#ml-papers March 2020

Automatic Detection and Diagnosis of Biased Online Experiments

Section 2: Triggered Analysis - Definitions

What is triggered analysis?

- in these experiments only impacted users are included to help separate the signal from noise
- it can be session vs user triggered; this paper/section deals with user triggered
- "triggered users" -> users affected by the experiment

Fully covered vs Partially covered metrics

- Fully covered metrics are fully nested under the trigger condition
- "In an experiment triggering on landing the profile page, profile views is a **fully-covered metric** as there is no way one can have a profile view without landing on the profile page. Non-triggered users all possess zero values in such metrics and triggered users also possess zero values on days they do not trigger. On the other hand, total page view is a **partially-covered metric** as one can contribute values to this metric through profile views as well as other channels such as home page. As a result, users can still have engagement on days they do not trigger"
- A good Stash example would be the learn recommendation endpoint, users who don't open the app and see the carousel/tab are not triggered

Rest of the section: they present math arguments for why user triggered analysis is beneficial for AB tests

Section 2: Triggered Analysis - Notation

- We show why triggered analysis is beneficial by evaluating the consequences of non-triggered analysis, where users that are not affected by the experiment are included (such as all-user analysis).

- This is broken down into 2 proofs, first they show that fully covered metrics have higher variance when performing all-user analysis and then they prove partially covered metrics have lower t scores when performing all-user analysis

- but first, some context..
  1. notation table
  2. what we're testing
  3. computing variance of effect using delta method

1

	Triggered Treatment	Triggered Control	All User Treatment	All User Control
Metric Sum	Σ_t	Σ_c	Σ'_t	Σ'_c
Sample Size	n_t	n_c	n'_t	n'_c
Variance	var_t	var_c	var'_t	var'_c

2

$$\Delta\% = \overline{X_t} \Big/ \overline{X_c} - 1 \qquad t = \frac{\Delta\%}{\sqrt{var(\Delta\%)}}$$

3

$$var(\Delta\%) = \frac{var_t}{\overline{X_c}^2 n_t} + \frac{var_c \overline{X_t}^2}{\overline{X_c}^4 n_c} \quad based \text{ on Delta method [13].}$$

Section 2: Triggered Analysis - Fully Covered Metrics

- non-triggered users have 0 values in fully covered metrics
- treatment effect is the same but variance is always higher for non triggered which means we have more noisy experiments with larger confidence bounds
- this conclusion feels very logical even without the breakdown, it makes sense that the variance would be bigger if we include the non-triggered users
- this proof is only for fully covered metrics.. what about partially covered ones?

$$E(\Delta\%') = E\left(\frac{\Sigma_t'/n_t'}{\Sigma_c'/n_c'} - 1\right) = E\left(\frac{\Sigma_t/n_t}{\Sigma_c/n_c} - 1\right) = E(\Delta\%)$$

And

$$\frac{var(\Delta\%')}{var(\Delta\%)} = \left(\frac{var_t' + r(1+\Delta\%)^2 var_c'}{var_t + r(1+\Delta\%)^2 var_c}\right)k$$

Note that $var_t' = \frac{1}{k}\left[var_t + \left(1 - \frac{1}{k}\right)\overline{X_t}^2\right]$ (similarly for var_c').

Hence

$$\frac{var(\Delta\%')}{var(\Delta\%)} = 1 + \frac{\overline{X_T}^2 + r(1+\Delta\%)^2\overline{X_C}^2}{var_t + r(1+\Delta\%)^2 var_c}\left(1 - \frac{1}{k}\right) > 1$$

Section 2: Triggered Analysis - Partially Covered Metrics

1. setup

2. variance inflation ratio: triggered users tend to have lower variance since they all share something in common (the trigger condition) so r_theta >= 1

3. plug in the variance formula

4. ratio between population t stat vs triggered t stat, the inequality holds because t`/t is monotone in r, the approximation holds because delta% is a very small number usually < 5%

1

can no longer assume $\Sigma_c' = \Sigma_c$.

$$\text{let } s = \Sigma_c' / \Sigma_c$$

$$var_t = var_c, var_t' = var_c',$$

2

$r_\sigma = var_c' / var_c$ be the variance inflation ratio.

3

$$\frac{var(\Delta\%')}{var(\Delta\%)} = \left(\frac{1 + r(1+\Delta\%/s)^2}{1 + r(1+\Delta\%)^2}\right)\frac{kr_\sigma}{s^2}$$

4

$$\frac{t'}{t} = \left(\frac{1 + r(1+\Delta\%)^2}{1 + r\left(1+\frac{\Delta\%}{s}\right)^2}\right)^{1/2} \frac{1}{kr_\sigma^{1/2}}$$
$$< \frac{1+\Delta\%}{1+\frac{\Delta\%}{s}} \frac{1}{kr_\sigma^{\frac{1}{2}}}$$
$$\approx \frac{1}{kr_\sigma^{\frac{1}{2}}}$$

STASH

Section 2: Triggered Analysis - Partially Covered Metrics

Last step:

- now we can rewrite like this, **SSc** is the sum of squares of the metric for the triggered members
- since **SS`c > SSc** and **n`c > nc** it follows that **kr^1/2_theta > 1** and **t`/t < 1**
- since the t stat of non triggered analysis is always smaller, the signal is weaker

$$kr_{\sigma}^{\frac{1}{2}} = \frac{n_c'}{n_c}\sqrt{\frac{var_c'}{var_c}}$$

$$= \sqrt{\frac{n_c'^2 var_c'}{n_c^2 var_c}}$$

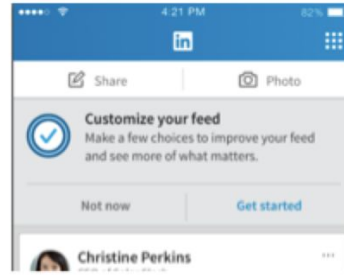
$$= \sqrt{\frac{n_c' SS_c'}{n_c SS_c}}$$

