Unbiased Experiments in Congested Networks

Bruce Spang Veronica Hannan Shravya Kunamalla Te-Yuan Huang Nick McKeown Ramesh Johari
We use A/B tests to see if an algorithm works in practice
What is an A/B test?

1. Randomly assign traffic to treatment/control (users, sessions, servers, etc...)
2. Collect data
3. Compare outcomes
A/B tests are used to generalize

We make decisions about deploying algorithms based on small A/B tests:

“This algorithm improves performance by 10%”

This assumes that the **outcome of one unit does not depend on other units**

This is called interference
Examples of interference

Lots of examples from causal inference

**Social networks:** a treatment that increases usage might also cause increased usage for friends in the control group.

**Online auctions/markets:** if treatment/control users bid against each other, making treated users more likely to win means that control users are more likely to lose.

And many more!
Interference exists in congested networks

Shared Queue

- Treatment
- Control

Shared Link
Interference raises two questions

1. Does it matter?
2. What can we do about it?
Interference can make A/B tests extremely misleading

We ran an experiment which demonstrates this.
Treatment: capping bitrate to reduce traffic

In response to COVID-19, Netflix reduced traffic by 25% by capping bitrates.

Capping bitrates means that Netflix will not serve the highest quality versions of a video
Videos are encoded at many different qualities

High quality
Segment 1  Segment 2  Segment 3  Segment 4  Segment 5  ...

Mid quality
Segment 1  Segment 2  Segment 3  Segment 4  Segment 5  ...

Low quality
Segment 1  Segment 2  Segment 3  Segment 4  Segment 5  ...

Time
Bitrate capping limits video quality we can send

High quality
- Segment 1
- Segment 2
- Segment 3
- Segment 4
- Segment 5
- ...

Mid quality
- Segment 1
- Segment 2
- Segment 3
- Segment 4
- Segment 5
- ...

Low quality
- Segment 1
- Segment 2
- Segment 3
- Segment 4
- Segment 5
- ...

Time
What could A/B tests look like with bitrate capping?

Originally:
Link is congested

Control
Congested Link

With Capping:
Link is not congested

Capped
Uncongested Link

Capping causes:
- Less bandwidth used
- Less congestion

One possibility:
Bitrate capping reduces congestion

A/B test results:
- Capped uses less bandwidth
- Level of congestion is the same (no congestion)

Another possibility:
Control traffic increases, link stays congested

A/B test results:
- Capped uses less bandwidth
- Level of congestion is the same (some congestion)

A/B tests results do not reveal what happens when we cap traffic
Imagine control throughput increases as traffic is capped

Per-session throughput

% Capped Traffic

0  25  50  75  100

Control

Capped
We want to measure the effect of capping Per-session throughput

<table>
<thead>
<tr>
<th>% Capped Traffic</th>
<th>Capped</th>
<th>Control</th>
<th>Total treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A/B tests look at one point on this graph

**Per-session throughput**

- **Control**
- **Capped**
- **50% A/B test**

% Capped Traffic:
- 0
- 25
- 50
- 75
- 100
A/B tests give biased estimates of total treatment effects

Per-session throughput

Bias of 50% A/B test

Total treatment effect

Control

50% A/B test

Capped

% Capped Traffic

0  25  50  75  100
With two measurements, we can measure capping effects and A/B test bias

*Per-session throughput*

![Graph showing per-session throughput with capped and control traffic at 5% and 95% A/B test, with percentage capped traffic on the x-axis and throughput on the y-axis.](image-url)
Comparing A/B tests with a pair of congested links

Found two reliably congested peering links with well-balanced traffic

Run two A/B tests on each link and compare:
- **Link 1**: 95% capped, 5% uncapped
- **Link 2**: 5% capped, 95% uncapped
Capping improves throughput, despite A/B test results

Normalized Throughput

- 117% (95% A/B Test)
- 112% (Control)
- 100% (Capped)
- 96% (5% A/B Test)

5% Capped (Link 2) vs. 95% Capped (Link 1)

Total treatment effect
A/B tests are also wrong about effects on RTT

Normalized RTT

- 5% A/B Test
- 95% A/B Test
- Control
- Capped

Total treatment effect

5% Capped
(Link 2)

95% Capped
(Link 1)
Per-session Throughput results

**Before experiment**

- Throughput decreases during congested peak hours
- Two links have identical throughput

**During experiment**

- Capping delays the onset of congestion for capped link
- Behavior is similar within a link
A/B tests do not reliably estimate TTE

<table>
<thead>
<tr>
<th>Metric</th>
<th>Total Treatment Effect</th>
<th>A/B Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round Trip Time</td>
<td>25% better</td>
<td>5-15% worse</td>
</tr>
<tr>
<td>Throughput</td>
<td>12% better</td>
<td>5% worse</td>
</tr>
<tr>
<td>Play Delay</td>
<td>10% better</td>
<td>Did not change</td>
</tr>
</tbody>
</table>

and more in the paper…
A/B tests are biased when run in congested networks

This is concerning!
Risks of congestion interference

Common development process:
1. Come up with idea
2. Implement idea
3. A/B test idea
   Could give up too early on a good idea, or continue with an approach that doesn’t work
4. Iterate
   ...
5. Deploy idea
   Could deploy things that don’t work as expected, leading to production issues or longer development time
We can run experiments that remove bias

Paired link experiment is just one example

In the paper we also discuss:

- Event studies
- Switchback experiments
Use event studies when deploying changes

Switch most traffic to treatment and compare before/after

**Pros:**
- Estimates TTE
- Easy to do when deploying changes

**Cons:**
- Seasonality issues
Use switchbacks for more accurate measurements

Switch back and forth between treatment/control

Pros:
- Estimates TTE
- More robust to seasonality

Cons:
- Carryover effects
Lots more to be done!

- Any A/B test using a congested network has the possibility of bias
- We encourage more measurement to tell if interference matters for your experiments.
- We would love to see total treatment effects measured for new algorithms
- Need for better experiment methodology for networks

Thank you!
Email: bspang@stanford.edu

ArXiv: