

Unbiased Experiments in Congested Networks

Bruce Spang
Stanford University
USA
bspang@stanford.edu

Veronica Hannan
Netflix
USA
vhannan@netflix.com

Shravya Kunamalla
Netflix
USA
skunamalla@netflix.com

Te-Yuan Huang
Netflix
USA
thuangu@netflix.com

Nick McKeown
Stanford University
USA
nickm@stanford.edu

Ramesh Johari
Stanford University
USA
rjohari@stanford.edu

ABSTRACT

When developing a new networking algorithm, it is established practice to run a randomized experiment, or *A/B test*, to evaluate its performance. In an *A/B test*, traffic is randomly allocated between a treatment group, which uses the new algorithm, and a control group, which uses the existing algorithm. However, because networks are congested, both treatment and control traffic compete against each other for resources in a way that biases the outcome of these tests. This bias can have a surprisingly large effect; for example, in lab *A/B tests* with two widely used congestion control algorithms, the treatment appeared to deliver 150% higher throughput when used by a few flows, and 75% lower throughput when used by most flows—despite the fact that the two algorithms have identical throughput when used by all traffic.

Beyond the lab, we show that *A/B tests* can also be biased at scale. In an experiment run in cooperation with Netflix, estimates from *A/B tests* mistake the direction of change of some metrics, miss changes in other metrics, and overestimate the size of effects. We propose alternative experiment designs, previously used in online platforms, to more accurately evaluate new algorithms and allow experimenters to better understand the impact of congestion on their tests.

CCS CONCEPTS

• **Networks** → **Network experimentation**; • **Mathematics of computing** → *Probability and statistics*.

ACM Reference Format:

Bruce Spang, Veronica Hannan, Shravya Kunamalla, Te-Yuan Huang, Nick McKeown, and Ramesh Johari. 2021. Unbiased Experiments in Congested Networks. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA*. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3487552.3487851>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IMC '21, November 2–4, 2021, Virtual Event, USA

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9129-0/21/11...\$15.00
<https://doi.org/10.1145/3487552.3487851>

1 INTRODUCTION

Engineers routinely run *A/B tests* when testing new network algorithms. In an *A/B test*, the experimenter randomly allocates a small fraction of traffic (say 1% or 5%) to a new algorithm, called the treatment group, and compares its performance against the control group running the old algorithm. *A/B tests* are widely used as the gold standard for understanding how a new algorithm will behave at scale. Almost all large tech companies routinely use *A/B tests* to evaluate changes before deploying them [18, 22, 34, 48, 53, 58, 63, 70, 76]. Networking research often includes the results of *A/B tests*, and uses them to justify new algorithms [17–19, 24, 25, 29, 42, 46, 49, 55, 57, 58, 60, 63, 72, 87].

So when we recently ran experiments to test whether *bitrate capping* reduces network congestion for Netflix, we ran *A/B tests*. Bitrate capping was introduced in response to COVID-19; major streaming services cooperated with governments to lower bitrates offered and reduce overall internet load [3, 33]. This caused a reduction in congestion in certain networks around the globe.

We decided to dig deeper, to understand exactly how bitrate capping reduces congestion, and how doing so impacts video quality metrics. While we had data from just before and after bitrate capping was deployed (and later when it was removed), these were during periods of lockdown and stay-at-home orders when the internet was changing rapidly. We wanted to conduct a more systematic study of its effects. Naturally, we ran an *A/B test* where we capped a fraction of traffic to a very congested network.

In this *A/B test*, capping didn't appear to reduce congestion at all! In fact, it appeared to make things worse: capped traffic experienced 5% lower throughput and 5% higher delay. The *A/B test* results were so marginal that if we had not had evidence showing that bitrate capping reduced congestion when widely deployed, we might have dismissed it and not explored further. How could a treatment that we *knew* reduced congestion at scale not also reduce congestion in an *A/B test*?

Stepping back, we realized the confusion could be caused by *interference*. Interference is when units in the treatment group interact with units in the control group. It is well known in causal inference that interference can bias experiment results [45]. In social networks, changing something for a user in the treatment group can impact the behavior of their friends in the control group and bias the results of an experiment [28]. In online marketplaces, increasing the price of items in a treatment group can increase the demand for the relatively cheaper items in the control group and

bias results [39]. There are many examples of interference bias from markets, education, disease, and more [21, 38, 41, 52].

Both treatment and control groups in our test used the same network, and their packets traversed the same links and same queues. There is a long line of networking research showing that algorithms compete with each other when sharing a congested network [2, 5, 8, 15, 16, 23, 43, 44, 50, 56, 71, 81, 82, 84–86]. If capping bitrates freed up bandwidth, the uncapped control traffic could take up that bandwidth and get better performance. This could make bitrate capping look worse than it would if the uncapped traffic were not present, even if it was improving congestion. This gave us reason to believe that interference may exist, which would explain our unexpected A/B test results.

In this work we show that interference exists in experiments run in congested networks, and biases the results of A/B tests at scale. We show that bitrate capping does reduce congestion, and that the misleading A/B test result was due to interference. In order to do this, we propose and test new experiment designs which more accurately evaluate new algorithms. Our results suggest that usual A/B testing practice paints an incomplete picture of the performance of new algorithms in congested networks, and should be complemented by additional experiments.

Without interference, A/B tests give us a way to safely and accurately evaluate performance using a very small fraction of traffic. But because of interference, A/B tests on small fractions of traffic do not accurately predict performance at scale. Interference therefore creates a tradeoff between safety and accuracy: the only way to accurately measure performance is to run an algorithm on 100% of traffic, but nobody would do this with an untested algorithm! Our goal in this paper is to make the networking community, both academic researchers and industry practitioners, aware of this tradeoff and to propose techniques to help mitigate it. We encourage the community to apply these techniques broadly and evaluate networking algorithms with alternate experiments. We encourage continued measurement and the development of new techniques to mitigate bias.

We begin with an overview of experiment design in Section 2. We describe how A/B tests are run, and which quantities they estimate. Using a framework from the field of causal inference, we define the relevant quantities of interest for new networking algorithms.

We then run small lab experiments in Section 3 to give examples of how networking A/B tests can be biased. We show that experiments using multiple parallel connections, packet pacing, and different congestion control schemes all exhibit bias. If we were to evaluate these algorithms using naïve A/B tests, we would make incorrect conclusions. We might prematurely abandon a good algorithm, or deploy an algorithm that behaves worse when widely deployed than in the experiment.

Returning to our bitrate capping experiments, in Section 4 we describe our joint experiments with Netflix. We study the performance of bitrate capping and report on the bias we found in our initial A/B tests. While measurements show that bitrate capping significantly reduces congestion, naïve A/B tests do not reflect this behavior. Naïve A/B tests miss changes in some metrics, overestimate or underestimate the changes in others, and even get the *direction* of improvement wrong for a few. We were able to carry out this analysis due to a unique network architecture at Netflix.

Using a pair of reliably congested links with well-balanced traffic, we ran different experiments on each link and compared the results.

Based on our experience, in Section 5 we investigate possible ways experimenters can accurately evaluate new algorithms at scale. We discuss two possible paths to managing the tradeoff between safety and bias. The first is to adapt the common process of gradual deployments to measure interference. The second involves the use of small-scale, targeted *switchback* experiments to more accurately measure the effects of a new algorithm while managing safety concerns. We use the results of our paired link experiment to *simulate* what the experimenter might have obtained in these alternate approaches, and show that both substantially reduce bias.

We believe this paper is just the beginning of work on unbiased network experimentation. There is much to explore in designing more effective experiments, improving the analysis of experiments we run, and understanding the way interference behaves in networks. We wonder how many effective algorithms have been abandoned because of the way we run experiments, and what ineffective algorithms have been deployed because we were misled by A/B tests? Accordingly, we situate our work within the broader context of related research in Section 6 and conclude in Section 7.

2 WHAT WE WANT TO MEASURE

Before discussing experiments in more detail, it will be useful to give some background on how they are run, and what they can measure. In this section we provide a formal statistical foundation for A/B testing. The presentation is borrowed from causal inference [45]. The description is simplified, but gives enough conceptual scaffolding for the remainder of our work.

Treatment assignment. When we evaluate a new algorithm there are some *units* which run the algorithm. Units may be users, sessions, flows, connections, servers, etc... We let U be the set of all units. Each unit $i \in U$ is allocated to either *treatment* where it runs the new algorithm or *control* where it does not. Let A be the vector of treatment assignments to all units. We denote treatment as $A_i = 1$, and the set of treated units as T . We denote control as $A_i = 0$ and the set of control units as C .

Potential outcomes. When evaluating a new algorithm, we are interested in how it improves various metrics. In the language of causal inference, these metrics are called outcomes. Let $Y_i(A)$ be the outcome of interest on unit i given the vector of treatment assignments A . $Y_i(A)$ might be the average throughput of unit i , the minimum latency, or the 99th percentile packet loss. $Y_i(A)$ can be a random variable, since we expect some variability due to randomness in algorithms and randomness in arrivals.¹

Randomized unit assignment. In an A/B test, we randomly assign units to treatment independently with probability p or control with probability $1-p$. In other words, each A_i is an independent Bernoulli(p) random variable. We refer to the probability p as the *treatment allocation*.

To make this point more explicit, we introduce some additional notation. Define $\mu_T(p)$ (resp., $\mu_C(p)$) to be the average outcome value over the randomness in the assignment of treatment (resp.

¹This approach to causal inference via potential outcomes was pioneered by Neyman [75] (a 1990 translation of the original 1923 publication) and Rubin [66]; see [45] for details.

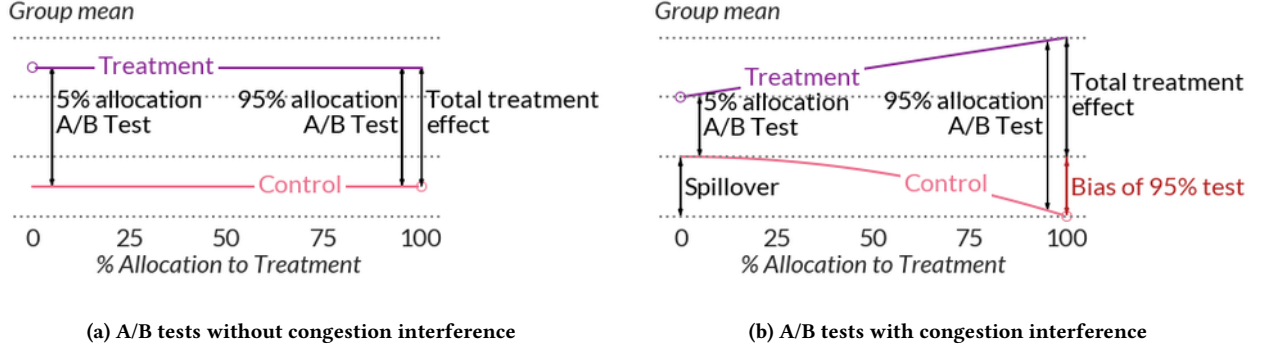


Figure 1: A/B tests are used to estimate the total treatment effect: how much better a treatment is than control if both were deployed globally. A/B Tests give accurate estimates of Total Treatment Effect (TTE) when there is no interference between sessions as in (a), but may be misleading when there is as in (b).

control), when the treatment allocation is p :

$$\mu_T(p) = \mathbb{E}_{T \subset U} \left[\frac{\sum_{i \in T} Y_i(A)}{|T|} \right].$$

Depending on the setting and the treatment, $\mu_T(p)$ may or may not depend on the treatment allocation p . This is visually depicted in Figure 1. $\mu_T(p)$ is the purple treatment line, and $\mu_C(p)$ is the pink control line.

Average treatment effect. An A/B test evaluates the average treatment effect. This is how much better the treatment group performs than the control group, when a p fraction of the traffic is allocated to treatment and $1 - p$ to control. It is defined as:

$$\tau(p) = \mu_T(p) - \mu_C(p), \quad (1)$$

This is visually depicted in Figure 1. The treatment effect at any point on the graph is the difference between the treatment and control lines.