Section 3.1 Diagnosing Biased Experiments:

I. **Dynamic Targeting**
   A. moving samples in/out of a group as a part of the algorithm
   B. e.g. "Jobs You May Be Interested In"
      1. separate models for active / passive job seeker
      2. for passive group, model lifted page views and clicks on jobs pages
      3. as part of the classification model, removed some members from the targeted experiment (not passive any more)

II. **Cool-off**
   A. cap # impressions on certain pages to prevent traffic distraction
   B. e.g. "rebuild news feed by following contents you care about"
      1. only treatment group gets "cool off"
      2. it's hard to count traffic from the very beginning
      3. after a few iterations, treatment group traffic shrinks

```
If (member in cool-off):        (1)
       Do not show widget       (2)
Else if (member in control):    (3)
       Do not show widget       (4)
Else:                           (5)
       Show widget              (6)
```



Section 3.1 Diagnosing Biased Experiments:

III. Residual Effect

- A. a former experiment contaminate user split
- B. e.g. "People You May Know"
  - a. mismatched sample size (more in experiment)
  - b. because the algorithm was "so good it made people come back more often!"

IV. Biased Implementation

- A. Similar to residual effect, also causes mismatch
- B. e.g. "Revamp Home URL"
  - a. two entry methods for treatment:
    - i. hit router & type linkedin.com
    - ii. enter the new homepage directly with designated Urls
  - b. mismatched sample size because people enter from type ii but only treatment are counted

V. Dependent Experiment

- A. e.g. posting & checkout page redesign
  - a. new design: hide fees in posting and make checkout page more simple and clear
  - b. checkout page sample mismatch
  - c. because parent page changed CTR

Section 3.1: Diagnosing Biased Experiments

Generalization for Diagnosis

1. **Run Sample Size Ratio Test on all targeted users**
   -- check if treatment assignment is independent

$$\text{chi-squared statistic } \frac{(n_t - E_t)^2}{E_t} + \frac{(n_c - E_c)^2}{E_c}$$

2. **Compute sample size ratios separately for users who trigger for the first-time and returned users**
   -- identify bias from feedback loop

3. **Identify tracking events that are independent of triggering code calls that would reproduce the desired trigger condition**
   -- use page views to track if bias is from implementation or user engagement

4. **Track Metadata to separate code calls at different places**
   -- Splitting triggered traffic by service allows us to identify which part of the code to look into

5. **For each experiment sharing hashID (experiment 1) with the problematic experiment (experiment 2), identify users that trigger both experiments**
   -- discover chronological dependency

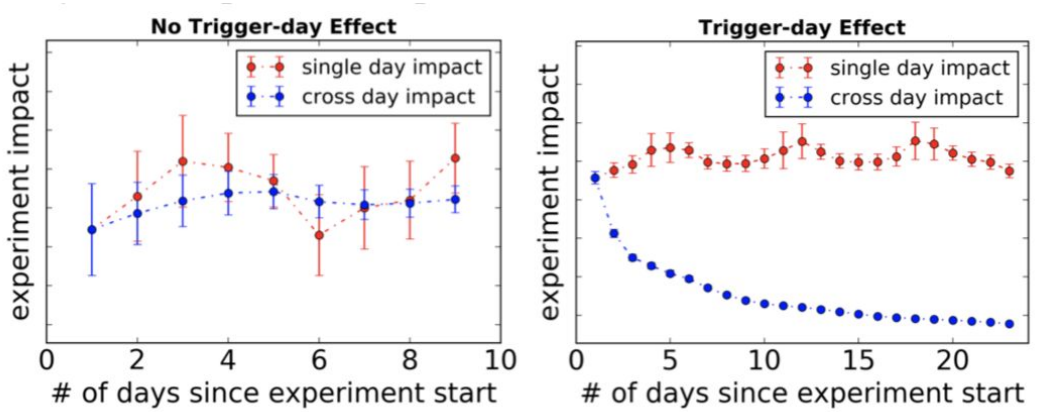
Section 3.2: Identifying Related Metrics

- Using a large number of in-house AB testing results, the describe a procedure to build a meta model to describe the correlation between two metrics.
- The primary usages are:
  - Identify why metrics move (e.g., look at correlations to possible experimental factors impacting both metrics)
  - Identify early indicators
    - E.g., metric X correlated with metric Y
    - Significant change in X detectable after 7 days
    - Significant change in Y only detectable after 30 days
    - Knowing that X and Y are correlated, we can detect changes in Y earlier

Section 3.3: Trigger-day Effect

- AB tests with triggers include users in the sample when they satisfy a trigger condition, e.g., landing on a specific page.
- If the trigger is activated, the user is included in the study for all days (not only the day they trigger).
- The **trigger-day effect** describes how the user behavior could be substantially different on the trigger-day.
    - Example: Pushing a Linkedin connection recommendation may increase connections by 100% on the day it is sent, but have no effect on days the recommendation is not pushed. In this example, the off-trigger impact is 0.

- Trigger day effect occurs when the **single-day impact** (over current day) differs from the **cross-day impact (**over all preceding days)
- The authors define an automated procedure to detect trigger-day effect.
- Identifying trigger-day effect can alert researchers that the effect of a test is evolving over time, and more time is needed for analysis.



STASH