, this problem is addressed through a search for more sensitive metrics and their transformations [28] or through the use of more appropriate statistical tests and OECs: learning sensitive combinations of metrics [22], statistical tests for two-stage A/B experiments [8], the optimal distribution decomposition approach [33], Bayesian approach for hypothesis testing [6, 9], sequential testing for early stopping [24, 9], extensive comparison of different evaluation statistics and statistical tests [16]. The most relevant study to ours in the sensitivity improvement context is [13, 15], where the sign-agnostic Fourier amplitudes of user engagement time-series from [12] were refined by the phases of Fourier sine waves to be able to detect the treatment effect via changes in the trend of the time series. In our paper, we compare these metrics with novel trend-aware metrics (the normalized difference and the slope of the linear regression line, see Sec. 4.1) by conducting a more extensive evaluation of their sensitivity (164 A/B tests vs 32 ones in [13]) and, more importantly, their directionality. To the best of our knowledge, no existing studies on more sensitive variants of metrics applied an empirical evaluation of their directionality on a wide set of real experiments (as in our work).

3. PRELIMINARIES

A/B testing background. A typical A/B test (also known as a randomized experiment) [30, 26, 28, 19, 32] compares the performance of a new variant B (*the treatment*) of a web service and the current production variant A (*the control*) by means of a *key metric* $\mathtt{M}$, which quantifies user behavior. Users, participated in the experiment, (a user set $\mathcal{U}$) are randomly exposed (assigned) to one of the two variants of the service (i.e., $\mathcal{U} = \mathcal{U}_A \sqcup \mathcal{U}_B$). Then, the *average treatment effect* (ATE) defined as $\mathrm{ATE}(\mathtt{M}) = \mathrm{E}(\mathtt{M} \mid B) - \mathrm{E}(\mathtt{M} \mid A)$, is estimated by the difference $\Delta(\mathtt{M}) = \mu_B(\mathtt{M}) - \mu_A(\mathtt{M})$, where, given the observations of the metric $\mathtt{M}$ for $\mathcal{U}$, $\mathrm{avg}_{\mathcal{U}_V}\mathtt{M} =$

[1]Moreover, these activity metrics can suffer from primacy and novelty effects [30, 26], when the treatment effect may be overestimated in a short A/B test.

$\sum_{u\in\mathcal{U}_V} \mathtt{M}(u)/|\mathcal{U}_V|$ is the *Overall Evaluation Criterion* (*OEC*, also known as the evaluation metric, etc. [30]), $V \in \{A, B\}$.

The absolute value $|\Delta(\mathtt{M})|$ of the estimator should be controlled by a statistical significance test that provides the probability (called *p-value* or the *achieved significance level*, ASL [16]) to observe this value or larger under *the null hypothesis*, which assumes that the observed difference is caused by random fluctuations, and the variants are not actually different. If the p-value is lower than the threshold $p_{\text{val}} < \alpha$ (α=0.05 is commonly used [30, 28, 13, 14, 35, 10]), then the test rejects the null hypothesis, and the difference $\Delta(\mathtt{M})$ is accepted as statistically significant. The pair of an OEC and a statistical test is referred [16] to as an *Overall Acceptance Criterion* (*OAC*). The widely applicable *two-sample t-test* [11, 38, 8, 13, 7, 14] is based on the *t-statistic*:

$$\Delta(\mathtt{M})/\sqrt{\sigma_A^2(\mathtt{M})\cdot|\mathcal{U}_A|^{-1}+\sigma_B^2(\mathtt{M})\cdot|\mathcal{U}_B|^{-1}}, \qquad (1)$$

where $\sigma_V(\mathtt{M})$ is the standard deviation of the metric M over the users $\mathcal{U}_V, V = A, B$. The larger the absolute value of the t-statistic, the lower the p-value. The additional details of the A/B testing framework could be found in [30, 19, 32].

Studied user engagement measures. In our work, we consider a search engine as a particular case of a web service, and the following user engagement (UE) measures (as in [14, 12, 33, 16]) are studied:

- the number of sessions (S);
- the number of queries (Q);
- the number of clicks (C);
- the presence time (PT);
- the number of clicks per query (CpQ or CTR);
- the absence time per absence (ATpA).

These measures are calculated over a time period (a day, a week, etc.) for a user[2]. Following common practice [21, 26, 17, 38, 3, 13, 15], a session is defined as a sequence of user actions (clicks or queries) whose dwell times are less than 30 minutes. The presence time PT is measured as the sum of durations (in seconds) of the user's sessions observed during a considered time period. The measure ATpA [16] is the average duration of absences, where an absence is a time period between two consecutive sessions of a considered user (as in [17]). The measure CpQ could be regarded as *the CTR of the search engine result pages* [14] as well.

Note that the measures S, Q, C, and PT are additive with respect to the time period, while ATpA and CpQ are non-additive (they are ratio measures). The measures S and ATpA represent the user loyalty, whereas the measures Q, C, PT, and CpQ represent the user activity aspects of user engagement [26, 36]. Activity metrics are known to be more sensitive than the loyalty ones [13, 14, 12, 16]. Additional details on these measures and analysis of their relationships, sensitivity, and persistence across time could be found in recent studies [13, 14, 16, 15]

[2]We use browser cookie IDs to identify users as done in other studies on user engagement and online A/B testing [39, 17, 38, 16, 20].

4. METRIC MODIFICATIONS

Let us consider any UE measure (e.g., the number of sessions S) and a N-day time period (e.g., the period of an A/B experiment). Then this measure calculated over the whole time period is referred to as the *source* or *total metric* [14, 35] and is considered as the *baseline* key metric in our study.

4.1 Trend in a time series

A notion of a trend intuitively designates a direction of a metric w.r.t. time: whether the measure growths or falls during the time period. Hence, if we know that users slowly change the key metric as a reaction to the treatment during an A/B test (i.e., there is a delay), then these changes are expected to affect a trend metric. For instance, if a user enjoys with the treatment version, then the number of sessions should grow during the experiment, and this growth along the whole experiment period should thus leave traces in a corresponding trend metric. Thus, if we have a non-ambiguous interpretation of the positiveness or negativeness of a growth of the source UE measure (i.e., it has clear directionality), then one can set up whether a trend is positive or negative (i.e., match "growing" and "failing" with "+" and "–"). A change of this trend in positive or negative direction (e.g., measured by an A/B test), in turn, defines a clear directionality of a trend metric. However, note that this does not guarantee a consistency between the trend directionality and the one of the source UE metric, that we study in Sec.6.

First of all, let $\mathbf{x}=(x_0, x_1, .., x_{N-1})$ be the daily time series that represents the measure calculated for each of the N consecutive days of the time period (e.g., the daily number of sessions). We consider the daily time series only for additive measures[3], i.e., for S, Q, C, and PT.

Difference. The straightforward way to determine the trend of the N-day time series $\mathbf{x}$ is to simply compare the first half of the series with the second one. Hence, we study the absolute and the normalized differences between the average value of the time series $\mathbf{x}$ over the last $[N/2]$ days and the similar one over the first $[N/2]$ days:

$$D := \sum_{n=0}^{[N/2]-1} \frac{x_n}{[N/2]} - \sum_{n=N-[N/2]}^{N-1} \frac{x_n}{[N/2]} = \sum_{n=0}^{[N/2]-1} \frac{x_n - x_{N-1-n}}{[N/2]}$$

and $D_N := D \cdot N/\sum_{n=0}^{N-1} x_n$ respectively.