Total Treatment Effect. When evaluating a new algorithm, we are often interested in what *would* happen if we were to deploy it widely. This is the *Total Treatment Effect*, or TTE: the difference between the average outcome when *all* flows are in treatment and when all flows are in control. In terms of our notation above:

$$\text{TTE} = \mu_T(1) - \mu_C(0).$$

This is depicted in Figure 1: it is the difference between the right-hand side of the treatment line (when all traffic is treated), and the left-hand side of the control line (when all traffic is allocated to control). Depending on the setting, it may or may not equal the average treatment effect.

Note that this definition of TTE is from the perspective of the experimenter, and not the internet. The experimenter may only control a small fraction of all traffic on the internet, and in this case TTE measures what happens if they switched all traffic under their control to a new algorithm. The TTE is also sometimes called the “global average treatment effect” in causal inference work (e.g., [51]), but we have avoided this name to avoid confusion around this point.

It is also reasonable to talk about TTE in specific groups of traffic. For instance, we may be interested in the TTE if we were to move all traffic globally to a new algorithm, but we may be also interested

in the TTE for a single network or a group of networks. This can be incorporated into the definition by changing the set of treatment and control flows.

Spillover. In addition to how well a new algorithm performs on its own, we are often also interested in how a new algorithm impacts existing algorithms. Recently, [84] defined the notion of the “harm” of a new algorithm, which is the negative effect caused by a new algorithm competing with an existing algorithm. This networking concept is similar to the concept of *spillovers* in the causal inference literature (e.g. [21, 37]). Formally, we define the spillover of treatment on control as the effect of increasing the treatment fraction to p on *control* units, relative to when the treatment units were not present. In terms of our notation:

$$s(p) = \mu_C(p) - \mu_C(0).$$

Spillover is non-zero when deploying a treatment algorithm has some impact on the control algorithm. This is shown in Figure 1b. Note that spillover is only defined for $p < 1$. If $p = 1$, there is no control traffic and no spillover can occur.

Spillovers may or may not be undesirable. It is possible that deploying a new algorithm can improve existing traffic, and we will see examples of this later.

Estimating from A/B tests All the quantities above are expectations over the distribution of all possible treatment assignments. Any experiment has only one set of treatment assignments and can only observe one set of potential outcomes—all other potential outcomes are missing. The fundamental problem in causal inference is to reason about these missing outcomes given what we observe.

In causal inference, we use the observed outcomes to estimate the quantities above. An estimator is called *unbiased* for some quantity if its expectation is equal to that quantity.

In an A/B test we randomly allocate units to treatment or control, and measure

$$\widehat{\mu}_T(p) = \frac{\sum_{i \in T} Y_i(A)}{|T|}.$$

This process gives an unbiased estimator of $\mu_T(p)$, since $\mathbb{E} \widehat{\mu}_T(p) = \mu_T(p)$, and similarly for $\mu_C(p)$. By linearity of expectation,

$$\widehat{\tau}(p) = \widehat{\mu}_T(p) - \widehat{\mu}_C(p)$$

is an unbiased estimator for $\tau(p)$, and we can define similar estimators \widehat{TTE} , and $\widehat{s}(p)$.

Congestion Interference In virtually all real-world experiments in networking today, experimenters run an A/B test. They infer that an improvement in the A/B test implies an improvement if the treatment were to be deployed. In our notation, this means that they use $\widehat{\mu}_T(p)$ and $\widehat{\mu}_C(p)$ as an unbiased estimate of the average treatment effect $\tau(p)$, and then interpret $\tau(p)$ as if it were the TTE. This is what we refer to as “naïve” A/B testing.

This process gives an unbiased estimate of TTE only in the very special case when the outcome of a unit does not depend on the fraction of other units allocated to treatment. This is part of the Stable Unit Treatment Value Assumption (SUTVA) [45], and requires that $TTE = \tau(p)$ for all p , and that spillovers are zero for all p . Visually, this process assumes that algorithms behave like Figure 1a and not Figure 1b.

Any A/B test that runs over a congested network has a clear pathway for interference between units in the treatment and control groups. Any explicit or implicit change in how the treatment group uses the congested network can create a different network condition for the control groups, which may lead to different behavior. This is *especially* true if the test explicitly changes the timing of how traffic is sent, or the amount of traffic that uses the network. Because of this, we will refer to violations of SUTVA as *congestion interference*.

Note on averages Average treatment effects, spillovers, and TTE are all defined as averages. Average here refers to the distribution of units in the A/B test, and not the outcome metric. The average treatment effect could measure the average difference in average latency, but it could also measure the variance of average latency or 99th percentile latency. Practitioners may also be interested in *quantile* treatment effects, e.g. the difference in 99th percentile latency between treatment and control. These are regularly estimated from A/B test results [1, 78]. It is straightforward to adapt our definitions to measure quantile treatment effects, and could be done by replacing $\mu_T(p)$ and $\mu_C(p)$ with quantile estimators.

3 SMALL LAB EXPERIMENTS

When interference is present, naïve A/B tests do not accurately describe the behavior of a new algorithm. They mispredict the TTE and give no estimate of spillover. To illustrate this, we set up a small test network in the lab. The lab setup gives us a global view of how a new algorithm performs at any fraction allocation, and lets us recreate Figure 1 for actual algorithms. With these results, we can look at the results of different A/B tests, estimate TTE, and measure spillover. These experiments do not tell us how different algorithms would behave at scale, but they provide easy-to-understand examples of how congestion interference causes bias in naïve A/B tests.

Lab Setup Our lab consists of two servers running Linux 5.5.0, each with an Intel 82599ES 10Gb/s NIC. Each NIC is connected to a port of a 6.5Tb/s Barefoot Tofino switch via $4 \times 10\text{Gb/s}$ breakout cables. The switch has a 1 BDP buffer. The sender server is connected to the Tofino with two 10G cables. The interfaces are bonded and packets are equally split between them, which ensures that congestion happens at the switch (otherwise we only see congestion at the sender NIC). We set MTUs to 9000 bytes so the servers can sustain a

10Gb/s rate. We add 1ms of delay at the sender using Linux’s traffic controller `tc`, and use `iperf3` to generate TCP traffic.

