Big Data Technology
Controlled Experiments

Eshcar Hillel
Yahoo!

Ronny Lempel
Outbrain

18 June 2017

Data Science Virtuous Cycle (Web)

Capture
Crawl, ingest feeds, record instrumented interactions...

Transfer
Move the data to a system capable of storing and processing it

Visualization

Experimentation & Metrics

Deploy/Serve
Tap output of previous step to improve user experience

Analyze/Model
Here data mining & machine learning take place
Business Problem

- We have an idea for some improvement that impacts user engagement (new algorithm, parameters, UX, ...)
- How do we validate that our idea actually improves upon the solution currently in production?
- Same question arises when we have two competing settings (not necessarily one “old” and one “new”) – how can we decide which setting is better?
- Different than, say, performance or stability changes that are transparent to users and whose measurements are typically (at least conceptually) simple

18 June 2017

Controlled Experiments

- Idea:
  - Identify metrics (KPIs) that determine which solution is better
  - Use some portion of the population as “control”
  - Give remaining portion a “treatment”
  - Measure difference in outcome between the populations to discern effectiveness of treatment
- A.K.A. “A/B testing”, “Bucket testing”
- Requires careful statistical analysis to distinguish statistically significant outcomes from random effects
### Controlled Experiments on the Web

**Control**

- Evaluation Criteria
  - engagement & monetization metrics

**Treatment**

- Statistical tests

**Data Driven Decision**

---

### Traffic Splitting

- Fraction of traffic to control is typically greater than or equal to that of the treatment

- “Sticky” experiments: typically done by randomly hashing user-ids or browser cookies
  - Must verify that the population is split in an unbiased manner to control and treatment

- “Non-Sticky” experiments: send every request to control/treatment independently, in a random manner

- Note that there may be multiple treatments in the experiment (not necessarily just one)
Interpreting Controlled Experiments

- We ran a controlled experiment
- The treatment produced metric value $x_T$, while the control exhibited that metric with value $x_C$
  - We are happy to observe that $x_T > x_C$
- Does that mean the treatment is better and we should switch to use it?
- We must decide whether $x_T$ is better than $x_C$ in a statistically significant manner (i.e. not by chance)

Hypothesis Testing

- In hypothesis testing language:
  - Null hypothesis: $E(x_T) = E(x_C)$, i.e. control and treatment have the same effect and what we've seen is some random result
  - Alternative hypothesis: treatment is indeed better than control
  - We must accept or reject the null hypothesis
- Type 1 error (false positive): we decide to reject the null hypothesis although it is actually true
  - Probability of this error is denoted by $\alpha$
  - $1-\alpha$ is called the confidence level of the test
- Type 2 error (false negative): we decide to accept the null hypothesis although it is actually false
  - Probability of this error is denoted by $\beta$
  - $1-\beta$ is referred to as the power of the test
Common issues in Hypothesis Testing

- For a given metric, what is a good statistical test?
  - T-tests, confidence interval methods, Chernoff bounds, Chi-Squared tests...
- The tradeoff between confidence level, power and sample sizes
  - fixing any parameter implies a tradeoff among the other two
- Given number of treatments to test, what confidence level should we test?
  - Setting to 0.95 (common value) means we incorrectly accept one out of 20 treatments as better than control; how many variations are we about to experiment with?

Duality of Estimation Variance and Confidence Intervals

- Let X be the random variable we’re trying to estimate
  - E.g. difference of some engagement metric between control and treatment
- We perform an A/B test, and sample the engagement of some users exposed to control and to treatment
- Let Y be an unbiased estimator of X, i.e. E[Y]=X, computed from the samples
Duality of Estimation Variance and Confidence Intervals (cont.)

- Let X be the random variable we're trying to estimate
- Let Y be an unbiased estimator of X, i.e. E[Y]=X, computed from the sampled users in an A/B test
- The variance of Y implies intuitively how far from the unknown value of X does Y typically stray
- The confidence interval around Y implies how far from the observed value of Y is X (with high probability)
- The two notions are both related and correlated

Example: Chernoff-Hoeffding Bound

- Let's assume a metric of interest ∈[0,1] per impression
- We estimate X_C – expected metric in control - with Y_C, the average of the metric over N_C users
- We estimate X_T – expected metric in treatment - with Y_T, the average of the metric over N_T users
- Assume Y_T-Y_C=2Δ
- Pr(X_T>X_C) ≥ Pr[ (X_C-Y_C)<Δ ] Pr [ (Y_T-X_T) <Δ ]
**Example: Chernoff-Hoeffding Bound**