Discrete Fourier transform. Another way to measure the trend in $\mathbf{x}$ is based on the *discrete Fourier transform* (the *DFT*) and was earlier applied in the context of A/B tests with UE metrics [12]. Let us remind the key points of this method. After application of the DFT to $\mathbf{x}$, we obtain the sequence of its coordinates in the harmonic basis $\{\mathbf{f}^k\}_{k=0}^{N-1}$:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\mathbf{i}\omega_k n}, \quad \omega_k = \frac{2\pi k}{N}, \quad k \in \mathbb{Z}_N,$$

where $\mathbf{f}^k = (e^{\mathbf{i}\omega_k n}/N)_{n\in\mathbb{Z}_N}$ is the sine wave (harmonic) with the frequency ω_k. Presenting each coordinate as a complex number in the polar form $X_k = |X_k|e^{\mathbf{i}\varphi_k}$, we obtain the *amplitude* $A_k := |X_k|/N$ and the *phase* φ_k, $k \in \mathbb{Z}_N$. The amplitude A_k encodes *the magnitude* of the sine wave $\mathbf{f}^k$

[3]All our time-series based modifications utilize an aggregation of the daily values $\{x_n\}_{n=0}^{N-1}$ that, for ratio measures, results in misleading interpretations since, e.g., the total ratio metric is not usually equal to the sum of her daily values.

Table 1: Correlations between transformations of daily time series for number of sessions S calculated over 1-week periods (the top-right triangle) and 4-week periods (the bottom-left triangle).

4w \ 1w	φ_1	$\mathrm{Im}X_1$	$\mathrm{Im}X_{N1}$	$\mathrm{Re}X_1$	$\mathrm{Re}X_{N1}$	D	D_N	R_1	Sum
φ_1	–	**0.613**	**0.701**	−0.018	0.004	**0.528**	**0.593**	**0.452**	−0.001
$\mathrm{Im}X_1$	**0.502**	–	**0.636**	−0.068	−0.025	**0.891**	**0.563**	**0.802**	−0.013
$\mathrm{Im}X_{N1}$	**0.688**	**0.488**	–	−0.026	−0.007	**0.56**	**0.866**	**0.504**	−0.014
$\mathrm{Re}X_1$	−0.003	−0.045	−0.009	–	**0.63**	**−0.451**	**−0.283**	**−0.404**	−0.048
$\mathrm{Re}X_{N1}$	0.043	−0.007	0.01	**0.481**	–	**−0.269**	**−0.418**	**−0.241**	−0.04
D	**0.479**	**0.961**	**0.464**	−0.132	−0.047	–	**0.638**	**0.949**	0.01
D_N	**0.648**	**0.457**	**0.907**	−0.058	−0.087	**0.489**	–	**0.612**	0.005
R_1	**0.407**	**0.893**	**0.428**	−0.131	−0.044	**0.94**	**0.46**	–	0.01
Sum	−0.004	−0.041	−0.015	−0.11	−0.055	−0.03	−0.009	−0.029	–

with the frequency ω_k, presented in the series $\mathbf{x}$, whereas the phase φ_k represents how this wave is *shifted* along the time axis. The relative magnitude of the sine wave $\mathbf{f}^k$ is measured by the normalized amplitude $A_{Nk} := A_k/A_0$, where A_0 is actually the average value of $\mathbf{x}$ over N days (e.g., the average number of sessions per day) and is thus the baseline total metric divided by the constant N.

The sine wave $|X_1|e^{\mathbf{i}\varphi_1}\mathbf{f}^1$ of the first frequency has the N-day period and, hence, has the sole minimum and the sole maximum, whose positions along the time axis define which half of the time series $\mathbf{x}$ has more amount of the UE measure than the other one. Thus, we expect that this wave should change as a reaction to a presence of the treatment effect in the trend of $\mathbf{x}$. Therefore, we can define the trend by determining its magnitude from the amplitude A_1 (or the normalized one A_{N1}) and its sign from the phase φ_1. In [12], these two vital components were combined in one real-valued metric $\mathrm{Im}X_1 := NA_1 \sin\varphi_1$ that continuously and monotonically encodes the trend, i.e., in such a way that the higher (lower) the metric's value the more positive (negative) the trend is. A similar normalized variant of the metric was proposed as well: $\mathrm{Im}X_{N1} := \mathrm{Im}X_1/A_0$. The idea behinds the metrics $\mathrm{Im}X_1$ and $\mathrm{Im}X_{N1}$ means informally that we extract from the time series $\mathbf{x}$ the main component responsible for the trend and remove all components with higher frequencies treating them as a noise.

Linear regression. Finally, one can determine the trend in the time series $\mathbf{x}$ as the slope of the straight line adjusted to fit the data points $\{(n, x_n)\}_{n=0}^N$ by solving the ordinary least squares problem (i.e., the linear regression). Let $g_1 := \sum_{n=0}^{N-1} x_n$, $g_2 := \sum_{n=0}^{N-1}(n-1)x_n$, $s_1 := \frac{N(N-1)}{2}$, and $s_2 := \frac{N(N-1)(2N-1)}{6}$, then the slope of the regression line is

$$R_1 := \frac{g_1 s_1 - g_2 N}{s_1^2 - s_2 N},$$

which we study in our work. To the best of our knowledge, this quantity was never previously applied to measure the user engagement treatment effect in A/B tests.

We refer to all considered metrics that encode the trend in time series (i.e., D, D_N, $\mathrm{Im}X_1$, $\mathrm{Im}X_{N1}$, and R_1) as *trend transformations (or modifications)* of the baseline metric.

4.2 Measurement over a delayed period

Let $[0, t_e]$ (in hours) be the studied N-day time period. Let us consider an example and assume for simplicity that it takes d hours for the baseline metric $\mathtt{M}_{[0,t_e]}$[4] to react on the treatment during an A/B test, i.e., $\mathrm{E}(\mathtt{M}_{[0,d]} \mid A) = \mathrm{E}(\mathtt{M}_{[0,d]} \mid B)$. For an additive measure M, this implies the following:

$$\mathrm{ATE}(\mathtt{M}_{[0,t_e]}) = \mathrm{ATE}(\mathtt{M}_{[0,d]}) + \mathrm{ATE}(\mathtt{M}_{[d,t_e]}) = \mathrm{ATE}(\mathtt{M}_{[d,t_e]}),$$

i.e., calculation of the measure over *the delayed period* $[d, t_e]$, on the one hand, preserves the average treatment effect. On the other hand, this may increase sensitivity since the removed part $\mathtt{M}_{[0,d]}$ can bear an unnecessary variance that affects the denominator of the t-statistics, see Eq. (1). Of course, in the general case, we do not know the actual d, but we can use the time d as a parameter to trade off between (1) the ablation of $\mathrm{ATE}(\mathtt{M}_{[0,d]})$ from the treatment effect with a risk to lose a vital information on user behavior during the period $[0, d]$ and (2) the probable variance reduction. Depending on the choice of d and the measure M, the point (1) may decrease sensitivity, while the point (2) may increase it.

Last days. The simplest way to get a delayed period is to consider the part of the given N-day time period that consists of several last days. In this case, the metric calculated over the last $k \in \mathbb{N}$ days is just equal to the sum $\sum_{n=N-k}^{N-1} x_n$ for our time series $\mathbf{x}$ of the additive measure. We refer to these metrics as *last-days modifications* of the baseline metric and study them for $k = 1, \ldots, 7$ in our paper. It is important to note that, in this approach, the delayed period is the same for all users relative to the starting time point of an A/B experiment, while users are usually assigned to the experiment set $\mathcal{U}$ during the whole experimentation [26, 35]. Therefore, the time between the first interaction with the treatment and the start of accounting of user actions in the metric (the period with an assumed low reaction on the treatment) considerably varies among the user population[5].

Delayed calculation. In order to make users equal in terms of the duration of the period in which we do not measure their behavior, we consider the following metric modification. Let $t_f(u) \in [0, t_e]$ be the time of the first action of a user $u \in \mathcal{U}$ since the start of the considered time period [35]. Then, given a delay d as a parameter, we calculate the measure over the period $[t_f(u) + d, t_e]$[6] for the user with $t_f(u) < t_e - d$, otherwise she is removed from $\mathcal{U}$[7]. We refer to these metrics as *delayed modifications* of the baseline metric and, in our work, study them for several representative $d \in [12, 144]$ (in hours). The delayed modifications are studied for our ratio metrics `CpQ` and `ATpA` as well, since they are not based on daily values.

Both last-days and delayed modifications differ from the baseline metric only in the domain time period, hence, their directional interpretation is straightforward and clear. To sum up, we study 6 baseline metrics, 7 types of modifications (with 7 and 9 variants of parameters for last-days and delay ones respectively), and, overall, 108 various key metrics.

[4] $\mathtt{M}_{[t_1,t_2]}$ denotes the measure M calculated over $[t_1, t_2]$.

[5] E.g., let an A/B test's duration = 9 days, $k = 3$, the first action of a user-1 (a user-2) occurs at the 1-st day (the 5-th day); then, for the user-1, the delay is 6 days, while, for the user-2, is only 2 days.

[6] The baseline metric equals to the delayed one with $d = 0$.

[7] Note that this criteria does not harm the requirement of independence of treatment assignment to the treatment (critical for A/B test methodology), since a user decides to make a first visit to a web service previously to be affected by the treatment.

Table 2: The number of experiments with detected treatment effect (and sensitivity rates) over 164 A/B experiments for main trend transformations and last-days modifications of the metrics S, Q, C, and PT.

α=0.05	Baseline	Trend modifications					Metric calculated over last days:						
Metric	*Sum*	D	D_N	ImX_1	ImX_{N1}	R_1	1 day	2 days	3 days	4 days	5 days	6 days	7 days
S	17 (10.4%)	12 (7.3%)	12 (7.3%)	10 (6.1%)	18 (11%)	8 (4.9%)	**21 (12.8%)**	17 (10.4%)	12 (7.3%)	15 (9.1%)	16 (9.8%)	14 (8.5%)	15 (9.1%)
Q	**34 (20.7%)**	15 (9.1%)	11 (6.7%)	14 (8.5%)	17 (10.4%)	16 (9.8%)	27 (16.5%)	27 (16.5%)	21 (12.8%)	22 (13.4%)	27 (16.5%)	29 (17.7%)	31 (18.9%)
C	**51 (31.1%)**	18 (11%)	17 (10.4%)	16 (9.8%)	19 (11.6%)	18 (11%)	39 (23.8%)	39 (23.8%)	41 (25%)	47 (28.7%)	45 (27.4%)	46 (28%)	44 (26.8%)
PT	**25 (15.2%)**	15 (9.1%)	14 (8.5%)	11 (6.7%)	15 (9.1%)	21 (12.8%)	18 (11%)	18 (11%)	13 (7.9%)	16 (9.8%)	16 (9.8%)	17 (10.4%)	21 (12.8%)

5. ANALYSIS

In this section, we use logs of Yandex over more than tens of millions of the web service's users from a period in March–May, 2013. In Table 1, we present the Pearson's correlation coefficient over users between transformations of daily time series for number of sessions S calculated over a 1-week and 4-week periods (the results for other UE measures and other durations are similar). We underline and highlight in **boldface** those correlations that are larger 0.8 and 0.2 correspondingly. First, we see that the growth of the length of time series weakens mostly the lowest and moderate correlations (< 0.8), but strengthen mostly the strongest ones (e.g., between ImX_1 with D and R_1). Second, there are highly positively correlated clusters: $\{\text{Im}X_1, D, R_1\}$ and $\{\text{Im}X_{N1}, D_N\}$. Third, there is one cluster with moderate positive correlations $\{\varphi_1, \text{Im}X_1, \text{Im}X_{N1}, D, D_N, R_1\}$, i.e., all the transformations that are assumed to quantify the trend in a time series. Conversely, the orthogonal to ImX_1 (ImX_{N1}) component ReX_1 (ReX_{N1}) is negatively correlated with all trend modifications, and all transformations are uncorrelated with the baseline metric (or, equivalently, the sum of daily measurements of S). Thus, we conclude that all trend modifications are actually similar in their behavior and all of them carry an additional information w.r.t. the total metric.

In Fig. 1, we plot the joint distributions[8] of users with respect to each pair of the trend metrics ImX_1, D, R_1, ImX_{N1}, and D_N (i.e., 10 heat maps in log-scale) calculated over 4-week periods for the number of sessions S. These heat maps straighten the observation made above in Table 1 that the clusters $\{\text{Im}X_1, D, R_1\}$ and $\{\text{Im}X_{N1}, D_N\}$ are highly positively correlated.

6. EXPERIMENTATION

Experimental setup. In order to experimentally evaluate and compare our novel key metrics, we use 164 large-scale A/B tests carried out on *the users* of Yandex in the period 2013—2014 (they were extensively verified against any possible issue by production analysts since 2014). The user samples used in these tests are all uniformly randomly selected, and the control and the treatment groups are approximately of the same size, according to a common practice of industrial A/B testing [30, 28, 16, 35]. Each experiment has been conducted over at least several hundreds of users (from 0.5M to 30M with the median equal to 4M users) at least 7 days (up to 30 with the median equal to 14 days). These A/B tests evaluate changes in main components of the search engine, that include the ranking algorithm, the user interface, the server efficiency, etc. Each of those changes is either an update of a search engine component, which is evaluated before being shipped to production [10], or its artificial deterioration (e.g., a swap of the second and the fourth results in the ranked list returned by the current ranking algorithm as in [13, 33]). Also we used a hundred of control experiments (they compare two identical variants of the web service and are also known as A/A tests) that are common means to verify correctness of an experimentation platform and statistical tests used in studied OACs [30, 5, 16].

Statistical tests. In our experimentation, we utilize two most popular statistical tests in A/B testing: the two-sample Student's t-test [11, 38, 8, 13, 14] and the Bootstrap test [18, Alg. 16.1] with $B = 1000$ iterations as in [37, 33, 16]. First, we validate applicability of these tests to our OECs by means of the A/A experiments [30, 5, 16]: they should be failed (i.e, the treatment effect is wrongly detected) in not more than $\alpha \cdot 100\%$ of cases for the p-value threshold α (e.g., 5% for $\alpha = 0.05$), since p-value should be uniformly distributed over $[0, 1]$ on A/A tests. The fraction of failed A/A tests is referred to as the *false-positive rate* (the type I error), and we find that both t-test and bootstrapping on all our OECs do not fail the false-positive rate threshold $\alpha \cdot 100\%$ for $\alpha = 0.05$ and 0.01. Second, Drutsa et al. have shown by the extensive empirical analysis that p-values calculated by means of t-test and Bootstrap are very close to each other for per-user engagement metrics [16]. In our study, we observe the same situation for all our metrics and their modification. In Fig. 2, we present joint distributions of our 164 A/B experiments on the plane (p-value of Bootstrap, p-value of t-test) for some representative metric modifications. One can see that, for each OECs, all marks are close to the main diagonal, and the statistical tests report thus nearly the same significance level. Therefore, due to the space constraints, all presented results are given only for t-test (the results for the Bootstrap test imply the same conclusions).