3.1 Test 1: Multiple connections

Web browsers, video streaming clients, and other applications request data over multiple TCP connections in parallel. Making simultaneous requests reduces head-of-line blocking, reduces page load time, and increases utilization [35, 36, 69, 73]. This behavior depends heavily on the particular ways an application uses TCP connections and the particular networks it traverses, and so would typically be evaluated with a large-scale A/B test.

However, using multiple TCP connections can also allow an application to outcompete its peers and achieve higher throughput, and so is often called “unfair” in the academic literature [8, 15]. This makes it an ideal example to illustrate how congestion interference can bias A/B tests.

We ran an experiment in the lab to illustrate this behavior and understand the bias it causes. We ran eleven tests in which ten applications used either one or two TCP Reno connections to transfer bulk data. We measured the average long-term throughput and retransmission rates experienced by each application.

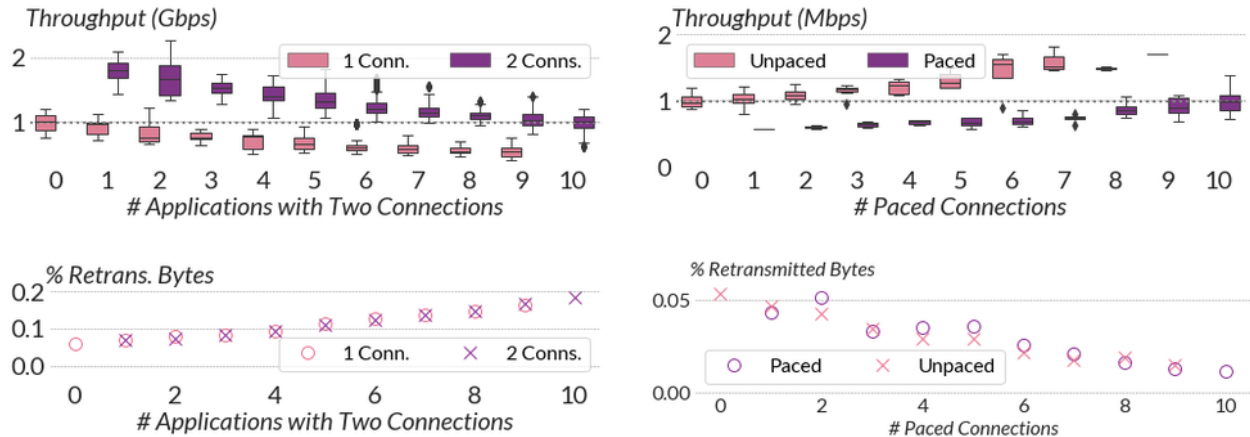
Figure 2a shows the results of the lab tests. Each test has two boxplots showing the average throughput for applications using one or two connections. Applications using two connections had 100% higher throughput and identical retransmission rates than applications using one. As more applications used two connections, their average throughput decreased. When all applications used two connections, their average throughput was identical to when all applications used one. Even worse, retransmission rates were higher when all applications used two connections.

These results are because of the way TCP fairly shares throughput between connections. If n identical TCP connections share a bottleneck link of capacity C , we expect each to receive a long-term average throughput of C/n . A group of flows with two connections should get a throughput of $2C/n$, 100% larger than C/n . But fundamentally, increasing the number of connections does not increase the capacity of the link so there can be no overall improvement.

This behavior is a well-understood consequence of TCP Reno’s throughput fairness. But suppose we followed common practice [17–19, 24, 25, 29, 42, 46, 49, 55, 57, 58, 60, 63, 72, 87] and ran an A/B test to measure how using two parallel connections performed. To illustrate the potential for bias, we will use the same data set interpreted in a different way.

In a naïve A/B test, we would randomly allocate some fraction of traffic to treatment and the rest to control. Treatment would use two connections and the rest would use one. We would compare the throughput and retransmissions of the treatment and control groups. *No matter what allocation we picked*, we would see that two connections have a 100% higher throughput than one, and that there was no impact on retransmission rates. The naïve interpretation is that we should always use two connections in production.

TTE and spillover give us a better idea of how two connections perform. The TTE shows that there would be no improvement in throughput and a 200% increase in the percentage of retransmitted bytes if all traffic were switched to two connections. Spillovers allow us to measure the impact of using two connections on other



(a) Units are applications using 1 or 2 long-lived TCP connections.

(b) Units are TCP connections which either pace traffic or not.

Figure 2: Throughput and retransmits in experiments where 10 units share a 10 Gb/s link. Every point on the x-axis is a different A/B test. All tests suggest a large change in throughput and no change in retransmissions, but the difference between 10 treated and 10 control units (TTE) is zero for throughput and large for retransmissions.

applications. When nine applications use two connections, the spillovers on the one remaining application using one connection are a 25% decrease in throughput and an almost 175% increase in retransmissions.

These results demonstrate that any single A/B test would not accurately measure the impact of changing the number of connections. But we should be careful not to extrapolate too much from the lab results. Applications may benefit from being more aggressive, but using multiple connections can also increase utilization. Without more experimentation, either could be a plausible explanation for a measured increase in throughput. Fundamentally, we believe that the only way to accurately measure the performance of such a policy would be to run an experiment at scale, on real traffic. We will discuss how to run such experiments later in Section 5.

3.2 Test 2: Pacing

Pacing is a generic, widely-used mechanism for reducing packet burstiness in a network [2, 17, 61, 67]. With pacing, a host adds delay between successive packets so that it sends a smooth, evenly paced stream of data into the network.

The Linux Kernel has supported pacing for TCP since 2013 [26, 27]. It adds delay between successive packets to ensure a rate of $2 \times cwnd/RTT$ during slow start and $1.2 \times cwnd/RTT$ during congestion avoidance [79].

Prior work, using ns-2, has shown that unpaced TCP traffic outcompetes paced traffic in terms of throughput [2, 86]. They recommend pacing at a rate of $(cwnd + 1)/RTT$, which is implemented by Linux. These fairness concerns suggest that spillover may be nonzero, which implies that there would be congestion interference in an A/B test.