- \( \Pr(X_T > X_C) \geq \Pr[(X_C - Y_C) < \Delta] \Pr[(Y_T - X_T) < \Delta] \)
- Chernoff-Hoeffding Theorem (simplified): let \( Z \) be the average of \( N \) observations of i.i.d. random variables in \([0,1]\) whose expectancy is \( \mu \). For any \( \Delta \),
  \[
  \Pr[Z - \mu \geq \Delta] \leq \exp\{-2 \Delta^2 N\}
  \]
  \[
  \Pr[\mu - Z \geq \Delta] \leq \exp\{-2 \Delta^2 N\}
  \]
- So: \( \Pr(X_T > X_C) \geq (1 - \exp\{-2 \Delta^2 N_C\})(1 - \exp\{-2 \Delta^2 N_T\}) \geq 1 - \exp\{-2 \Delta^2 N_C\} - \exp\{-2 \Delta^2 N_T\} \)

---

**A/A Testing**

- A good sanity check for the experimentation platform
- Split the users but expose both portions to control
- Useful for estimating variance of the metric over time/traffic
- The null hypothesis should, of course, prevail
  - However, don’t panic - it will be rejected occasionally (w.p. \( \alpha \))
- If significant change in metrics is consistently detected, this means that:
  - Instrumentation may be buggy
  - Metrics may be incorrectly computed or compared
  - Something in the supposedly identical setups is actually different
Pitfalls

- Humbling: even experienced professionals’ intuitions about experiments’ outcomes are often wrong
  - In many reputable companies, over 80% of tests fail to improve business metrics or lead to business changes
- Picking an evaluation criterion that’s easy to improve (in the short term) by doing something that is clearly wrong
  - Search engine example: bad algorithmic results increase both monetization and query share in the short term
  - Thus, Bing’s main evaluation criterion: search sessions per user
- Primacy and newness effects: changes in the UX may
  - ...take time getting used to, leading to low initial engagement [primacy]
  - ...seem new and exciting, leading to high initial engagement [newness]

Pitfalls (cont).

- Carryover effects – keeping user splits fixed may cause biases, where the same set of users exposed to some treatment would be exposed to a follow-up treatment
- Simpson’s paradox: uneven sampling of the population may cause an alternative to win all the battles, yet lose the war:

<table>
<thead>
<tr>
<th></th>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Stones</td>
<td>Group 1: 93% (81/87)</td>
<td>Group 2: 87% (234/270)</td>
</tr>
<tr>
<td>Large Stones</td>
<td>Group 3: 73% (192/263)</td>
<td>Group 4: 69% (56/80)</td>
</tr>
<tr>
<td>Both</td>
<td>78% (273/350)</td>
<td>83% (289/356)</td>
</tr>
</tbody>
</table>

Given a partitioning of the data, a trend may appear to hold in all partitions but is reversed when the entire data is examined.
Pitfalls (cont).

- Confirmation bias: the tendency to search for, interpret, favor, and recall information in a way that confirms one's beliefs or hypotheses, while giving disproportionately less consideration to alternative possibilities.

- Semmelweis Reflex: a reflex-like rejection of new knowledge because it contradicts entrenched norms, beliefs, or paradigms.

Concurrent and Overlapping Experimentation: Motivation

- Business needs may dictate the need to perform hundreds of concurrent experiments.
- Every treatment in every A/B test needs to get some non-negligible fraction of traffic:
  - The smaller the overall traffic, the shorter the desired test period, or the smaller the required error rates – the higher this fraction.
- If each treatment is tested separately against control, the throughput of the experimentation platform – and the rate of innovation of the site - is severely limited.
- Can experimentation throughput scale reliably?

Based on:
"Overlapping Experiment Infrastructure: More, Better, Faster Experimentation", Tank et al., KDD 2010
"Online Controlled Experiments at a Large Scale", Kohavi et al., KDD 2013
Full Factorial Designs

- Assume the need to run \( n \) experiments, \( E_1, \ldots, E_n \)
- Experiment \( E_j \) has control + \( T_j \) treatments, with traffic allocations \( A_j(c) \) and \( A_j(1), \ldots, A_j(T_j) \)
  - Summing up to 100%; typically treatments get equal allocations
- Using \( n \) independent random assignment functions, one can test all possible combinations of treatments:

Concurrent and Overlapping Experimentation: System Design

- The following is a simplified view of Google’s system for overlapping experiments in a multi-factorial design
- Experiments are grouped into sets (“layers”) of potentially conflicting experiments; users are exposed to no more than one treatment per any such layer
Concurrent and Overlapping Experimentation: Misc.