6.1 Sensitivity

Following [13, 14, 12, 33, 16, 35, 22], we compare sensitivity of our key metrics in terms of the *success sensitivity rate*, which is the fraction of A/B tests whose treatment effect is detected by an OAC (i.e., by a key metric together with a statistical test) [16, 35]. In Table 2, we present the number of experiments with detected treatment effect (w.r.t. the state-of-the-art threshold $\alpha = 0.05$ of p-value) and corresponding success sensitivity rate over our 164 A/B experiments for each of the trend transformations and the last-days modifications of the additive user engagement metrics S, Q, C, and PT. The highest sensitivity rate in each row is highlighted in **boldface**. We see that only the loyalty measure S gains profit in terms of sensitivity from modifications: the measurement of S over the last day and the trend transformation ImX_{N1} of time series outperform the baseline total number of sessions. Activity metrics Q, C, and PT, in turn, noticeably suffer from the considered transformations.

[8]We hide all absolute values for confidentiality reasons.

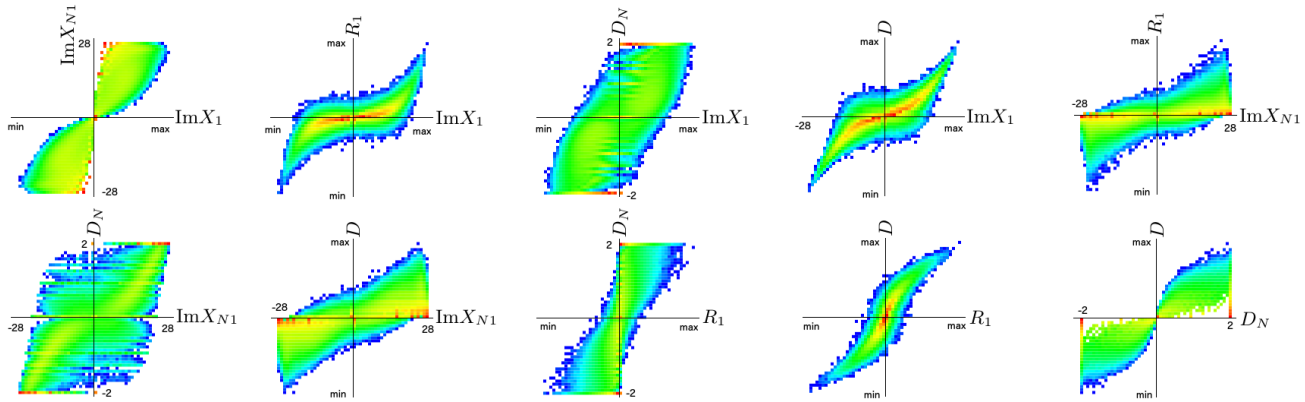

Figure 1: The joint distributions of users w.r.t. each pair of main trend metrics $\{\mathrm{Im}X_1, D, R_1, \mathrm{Im}X_{N1}, D_N\}$ calculated over 4-week periods for the measure S.

Table 3: The number of experiments with detected treatment effect (and sensitivity rates) over 164 A/B experiments for baseline and delayed variants of the metrics S, Q, C, PT, CpQ, and ATpA.

α=0.05	Baseline	Metric calculated with a delay in:								
Metric	Sum	12 hrs	24 hrs	36 hrs	48 hrs	60 hrs	72 hrs	96 hrs	120 hrs	144 hrs
S	17 (10.4%)	16 (9.8%)	16 (9.8%)	16 (9.8%)	19 (11.6%)	18 (11%)	18 (11%)	17 (10.4%)	**20 (12.2%)**	18 (11%)
Q	**34 (20.7%)**	30 (18.3%)	29 (17.7%)	29 (17.7%)	28 (17.1%)	27 (16.5%)	25 (15.2%)	24 (14.6%)	25 (15.2%)	25 (15.2%)
C	**51 (31.1%)**	46 (28%)	45 (27.4%)	43 (26.2%)	41 (25%)	40 (24.4%)	41 (25%)	42 (25.6%)	40 (24.4%)	42 (25.6%)
PT	**25 (15.2%)**	24 (14.6%)	**25 (15.2%)**	22 (13.4%)	23 (14%)	21 (12.8%)	20 (12.2%)	18 (11%)	21 (12.8%)	16 (9.8%)
CpQ	80 (48.8%)	82 (50%)	81 (49.4%)	82 (50%)	81 (49.4%)	82 (50%)	82 (50%)	**86 (52.4%)**	77 (47%)	74 (45.1%)
ATpA	10 (6.1%)	7 (4.3%)	**17 (10.4%)**	11 (6.7%)	12 (7.3%)	8 (4.9%)	11 (6.7%)	13 (7.9%)	11 (6.7%)	8 (4.9%)

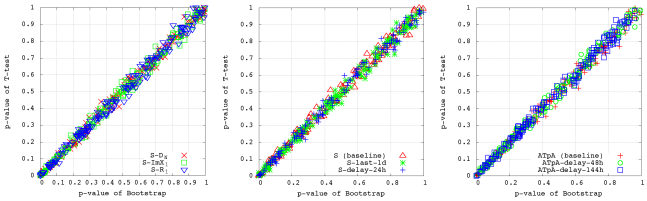

Figure 2: Joint distributions of 164 A/B tests on the plane (p-value of Bootstrap, p-value of t-test) for some representative metrics and modifications.

The DFT transformation $\mathrm{Im}X_{N1}$ demonstrates the best sensitivity among other trend modifications for the count-like measures S, Q, and C, while, for the time-based measure PT, the best one is the linear regression term R_1 (which, however, shows the worst sensitivity for S). All other trend transformations have relatively similar sensitivity for each measure. If we compare the last-days modifications and the trend transformations, then the former ones have either noticeably better sensitivity than the latter ones (generally, for S, Q, and C), or roughly similar performance.

In Table 3, we present the sensitivity rates (similarly to Table 2) for baseline and delayed variants of all studied metrics (i.e., S, Q, C, PT, CpQ, and ATpA). We see that, for the additive measures S, Q, C, and PT, delayed modifications mostly are not considerably better than the last-days ones (for PT only, some delay variants outperform the latter ones up to the sensitivity rate of the baseline total presence time). For the ratio measures, the delayed calculation of a metric demonstrates promising results: the modifications with most values of the delay outperform the baseline variant. The best improvement of the baseline sensitivity rate is achieved by the 24-hour delay for ATpA (+70%) and the 96-hour delay for CpQ (+7.5%).

Finally, in Fig. 3, we present joint distributions of A/B tests on the plane (p-value of M, p-value of S) for some representative modifications M of the metric S. First, we see that the p-values of trend modifications are completely uncorrelated with the ones of the baseline metric. Presumably, they detect the treatment effect different to the one of S (this is further confirmed by study of Δ in the next subsection). On the contrary, the other modifications have a relationship with S, and the smaller the calculation delay in such a modification the closer the p-values of the modified metric are to the ones of the baseline.