We ran pacing A/B tests in our lab to measure whether this interference still exists and if it would impact the results of an A/B test. Figure 2b shows the results. Paced traffic (the treatment)

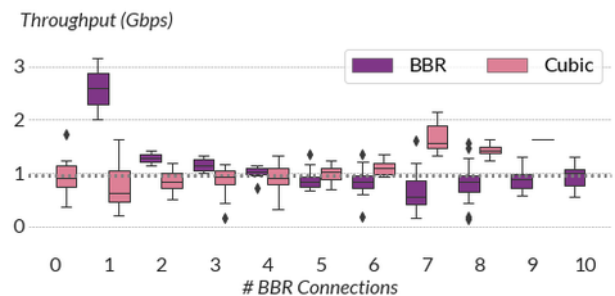


Figure 3: Experiments where 10 TCP connections using Cubic or BBR share a 10 Gb/s link. Throughput is the same if everyone uses either algorithm, but A/B tests suggest that both are improvements.

obtains 50% lower throughput than unpaced traffic (the control) in any A/B test, regardless of allocation. In each A/B test, we observed essentially no reduction in retransmissions for pacing.

Applying usual A/B testing practice to these results might have led us to decide not to deploy pacing. However, if we did deploy pacing, we would be pleasantly surprised to see no impact on throughput and a large decrease in retransmissions. The A/B tests also miss that pacing is good for other traffic: the spillovers from pacing are an increase in throughput and a decrease in retransmissions.

Pacing highlights the importance of estimating TTE when experimenting with networking algorithms. It is not obvious that pacing changes the way connections compete with each other: we expected it would smooth out bursts and cause lower RTT and loss with no impact on throughput. Without careful experiment design, an experimenter could be easily misled into thinking that pacing is not useful, or waste effort chasing a non-existent bug.

3.3 Test 3: Congestion Control Algorithms

There has been extensive study of the fairness of congestion control algorithms (e.g. [5, 15, 16, 23, 43, 44, 56, 71, 81, 82, 84, 85]). A treatment algorithm is often said to be unfair if it gets a larger share of throughput when competing against a control algorithm. In terms of our metrics, this would be if the spillover on control traffic is a decrease in throughput.

An A/B test will not accurately measure the TTE for an unfair algorithm. The treatment algorithm will take throughput away from the control, making the control perform worse than if the treatment were not present. Most widely-used congestion control algorithms are known to be unfair to at least some other algorithms in certain settings. The resulting biases undermine A/B tests on new congestion control algorithms at scale.

As an example, it's been widely reported that BBR is unfair to Cubic in certain situations [16, 43, 44, 71, 84, 85]. This unfairness suggests congestion interference, so we ran simulated A/B tests in our lab. We ran ten long-lived TCP connections, and allocated some fraction of them to BBR and the rest to Cubic. Figure 3 shows our results. If we were interested in deploying BBR in this setting and ran a 10% allocation, we would see a huge improvement in throughput. If instead we were interested in deploying Cubic and ran a 10% allocation, we would also see a huge improvement! But in this setting there is no difference in throughput between a global allocation to either BBR or Cubic.

4 PAIRED LINK EXPERIMENT WITH BITRATE CAPPING

In response to the increased network usage during the beginning of the COVID-19 pandemic, Netflix worked with various governments to reduce load on the Internet, and rolled out a bitrate capping program which reduced video quality [30]. This program capped the video bitrate delivered to clients, while preserving the video resolution based on their subscription plans. It was observed that between March and June 2020, capping the bitrate reduced Netflix traffic in many countries by 25%, and reduced congestion for a number of ISPs.

In this section, we will describe a controlled experiment we ran to accurately measure the effects of bitrate capping. Given that bitrate capping reduced Netflix traffic by 25%, we suspected it would decrease congestion. Our preceding lab studies also led us to suspect that standard A/B tests may give biased results. So our goals with this experiment were to:

- (1) Measure the impact of bitrate capping on network performance and video quality of experience, by estimating TTE and spillover effects.
- (2) Estimate the bias of naïve A/B tests on these measurements, and
- (3) Evaluate whether alternate experiment designs would reduce this bias.

These are challenging goals to accomplish simultaneously. To evaluate the bias of a naïve A/B test and newer experimental designs, we need to measure what happens when all traffic is treated. But if we treat all traffic, we have nothing to compare against! We could run sequential experiments and compare their results, but this makes strong assumptions about how the system behaves over

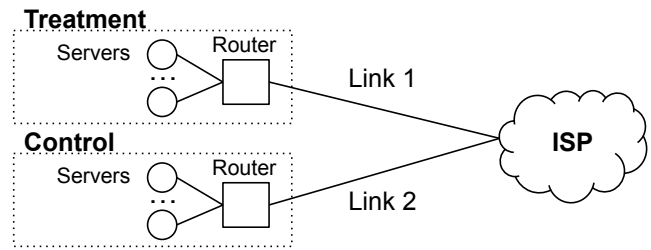


Figure 4: Diagram of the paired link experiment.

time. These would be useful assumptions to make when running alternate experiment designs, and we wanted to use this experiment to evaluate these assumptions.

In this section we describe the experiment we ran to achieve these goals. In Netflix's network, there are a pair of 100 Gb/s peering links to an ISP. The links are reliably congested during peak hours, and are statistically very similar. We treat these two links as "parallel universes," and can compare the outcomes of different experiments to investigate A/B test biases and congestion interference.

Our results are striking and sobering. Bitrate capping reduced congestion at the cost of slightly lower video quality, and improved the performance of uncapped traffic. This was almost completely undetected by naïve A/B tests which underestimated some treatment effects, failed to detect others, and, as we will see, even inferred the wrong *direction* of improvement for certain metrics.

4.1 Paired peering links

Netflix has a location with a pair of identical clusters, replicated for scale and redundancy. Each cluster is identically configured with a router and a number of cache servers. Each router connects to a partner ISP via a 100 Gb/s peering link. This setup is depicted in Figure 4.

During peak viewing hours, demand from users connecting via this ISP increases until eventually a large standing queue builds up on both links. Latency increases, and throughput and video quality decrease. The congestion has a large impact on the quality observed by traffic, and we suspected strong congestion interference between connections sharing the same link.