- Requires the ability to detect (unexpected) interactions between experiments
  - In reality, there are few such interactions
- Requires an alerting and monitoring system to detect (perhaps sudden or degrading) adverse user experience
  - "Fail fast"
  - But false alarms also have a cost!
- Has negative impact on cache hit rates of page elements that are the subject of experimentation
- System must be accompanied by extensive visual dashboards supporting data driven decision making

18 June 2017 236620 Big Data Technology

Interleaved Rankings

- Useful when testing competing top-k ranking algorithms
- Users aren't split; the ranking they see is a combination of the control and the treatment
- By observing the items that garnered the most interaction, the winning option is determined
- Approach was shown to be more sensitive than A/B testing
  - Reaching the same conclusions quicker, i.e. after fewer interactions

18 June 2017 236620 Big Data Technology
Interleaved Rankings - Theory

- Rationality assumption: users tend to click on relevant results more than on irrelevant results
  - When accounting for position
- Two basic questions:
  - Given rankings A and B, how should they be combined to produce interleaved ranking I?
    - What is actually constructed is a distribution over a set of allowed interleavings AI={I₁, I₂, ...} of A and B
  - Given clicks by a user on ranking I, how are those interpreted to a preference between A and B of that user?
  - The approach in the next slides due to Radlinski & Craswell, “Optimized interleaving for Online Retrieval Evaluation”, WSDM’2013

Interleaved Rankings – Construction

- To ensure that the experiment does not substantially alter the search experience, the following criteria should be met by each interleaving I∈AI:
  - Whenever A == B, their interleaving I should also be identical
  - The above is true for any identical prefix of A and B – I must begin with whatever results A and B agree on
  - If d₁ is ranked above d₂ in both A and B, the same order must also hold in I
  - Any document in I’s top-k must be in the top-k of either A or B
    - Is this necessarily a wise requirement?
Interleaved Rankings – Click Interpretation

- How should a user’s clicks on the interleaved results be interpreted as a preference between A and B?
- Two principles:
  1. For any clicked document d, the input ranking that had d ranked higher should get more credit
  2. The expected credit of each input ranking by a randomly clicking user should be the same, i.e. random clicking should not introduce preference bias
- Principle #2 embodies the constraints imposed on the distribution over the allowed interleavings in the set AI

Interleaved Rankings – Credit Functions

- Binary credits do not work!
  - Let A=(d1, d2, d3) and B=(d2, d3, d1)
  - AI = { (d1, d2, d3), (d2, d3, d1), (d2, d1, d3) }
  - Clicks on d2 and d3 favor B; only clicks on d1 favor A
  - There is no way to assign probabilities over AI such that random clicking will be agnostic – random clicking will favor B
- The following credit functions empirically work:
  - Credit\(_a(d)_i\) = -Credit\(_b(d)_i\) = rank(d, in A) – rank(d, in B)
  - Credit\(_b(d)_i\) = -Credit\(_a(d)_i\) = 1/rank(d, in B) – 1/rank(d, in A)
A/B Testing in Social Networks

- Standard A/B testing methodology faces challenges in social networks, as one’s experience may be influenced by the experiences and abilities of the connections.
- Specifically, one’s response to a treatment may depend on the number or proportion of connections exposed to the same treatment.
- Hence, the splitting of users between control and treatment should consider network topology.

Based on:
- Backstrom-Kleinberg, Network bucket testing, WWW’2011
- Katzir-Liberty-Somekh, Framework and algorithms for network bucket testing, WWW’2012
- Ugander-Karrer-Backstrom-Kleinberg, KDD’2013

A/B Testing in Social Networks: Network Exposure

- Exposure models: a user is said to be network-exposed to the treatment if:
  - The user & all of his/her neighbors are exposed to the treatment
  - The user along with at least some fraction/number of his/her neighbors are exposed to the treatment
  - Some more complex conditions exist in the literature
- Both treatment’s and control’s effects are measured only on network-exposed users, not all exposed users.
- Data of users who are not be network-exposed in either control or treatment, is left out of the analysis.
A/B Testing in Social Networks – Research Questions

- Given an exposure budget B, can we expose no more than B users so as to maximize the number of network-exposed users?
  - Equivalently, to increase the power of the test
- Are there network topologies that are easier to test, i.e. have low variance estimators?

A/B Testing in Social Networks – Flavor of Results

- Ugander et al. propose the following exposure protocol:
  - Cluster the graph somehow
  - Assign each cluster as a whole to either control or treatment
  - Nodes whose local neighborhood is (mostly) captured within their cluster become network exposed with high probability
- Good news: when the clusters are of constant size and the max degree of the graph is O(1), there is a way to compute an unbiased estimator to the effect of the treatment whose variance is O(1/N)
- Bad news: conversely, the variance of the estimator may grow exponentially in the degrees of the nodes
Ethics of Controlled Experiments

- In medicine, there’s a strict code of ethics regarding controlled experiments, and a rigorous approval process.
- For online controlled experiments, there’s no code of ethics often no rigorous approval process.
- Certain publicized online experiments (Facebook, okCupid) have raised questions regarding the etiquette of online controlled experiments.
- Not a life or death situation, but – for example - no informed consent.
- Is there an issue or is this an overreaction by people seeking a wrong parallel?