6.2 Directionality

In this subsection, in turns, we examine directionality [10] (also known as interpretability [33]) of novel metrics, the second important property of a good key metric. In the previous subsection, we found that mostly loyalty metrics S and ATpA gain sensitivity improvements from the studied modifications. Besides, these metrics are believed to align with long-term goals of a web service [26, 17, 10], therefore we further consider only these metrics and their modifications. Since we are strongly confident in the correctness of the directions of S and ATpA, we assume that the sign of the statistically significant difference Δ(S) (Δ(ATpA)) of the baseline metric S (ATpA[9]) is the ground truth that determines the actual sign of the treatment, i.e., positiveness or

[9]They are consistent with each other as shown in Fig. 5.

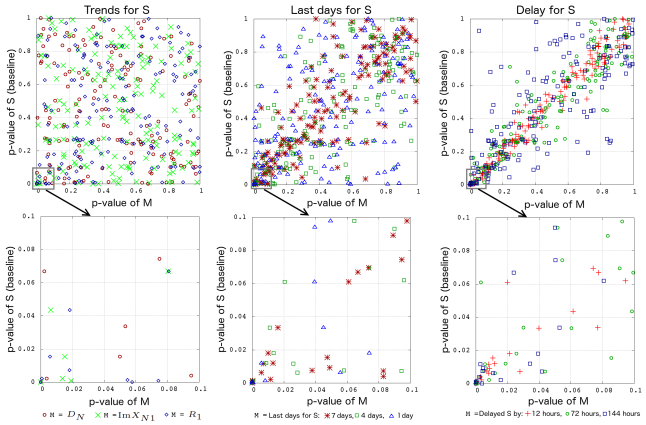

Figure 3: Joint distributions of A/B tests on the plane (p-value of M, p-value of S) for some representative modifications M of the metric S.

negativeness (w.r.t. user experience) of the service change evaluated by an A/B test[10]. Then, for each studied metric modification M, we compare the sign of its statistically significant difference $\Delta(\mathtt{M})$ with this ground truth. In order to compare studied metrics in terms of the magnitude and the sign of the average treatment effect, we calculate the scaled relative difference $\mathrm{Diff}_{\mathrm{pc}} = \varkappa\Delta/\mu_A$, where the constant $\varkappa$ is randomly chosen once in our study to hide real values for confidentiality reasons.

First of all, we take all A/B tests whose treatment is detected by at least one baseline, 23 in total (only for them we have the ground truth labels), and, for each of them, we report $\mathrm{Diff}_{\mathrm{pc}}$ of all studied metrics based on the measures S and ATpA in Fig. 4. A cell is highlighted in green (red) color for a positive (negative) effect detected (w.r.t. $\mathrm{p_{val}} < \alpha = 0.05$) by the corresponding key metric in the corresponding A/B test. We remind that the sign of the effect for ATpA-based metrics is opposite to the sign of $\mathrm{Diff}_{\mathrm{pc}}$, since a decrease of absence time is better w.r.t. user experience than its increase [17]. We see that most trend modifications ($\mathrm{Im}X_1$, $\mathrm{Im}X_{N1}$, and R_1) disagree with the ground truth: there are A/B tests both where the treatment effect sign of a modification matches with the baseline one and where the signs do not match each other[11]. On the contrary, all last-days modifications and the ones with delayed calculation demonstrate *absolute agreement with the ground truth*: there is no any contradiction for them on each A/B experiment. Note also that the magnitudes of $\mathrm{Diff}_{\mathrm{pc}}$ for these modifications are of the same order as the baselines S and ATpA, while the trend modifications have the magnitudes of $\mathrm{Diff}_{\mathrm{pc}}$ that are mostly higher by several orders.

Since the quantity of A/B tests with the ground truth labels is relatively small, we additionally evaluate agreement of novel metrics with the baseline ones by studying the correlations of their $\mathrm{Diff}_{\mathrm{pc}}$ over the whole set of 164 A/B experiments. In Table 4, we present the Pearson's correlation coefficients over the A/B tests between $\mathrm{Diff}_{\mathrm{pc}}$ of the baseline loyalty metrics (S and ATpA) and $\mathrm{Diff}_{\mathrm{pc}}$ of their representative modifications. In each column, the highest non-diagonal absolute correlation is highlighted in **boldface**, while the lowest one is underlined. First, the presented results support the observations made earlier on the A/B tests with ground truth: the relative ATE estimator of any trend metric has a very poor correlation with the one of any other key metric, while the last-days and delay modifications are strongly correlated with the baseline ones. For the additive metric S, this correlation decreases with the growth of the delay (with the decrease of the number of last days), and the most correlated modification is the calculation of S with the 12-hour delay. For the ratio metric ATpA, correlations of the delayed variants with the baseline one are high (the best is for the 24-hour delay), but are lower than in the case of S.

Second, we see that the trend modifications are all poorly correlated with each other in terms of the relative differences $\mathrm{Diff}_{\mathrm{pc}}$. That is quite surprising since these metrics are strongly correlated in terms of their values over users (see Table 1). Hence, we conclude that *a strong correlation of metrics over users does not mandatory imply that their ATEs are consistent*. On the contrary, last-days and delayed modifications are highly correlated with each other in terms of $\mathrm{Diff}_{\mathrm{pc}}$. These metrics are in fact the same metric but measured over different time periods, that presumably explain their high correlation in terms of both their ATEs over A/B tests and their actual values over users (see [14]).

In Fig. 5, we present joint distribution of A/B tests w.r.t. $\mathrm{Diff}_{\mathrm{pc}}$ of some representative pairs of metrics (the axes are logarithmically scaled). These plots demonstrate visually what lies behind the correlations from Table 4 and additionally straighten our conclusions. Moreover, the positions of magenta stars (the marks that correspond to the A/B tests whose treatment is detected by both metrics) clearly show which pair of metrics has consistent sign interpretations.

6.3 Discussion and lessons learned

First of all, our analysis shows that utilization of the delay property of the treatment effect is profitable for the sensitivity of the user loyalty measures (e.g., the state-of-the-art sessions-per-user), while the additive metrics of user activity become less sensitive with the same modifications. This result has been expected as it aligns with the knowledge that the metrics of user activity usually react quickly to web service changes (without a delay from the first user interaction with the treatment) [26, 28] (see Sec. 1). So, when we try to catch a delayed effect in their time series (e.g., by finding a trend or calculating the metric over a delayed time period) we may actually lose some information on user experience comprised in the time period right after the first user interaction. This is confirmed by the following dependences of the sensitivity rate on the parameters of the last-days and delayed modifications: the smaller the number of last days used (or the larger the delay value) the smaller the sensitivity rate of additive activity metrics is (see Tables 2 and 3). Here, for the loyalty metrics based on S, we see an opposite dependence, that, in turn, agrees with the knowledge that a user generally slowly accumulates her long-term experience on interactions with a web service and, then, shifts (learns) her loyalty to this service within weeks [38, 28, 14].

[10]Moreover, if an evaluated service change has an a-priory known effect on users, the sign of a significant difference of a loyalty metric agrees with this knowledge (e.g., in the case of an artificial deterioration, S has a negative difference Δ).

[11]The modification D_N is consistent with the ground truth, but we further show that its $\mathrm{Diff}_{\mathrm{pc}}$ does not correlate with the one of the baseline S (see, Fig. 5 and Table 4).