A priori, we are not guaranteed that the two links will be similar to each other, since the system is optimized to serve video and not to run experiments. The content available on the two clusters is not identical, and different traffic is routed to the servers across each link. To validate statistical similarity between the two links, we collected data on both links during a week-long baseline period, comprising over five million sessions: 50.8% on link 1, and 49.2% on link 2. Netflix collects client- and server-side data on video performance. We looked at 24 important metrics including ones related to network performance (throughput, RTT, etc...) and video QoE (perceptual quality, stability, etc...). For each metric, we used the analysis approach described in Appendix B to compare links 1 and 2. We will discuss the most relevant subset of these metrics.

We obtained the following results, reported as means and 95% confidence intervals. Relative to link 2, link 1 had 5% (0.5%-10%) more overall bytes sent, a 2% (0.1%-3%) higher video stability metric,

and 0.1% (0.03%-0.25%) lower perceptual quality. The largest differences were related to rebuffers. Rebuffers are moments when video playback is interrupted because the client is unable to download a piece of video from the server. Relative to link 2, link 1 had 20% (13-27%) more sessions with rebuffers; there were four additional metrics related to rebuffers that also exhibited similar differences. All other metrics did not have statistically significant differences. Notably, we did not see differences in most metrics we will discuss in our experiment below, including RTT, throughput, video bitrate, cancelled starts, or packet retransmissions.

Traffic on these links is not perfectly balanced, but it is clearly quite similar. Although the pre-existing differences in rebuffers is large, it is important to note that in absolute terms rebuffers are rare. Given the similarity in other metrics, we believe they are caused by some other difference, such as the content served on the two links. Nevertheless, we carefully discuss our experimental findings regarding rebuffers in Section 4.3, where our observations suggest this difference in fact causes us to *underestimate* the extent to which naïve A/B tests are biased.

Being able to run an experiment like this is an extremely unusual situation. Operators work hard to avoid persistent congestion, so it is rare to have a pair of congested peering links. It is even rarer for the traffic to be balanced, and to be able to run separate experiments on each link. Netflix has hundreds of locations and thousands of peering links worldwide, but only *two* were suitable for this experiment.

4.2 Experiment design and analysis

We now describe the experiment we ran. Our goal was to estimate the effects when most traffic was capped, the TTE, and compare this to the results of A/B tests. We also wanted to measure the spillover of capped traffic on uncapped traffic.

To accomplish this, we ran a pair of A/B tests on the two links. On link 1, we allocated 95% of flows to treatment ($p = 0.95$). On link 2, we allocated 5% to treatment. Computing the naïve $\hat{\tau}(p)$ estimator on sessions *within* each link allows us to calculate $\hat{\tau}(0.95)$ and $\hat{\tau}(0.05)$. By comparing the mean of the 95% *treatment sessions on link 1* to the 95% *control sessions on link 2*, we obtain an approximate estimate of TTE. By comparing the mean of the 5% control sessions on link 1 to the 95% control sessions on link 2, we can obtain an approximate estimate of the spillover of capping. With this design, we ran A/B tests simultaneously on the pair of links. The experiment ran for five days, and included about fourteen million video sessions. We analyzed the experiment using techniques described in Appendix B.

In practice, network experiments are usually run in one of two settings. The first is an initial experiment with a relatively low level of initial treatment allocation, corresponding to the 5% A/B test. The second is a long-term holdback test, where almost all traffic is treated. We might naively hope that by treating more traffic, we would reduce congestion interference, and this corresponds to the 95% A/B test.

This experiment may at first appear a bit odd. We are measuring the difference in behavior when *almost* all traffic is capped and *almost* all is uncapped. This is an interesting quantity which tells us a lot about the behavior of bitrate capping during congestion, but it

is only an approximation to TTE. The most straightforward way to estimate TTE in this network would be to cap 100% of sessions on link 1 as treatment, and uncapped 100% of sessions on link 2 as control. We could then compare the means of each group to estimate TTE. However, if we did this, we would have no instances where capped and uncapped traffic shared a link, and we would be unable to compare the results to an A/B test or measure spillover. We could run other experiments other times on the links and compare the results, but we would be making strong assumptions about time invariance. This would require careful experimental design and analysis, and one of our goals here was to *validate* these designs.

Putting it another way: one of our goals is to test the SUTVA assumption, and check whether treatment effects as measured by A/B tests give good predictions of what happens when an algorithm is widely deployed. If SUTVA holds, as in Figure 1a, spillover must be zero, and there must be no difference between the results of the two A/B tests and the approximate TTE we measure. If there is any difference between these quantities in our experiments, SUTVA cannot hold. Knowing that SUTVA does not hold, we would not expect slightly increasing the fraction of capped traffic to fix this problem.

4.3 Results

Our results can be summarized as follows: bitrate capping substantially reduced congestion and improved performance of uncapped traffic, and yet the naïve estimator would have largely failed to detect this.

Figure 5 reports our estimates of treatment effects and 95% confidence intervals for several important video streaming and network metrics. We report the results of 5% and 95% Naïve A/B test results (*i.e.*, $\hat{\tau}(0.05)$ and $\hat{\tau}(0.95)$), as well as our estimate of approximate TTE and our estimate of spillover. The naïve estimators are also wrong about the direction of improvement for minimum RTT and average throughput, and the magnitude of average play delay and video bitrate. The spillover is non-zero for most metrics.

Taking the example of average throughput, the two naïve A/B tests predicted a 5% *decrease* in throughput, which naively suggests that capping increased congestion. However, the TTE tells a very different story: that capping *increased* average throughput by 12%. Spillover shows that capping also benefited other traffic sharing the link: control traffic on the mostly capped link had 16% higher throughput than that on the mostly uncapped link.

These results can be explained by the way bitrate capping reduced congestion. There was significantly less capped traffic, so it took a larger number of users for the link to become congested. Since user demand was the same on both links, congestion started later, ended earlier, and was less severe on the majority-capped link. The naïve estimators were unable to detect this because both capped and uncapped traffic used the same congested link, and therefore saw similar performance.

This becomes clearer if we take a closer look at how the average throughput of sessions changes in Figure 6b, which can be contrasted with how the behavior during the baseline period in Figure 6a. We report the average of all client throughputs during each hour, normalized by the largest hourly throughput. Throughput slowly decreases as overall traffic increases throughout the day, and

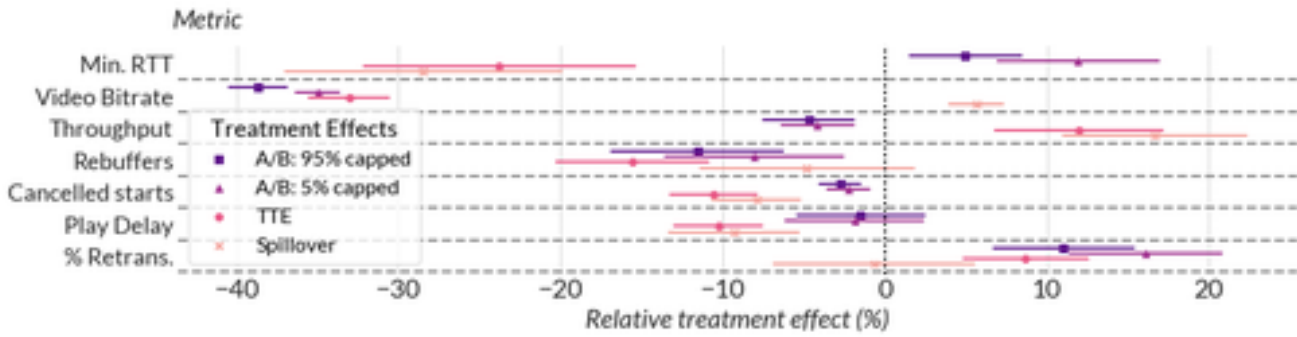
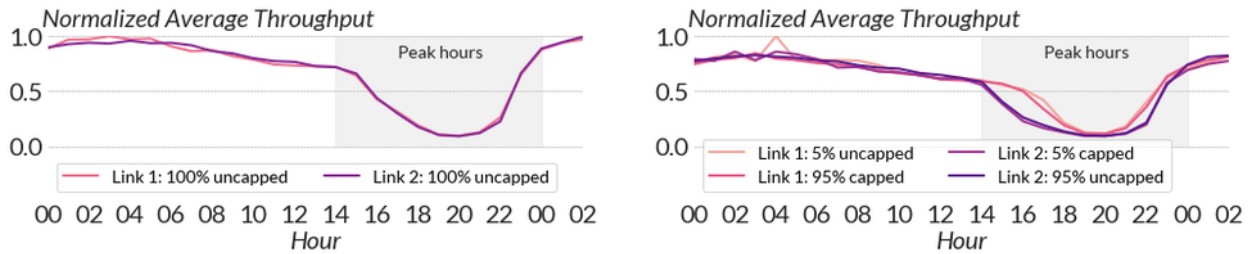


Figure 5: Treatment effects with 95% confidence intervals in our bitrate capping experiments. Each row is a metric of interest, with the naïve A/B Test estimates, and TTE and spillovers as estimated by the paired link experiment.



(a) Average throughput for the Saturday of the baseline test period. (b) Average throughput for the Saturday of the main experiment.

Figure 6: Client-reported average throughput over time in the experiments, normalized to the largest hourly average. During peak hours, the links become congested and throughput decreases. Capping the majority of traffic in (b) causes Link 1 to be less congested and have higher throughput during most of the peak hours.

then suddenly drops when the link becomes congested during peak hours. During the baseline period, there is no difference between throughputs for the two links. During the main experiment, the mostly capped link remains uncongested for longer during peak hours, and has higher throughput before and after the most heavily loaded hours. Despite this difference, the capped and uncapped traffic on the same link have very similar performance.

In Figure 7, we show the four outcomes of throughput in the experiment: for capped and uncapped traffic as a function of allocation percentage. Both A/B tests confidently report that capped traffic reduces throughput relative to uncapped traffic. However by capping the majority of traffic, we improve throughput for all traffic using the link. This leads to an improvement as measured by TTE, and a positive spillover.

If we considered just one of the A/B tests in isolation, we would falsely conclude that capping traffic makes throughput slightly worse. This is our “smoking gun”—the confusion arises because treatment and control interfere with each other via congestion on the link.

We observed similar behavior for round-trip times in the experiment, as shown in Figure 8. During congested hours, large queues build up at the congested link, which causes all packets in a session to be delayed, and leads to a sharp increase in the minimum RTT

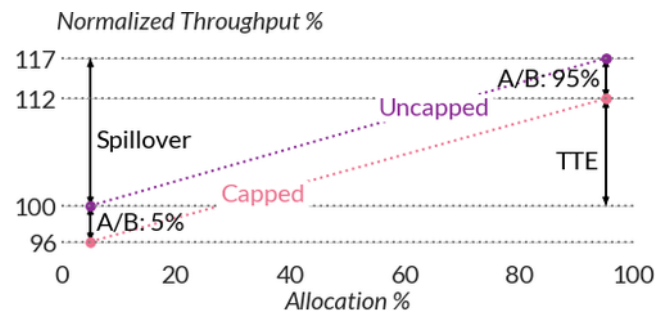


Figure 7: Average values of throughput in the cells in this experiment, with estimands of interest.

observed during each session. However, because bitrate capping delayed the onset of congestion, the majority-capped link (link 1) had empty queues for more time. The total treatment effect was a 24% *improvement* in the minimum RTT for the bitrate-capped sessions. The spillover was positive: capping traffic improved the minimum RTT by 27% for uncapped traffic. Again this was incorrectly estimated by the naïve A/B tests which both reported a 5% and 12% *increase* in minimum RTT.

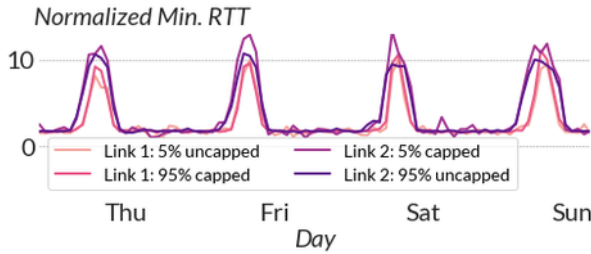


Figure 8: Average of minimum RTT in each connection, normalized to smallest cell value.