	S	Trends for S				Last days for S (in days)							Delay for S (in hours)									ATpA	Delay for ATpA								
# Exp	base	D_N	ImX_1	ImX_N1	R_1	1	2	3	4	5	6	7	12	24	36	48	60	72	96	120	144	base	12	24	36	48	60	72	96	120	144
1	-4,84	-85,56	40,84	90,82	-62,25	-6,53	-6,17	-5,97	-6,53	-6,47	-6,21	-5,98	-5,29	-5,30	-5,28	-5,22	-5,13	-5,14	-5,20	-5,20	-5,15	3,82	3,23	3,10	3,44	3,41	3,55	3,30	3,99	4,01	3,83
2	-3,87	-2,58	-4,89	0,27	-10,49	-4,56	-4,37	-4,06	-4,40	-4,16	-4,18	-3,87	-4,05	-3,91	-3,98	-3,95	-3,72	-3,70	-3,13	-2,18	-1,75	2,81	1,41	1,76	2,30	2,27	2,30	2,68	3,03	3,29	3,79
3	1,64	-18,45	-16,07	-13,73	-224,30	2,70	2,17	2,40	2,19	2,22	2,00	1,82	1,88	1,99	2,05	2,16	2,27	2,40	2,65	2,75	2,67	-0,35	0,11	-0,03	-0,42	-0,58	-0,75	-0,38	0,13	-0,47	-0,70
4	-1,10	-29,32	-2,74	-0,81	-179,63	-1,27	-1,42	-1,30	-1,33	-1,30	-1,32	-1,25	-1,32	-1,40	-1,54	-1,53	-1,45	-1,46	-1,31	-1,29	-1,13	-0,17	0,75	0,80	0,91	0,59	0,49	0,41	0,92	0,83	0,31
5	-1,10	-33,50	34,15	25,81	-45,17	0,27	0,00	-0,49	-0,61	-0,71	-0,85	-0,94	-1,17	-1,23	-1,16	-1,26	-1,12	-1,09	-0,91	-0,96	-0,93	1,05	0,22	0,35	0,85	0,61	0,52	-0,02	0,09	0,76	-0,04
6	-1,47	-19,95	46,32	-104,59	67,05	-0,16	-0,27	-0,54	-0,92	-1,11	-1,06	-0,98	-1,61	-1,62	-1,66	-1,67	-1,71	-1,72	-1,74	-1,76	-1,79	-0,29	-0,09	-0,01	0,28	0,01	0,16	0,22	-0,36	-0,59	-0,53
7	-1,02	-6,99	-14,19	15,58	-9,74	-1,06	-0,97	-0,96	-0,97	-1,05	-1,09	-1,14	-1,09	-1,14	-1,12	-1,12	-1,07	-1,08	-1,11	-1,13	-1,13	-0,64	-0,46	-0,51	-0,54	-0,56	-0,61	-0,09	0,44	-0,06	-0,19
8	-0,91	7,08	4,62	1,93	75,05	-0,49	-0,60	-0,64	-0,65	-0,67	-0,63	-0,63	-0,82	-0,81	-0,78	-0,79	-0,78	-0,80	-0,77	-0,80	-0,82	1,22	1,06	0,69	0,76	0,82	0,78	0,57	0,31	0,36	0,41
9	-1,37	-3,08	-1,42	39,73	16,62	0,10	-0,18	-0,75	-1,12	-1,17	-1,21	-1,18	-1,46	-1,50	-1,55	-1,58	-1,60	-1,61	-1,62	-1,67	-1,68	0,18	0,57	0,59	0,71	0,42	0,70	1,02	0,40	0,38	0,56
10	-1,33	-14,33	-16,84	-36,50	7,34	0,40	-0,10	-0,62	-0,92	-0,96	-1,09	-1,11	-1,43	-1,45	-1,48	-1,52	-1,53	-1,57	-1,57	-1,62	-1,66	0,40	0,89	1,31	1,10	0,70	0,74	0,51	0,25	0,23	0,38
11	1,21	9,05	13,39	-5,83	7,81	0,49	0,78	1,00	1,21	1,30	1,23	1,38	1,38	1,41	1,48	1,49	1,52	1,52	1,64	1,69	1,64	0,48	-0,21	-0,38	-0,13	-0,01	0,10	0,12	-0,11	0,36	-0,25
12	2,42	-12,51	-12,81	-48,10	-16,43	4,56	4,13	3,71	3,09	2,78	2,69	2,65	2,64	2,66	2,69	2,66	2,66	2,60	2,55	2,57	2,41	-0,25	-1,85	-2,30	-1,94	-1,65	-2,18	-1,03	-2,74	-2,37	-2,52
13	-0,87	-23,51	33,60	15,71	-37,73	0,51	-0,19	-0,48	-0,45	-0,40	-0,62	-0,74	-1,02	-1,02	-0,96	-0,91	-0,82	-0,71	-0,63	-0,66	-0,62	0,48	0,46	0,78	0,94	0,65	0,26	0,39	0,43	0,27	-0,59
14	1,66	-22,19	90,57	178,68	359,59	1,80	2,20	2,02	2,00	2,05	2,06	1,97	1,74	1,78	1,73	1,78	1,76	1,76	1,70	1,73	1,67	0,00	0,69	0,32	0,33	0,47	0,50	0,32	-0,49	-0,48	-0,57
15	1,49	-28,60	83,68	327,40	409,71	2,02	2,37	2,21	2,06	1,89	1,87	1,84	1,55	1,57	1,55	1,55	1,46	1,48	1,42	1,30	1,29	-1,56	0,10	0,24	0,80	0,59	0,61	0,53	0,40	0,23	0,30
16	-0,75	-2,18	16,80	-4,84	3,85	0,59	0,42	0,29	0,34	0,20	0,19	0,23	-0,70	-0,75	-0,76	-0,86	-0,84	-0,93	-0,98	-0,94	-0,90	-0,25	0,55	0,71	1,01	0,55	0,68	0,73	0,98	1,31	0,61
17	-0,99	-29,23	-25,41	-48,92	57,52	-1,04	-0,46	-0,29	-0,31	-0,36	-0,47	-0,55	-1,01	-0,96	-0,96	-0,97	-0,95	-0,93	-0,80	-0,69	-0,67	0,31	0,55	0,05	-0,18	0,42	-0,06	0,09	0,46	0,28	0,32
18	-0,53	-0,10	-1,52	0,22	-1,68	-0,54	-0,57	-0,72	-0,69	-0,72	-0,62	-0,54	-0,59	-0,61	-0,64	-0,65	-0,59	-0,58	-0,51	-0,55	-0,61	1,11	0,60	0,57	0,79	1,09	0,84	0,67	0,63	0,56	0,67
19	0,69	5,80	-1,67	0,56	1107,96	1,57	0,82	0,83	0,70	0,76	0,82	0,69	0,81	0,82	0,83	0,86	0,90	0,84	0,70	0,59	0,11	-1,19	0,04	0,04	-0,94	-0,83	-0,70	-0,27	-0,66	-1,64	1,70
20	0,58	-46,42	-6,59	-15,43	85,27	-0,61	-0,42	-0,06	0,21	0,31	0,32	0,37	0,64	0,65	0,62	0,64	0,69	0,73	0,75	0,71	0,71	-1,74	-0,64	-0,73	-0,79	-1,08	-0,97	-0,96	-0,88	-0,43	-0,64
21	-0,57	-255,52	-51,75	42,76	66,70	-0,96	-0,98	-0,79	-0,76	-0,64	-0,87	-0,76	-0,56	-0,53	-0,51	-0,56	-0,59	-0,61	-0,59	-0,60	-0,62	1,22	0,37	0,58	0,70	1,04	0,40	0,23	0,66	0,57	0,16
22	0,17	6,75	4,10	3,73	18,74	0,02	0,44	0,19	0,08	0,14	0,10	0,17	0,19	0,22	0,23	0,19	0,19	0,23	0,28	0,17	0,16	-0,87	-0,30	-0,23	-0,35	-0,42	-0,82	-0,24	-0,35	-1,03	-1,00
23	0,17	-2,39	-4,05	-3,90	-303,90	-0,55	-0,11	-0,13	0,13	0,05	0,09	0,17	0,18	0,28	0,23	0,27	0,11	0,17	0,16	0,14	0,14	1,35	-0,62	-0,49	-0,08	-0,42	-0,79	-1,10	-0,37	-0,31	0,79

Figure 4: Diff_{pc} **of all studied metrics based on the measures** S **and** ATpA **for all those A/B tests whose treatment is detected by at least one baseline; if a key metric detects a treatment (w.r.t.** $\alpha = 0.05$**), the corresponding cell is highlighted in green (red) color for a positive (negative) effect, see Sec. 6.2.**

Table 4: Pearson's correlations between Diff_{pc} **of** S**,** ATpA**, and their modifications over our 164 A/B tests.**

		S	Trends for S				Last days for S			Delay for S			ATpA	Delay for ATpA		
		(base)	D_N	$\text{Im}X_1$	$\text{Im}X_{N1}$	R_1	1 day	4 days	7 days	24 hrs	48 hrs	144 hrs	(base)	24 hrs	48 hrs	144 hrs
S	(base)	1	0.003	-0.087	0.01	0.089	0.811	0.923	0.955	0.992	0.985	0.917	**-0.545**	-0.415	-0.479	-0.464
	D_N	0.003	1	-0.153	-0.054	0.127	-0.061	-0.086	-0.05	0.006	0.009	-0.037	0.02	0.163	0.035	-0.005
	$\text{Im}X_1$	-0.087	-0.153	1	0.048	0.026	-0.093	-0.059	-0.062	-0.094	-0.103	-0.08	0.149	0.098	0.149	0.166
	$\text{Im}X_{N1}$	0.01	-0.054	0.048	1	0.052	0.004	0.04	0.035	-0.003	-0.007	-0.032	0.047	0.191	0.164	0.058
	R_1	0.089	0.127	0.026	0.052	1	0.162	0.107	0.101	0.089	0.085	0.048	-0.178	-0.01	-0.06	-0.014
	last-1d	0.811	-0.061	-0.093	0.004	0.162	1	0.872	0.848	0.811	0.805	0.777	-0.424	-0.378	-0.44	-0.424
	last-4d	0.923	-0.086	-0.059	0.04	0.107	**0.872**	1	**0.973**	0.921	0.917	0.869	-0.53	-0.437	-0.491	**-0.505**
	last-7d	0.955	-0.05	-0.062	0.035	0.101	0.848	**0.973**	1	0.952	0.946	0.905	-0.543	-0.446	-0.5	-0.499
	del-24h	**0.992**	0.006	-0.094	-0.003	0.089	0.811	0.921	0.952	1	**0.996**	0.92	-0.524	-0.434	-0.496	-0.447
	del-48h	0.985	0.009	-0.103	-0.007	0.085	0.805	0.917	0.946	**0.996**	1	**0.923**	-0.521	-0.439	-0.504	-0.454
	del-144h	0.917	-0.037	-0.08	-0.032	0.048	0.777	0.869	0.905	0.92	0.923	1	-0.477	-0.411	-0.475	-0.439
ATpA	(base)	-0.545	0.02	0.149	0.047	**-0.178**	-0.424	-0.53	-0.543	-0.524	-0.521	-0.477	1	0.45	0.522	0.487
	del-24h	-0.415	**0.163**	0.098	**0.191**	-0.01	-0.378	-0.437	-0.446	-0.434	-0.439	-0.411	0.45	1	**0.813**	0.486
	del-48h	-0.479	0.035	0.149	0.164	-0.06	-0.44	-0.491	-0.5	-0.496	-0.504	-0.475	0.522	**0.813**	1	0.489
	del-144h	-0.464	-0.005	**0.166**	0.058	-0.014	-0.424	-0.505	-0.499	-0.447	-0.454	-0.439	0.487	0.486	0.489	1

Overall, in the case of the last-days and delay modifications, the parameter d of such modification (the number of days or the delay value) is responsible for the trade off between the possible loss of the information on the user experience in the period right after the first user interaction and the variance reduction occurred in the key metric due to elimination of the part with this information (see Sec. 4.2). To sum up, we conclude that *exploitation of the delay property of a treatment effect can improve sensitivity of engagement metrics, while preserving their directionality*, which is the answer to the main research question of our work.

In addition to sensitivity, we examined directionality [10] that was never considered in the previous studies on transformed engagement metrics (e.g., [13, 12, 15]). For trend transformations, we have derived a clear directional interpretation (see Sec. 4): we can easily say whether a trend in a time series (and, thus, its change) is positive or negative based solely on the directional interpretation of the measure that underlie the modifications [12]. But, then, we discovered (Fig. 4) that the sign of their treatment effect does not align with the one of the state-of-the-art loyalty metrics S and ATpA (which we consider as ground truth w.r.t. user experience and long-term goals of a web service [26, 28]). Hence, we showed that *even if a metric can have a clear directional interpretation (e.g., derived from the one of the baseline metric), the sign of its treatment effect may actually disagree with the one of the baseline metric and, thus, poorly encode user experience feedback.*

Summarizing, we conclude that *only the modifications based on the delay principle (i.e., a sum of per-day metric values over last days of an experiment or a calculation of a metric over a delayed period for each user) have such directional interpretation that is clear and consistent with the one of the source baseline metric.* Thus, if one needs to improve sensitivity of a loyalty metric without harming its directionality, then we recommend to use last-days or delay modifications: for instance, based on our data, the last-day variant or the 5-day delay modification for S and 1-day delay one for ATpA. Nonetheless, if we are aimed to evaluate a service update in terms of a trend in time series of an additive metric, then $\text{Im}X_{N1}$ and R_1 are the best candidates (in terms of sensitivity) to be used as key metrics. But one should always keep in mind that their treatment effects (if being detected)

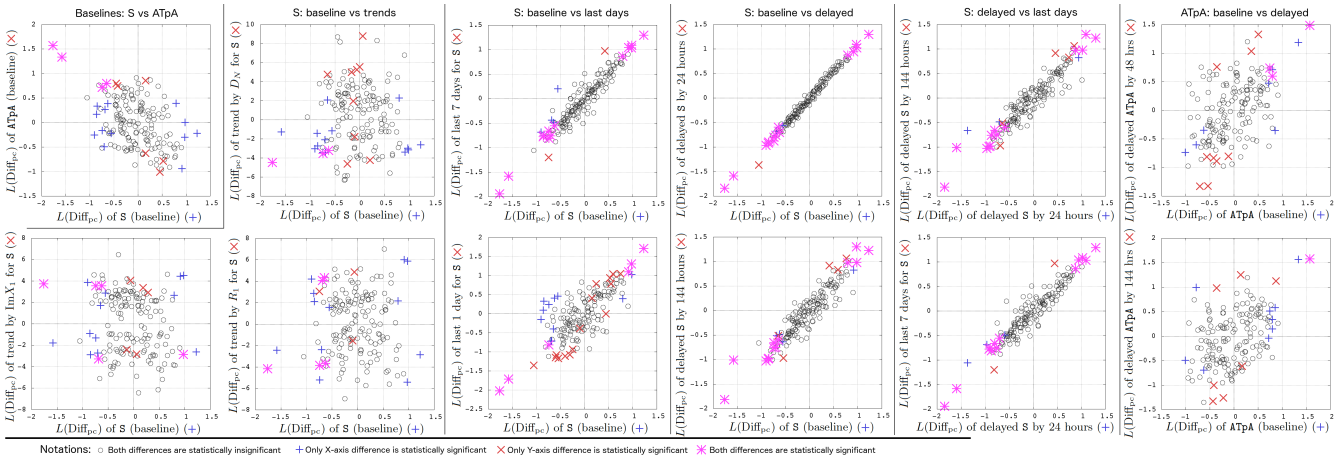

Figure 5: Joint distribution of our 164 A/B tests with respect to $L(\mathrm{Diff_{pc}})$ for some representative pairs of metrics, where $L(x) = \mathrm{sign}(x)\ln(|x|+1)$ and the statistical significance is measured with the threshold $\mathrm{p_{val}} < 0.05$.

may not align both with each other and with the ATE of the source metric that underlies the trend ones.

7. CONCLUSIONS AND FUTURE WORK

In our work, we focused on the problem of exploiting a delay in user learning of an evaluated treatment to improve sensitivity of some state-of-the-art user engagement metrics. We studied 21 variants of different metric modifications that are either based on methods of quantifying a trend in a metric's daily time series or represent a calculation of the metric over a delayed time period. We evaluated and compared their properties on a large and diverse set of 164 real large-scale A/B tests. of one of the global web search engines. First, we have shown that our novel metric modifications can improve sensitivity of the loyalty metrics. Second, we have found among these modifications the ones that preserve the directionality of the source metrics and, thus, agree with long-term goals of the web service. Hence, our study produces essential results that align with ongoing development of the best online metrics in modern Internet companies.

Future work. First, we can improve novel metrics by applying other methods of increasing sensitivity (like linear or boosted decision tree regression adjustment [35]). Second, one can study combinations of trend modifications and a baseline metric in order to find the one that is both sensitive to changes of trend in time series of the metric and consistent with the baseline treatment effect. Third, we can study more complicated methods to detect a trend in a time series.

8. REFERENCES

[1] O. Arkhipova, L. Grauer, I. Kuralenok, and P. Serdyukov. Search engine evaluation based on search engine switching prediction. In *SIGIR'2015*, pages 723–726. ACM, 2015.

[2] E. Bakshy and D. Eckles. Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods. In *KDD'2013*, pages 1303–1311, 2013.

[3] S. Chakraborty, F. Radlinski, M. Shokouhi, and P. Baecke. On correlation of absence time and search effectiveness. In *SIGIR'2014*, pages 1163–1166, 2014.

[4] S. Chawla, J. Hartline, and D. Nekipelov. A/B testing of auctions. In *EC'2016*, 2016.

[5] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In *KDD'2009*, pages 1105–1114, 2009.

[6] A. Deng. Objective bayesian two sample hypothesis testing for online controlled experiments. In *WWW'2015 Companion*, pages 923–928, 2015.

[7] A. Deng and V. Hu. Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *WSDM'2015*, pages 349–358, 2015.

[8] A. Deng, T. Li, and Y. Guo. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *WWW'2014*, pages 609–618, 2014.

[9] A. Deng, J. Lu, and S. Chen. Continuous monitoring of A/B tests without pain: Optional stopping in bayesian testing. In *DSAA'2016*, 2016.

[10] A. Deng and X. Shi. Data-driven metric development for online controlled experiments: Seven lessons learned. In *KDD'2016*, 2016.

[11] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM'2013*, pages 123–132, 2013.

[12] A. Drutsa. Sign-aware periodicity metrics of user engagement for online search quality evaluation. In *SIGIR'2015*, pages 779–782, 2015.

[13] A. Drutsa, G. Gusev, and P. Serdyukov. Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *WSDM'2015*, pages 27–36, 2015.

[14] A. Drutsa, G. Gusev, and P. Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW'2015*, pages 256–266, 2015.

[15] A. Drutsa, G. Gusev, and P. Serdyukov. Periodicity in user engagement with a search engine and its application to online controlled experiments. *ACM Transactions on the Web (TWEB)*, 11, 2017.

[16] A. Drutsa, A. Ufliand, and G. Gusev. Practical aspects of sensitivity in online experimentation with user engagement metrics. In *CIKM'2015*, pages 763–772, 2015.

[17] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *WSDM'2013*, pages 173–182, 2013.

[18] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[19] D. A. Freedman, D. Collier, J. S. Sekhon, and P. B. Stark. *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press, 2010.

[20] H. Hohnhold, D. O'Brien, and D. Tang. Focusing on the long-term: It's good for users and business. In *KDD'2015*, pages 1849–1858, 2015.

[21] B. J. Jansen, A. Spink, and V. Kathuria. How to define searching sessions on web search engines. In *Advances in Web Mining and Web Usage Analysis*, pages 92–109. Springer, 2007.

[22] E. Kharitonov, A. Drutsa, and P. Serdyukov. Learning sensitive combinations of a/b test metrics. In *WSDM'2017*, 2017.

[23] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. Optimised scheduling of online experiments. In *SIGIR'2015*, pages 453–462, 2015.

[24] E. Kharitonov, A. Vorobev, C. Macdonald, P. Serdyukov, and I. Ounis. Sequential testing for early stopping of online experiments. In *SIGIR'2015*, pages 473–482, 2015.

[25] R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, and T. Melamed. Online experimentation at microsoft. *Data Mining Case Studies*, page 11, 2009.

[26] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD'2012*, pages 786–794, 2012.

[27] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *KDD'2013*, pages 1168–1176, 2013.

[28] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *KDD'2014*, 2014.

[29] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD'2007*, pages 959–967, 2007.

[30] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.*, 18(1):140–181, 2009.

[31] R. Kohavi, D. Messner, S. Eliot, J. L. Ferres, R. Henne, V. Kannappan, and J. Wang. Tracking users' clicks and submits: Tradeoffs between user experience and data loss, 2010.

[32] S. L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

[33] K. Nikolaev, A. Drutsa, E. Gladkikh, A. Ulianov, G. Gusev, and P. Serdyukov. Extreme states distribution decomposition method for search engine online evaluation. In *KDD'2015*, pages 845–854, 2015.

[34] E. T. Peterson. *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram, 2004.

[35] A. Poyarkov, A. Drutsa, A. Khalyavin, G. Gusev, and P. Serdyukov. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *KDD'2016*, pages 235–244, 2016.

[36] K. Rodden, H. Hutchinson, and X. Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *CHI'2010*, pages 2395–2398, 2010.

[37] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR'2006*, pages 525–532, 2006.

[38] Y. Song, X. Shi, and X. Fu. Evaluating and predicting user engagement change with degraded search relevance. In *WWW'2013*, pages 1213–1224, 2013.

[39] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD'2010*, pages 17–26, 2010.

[40] H. Xie and J. Aurisset. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *KDD'2016*, 2016.

[41] Y. Xu and N. Chen. Evaluating mobile apps with A/B and quasi A/B tests. In *KDD'2016*, 2016.

[42] Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin. From infrastructure to culture: A/b testing challenges in large scale social networks. In *KDD'2015*, 2015.