We saw similar effects in start play delay, which is the time it takes a video to start playing. This is not surprising: improving throughput and reducing queueing delay should cause videos to load faster. Neither A/B test predicted a significant decrease in start play delay, whereas there was actually a 10% improvement in total play delay effect. The spillover was also positive: capping traffic reduced play delay by 9% for both itself and for uncapped traffic.

We measured a 33% reduction in video bitrate, with positive spillover. Capping the majority of traffic meant that the *uncapped* traffic was able to take up more bandwidth and achieve higher bitrates. It is surprising that despite the spillover, the two A/B tests still give reasonably good estimates of TTE. We believe this is because the majority of the reduction in bitrate comes from the artificial cap, which is applied independently of how other traffic behaves. The spillover is small relative to this effect, but might explain the difference between the 95% treatment effect and TTE.

We observed the total treatment effect for capping was a 10% *increase* in the fraction of sent bytes that were retransmitted. This was driven by a 16% *increase* in the fraction of retransmitted bytes during off-peak hours, and a 20% *decrease* during peak hours as shown in Figure 9. This may seem surprising since bitrate capping reduced congestion, but in fact retransmits did not get worse. Capping reduced the *absolute* number of bytes retransmitted during both during peak and off-peak hours. The apparent increase in the percentage was caused by the absolute number of sent bytes decreasing more than the absolute number of retransmitted bytes. Although odd, Netflix observed similar behavior in a number of ISPs when removing bitrate capping.

Finally, we discuss the impact on rebufferers. Recall from Section 4.2 that we observed a 20% difference in rebufferers between the links from our baseline analysis prior to the experiment. Based on our experiment, we believe bitrate capping had at least some impact on rebufferers: we see a 15% decrease in rebufferers in the A/B tests within each link. We also measured that rebufferers for the mostly capped traffic in link 1 were 18% lower compared to the mostly uncapped traffic in link 2.

Given that rebuffer rates were not identical pre-experiment, we investigated further and measured rebuffer rates for both links during the month after we ran the experiment. We consistently found a difference: link 1 had on average 15% more rebufferers. In 70% of all hours, and in all but one peak hour, link 1 had more rebufferers than link 2. While we are not certain of the underlying reason

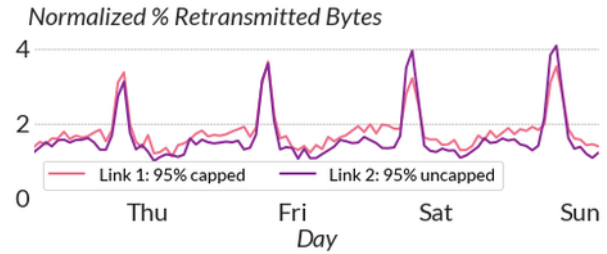


Figure 9: Capping bitrate generally reduced the fraction of retransmitted bytes during congested hours, but caused an increase in uncongested hours.

for the difference, we believe an 18% improvement is probably an *underestimate* of the improvement of rebufferers. If we account for the underlying difference between links 1 and 2, it is closer to a 20%-30% improvement (rather than 15% improvement from the naïve estimate), suggesting congestion interference.

We conclude by highlighting one reason our results may underestimate the amount of congestion interference. As discussed in Appendix B, A/B test analysis usually assumes that sessions from different users are statistically independent of each other. By estimating standard errors only on data aggregated to the hourly level, our analysis effectively makes a nearly worst-case assumption that sessions in the same hour are *perfectly correlated*. This dramatically increases the size of the confidence intervals we report for TTE and spillover.

5 UNBIASED EXPERIMENTS AT SCALE

We care about two different things when evaluating a new algorithm: testing it safely and accurately measuring its performance. We want to experiment safely: if a new algorithm works so poorly that it could cause material harm to the service, we want to detect it quickly and avoid deploying it widely. We also want to be accurate: the goal of a new algorithm is usually to improve some metric, and we need to accurately evaluate whether it succeeded.

A/B tests are used today with the assumption that they are both safe and accurate. If the SUTVA assumption held, we can accurately estimate performance by running an A/B test on a very small fraction of users. This allows us to predict the performance of an algorithm at scale, without broadly deploying a harmful algorithm.

But in the worst case, congestion interference means that an A/B test is neither safe nor accurate. An algorithm which performs well in an A/B test might cause significant harm when it is deployed globally. But if an algorithm has marginal A/B test results and we do not deploy it globally, we may miss out on extremely effective algorithms.

This is a fundamental tradeoff with congestion interference, and what makes it so difficult to work with in practice. If we want to get a completely unbiased estimate of TTE, we need to allocate 100% of traffic to a treatment. But for safety reasons we would never allocate 100% of traffic to an untested or poorly performing algorithm.

In this section, we provide some guidance on how to run experiments in practice. We will not be able to completely resolve this

tradeoff, but we will describe two ways of measuring congestion interference despite it.

Naïve A/B tests are biased in congested networks because of the combination of the A/B experiment design itself, and the flawed causal inference used when interpreting the results of that design. We will propose modifications to the A/B experiment design, and describe the improved causal inference that these modifications allow. First, we propose slightly modifying existing deployment practices to look for congestion interference. This is easy to do and helps build intuition around when congestion interference exists, at the cost of time-related bias and rejecting effective algorithms. To counter this, we also propose running small-scale, targeted switch-back experiments to measure how a new algorithm behaves in a specific network.

5.1 Measure deployed algorithms with event studies

When deploying an algorithm, it is important to get an accurate estimate of TTE. Optimistically, an algorithm might perform better at scale than it did in small-scale evaluations. Perhaps when an algorithm is run by a larger fraction of traffic, it even further reduces congestion and improves performance than it did in small-scale experiments. Accurately quantifying the improvement is important to understanding its behavior and giving the team working on the algorithm the credit they deserve.

Pessimistically, a new algorithm might perform worse at scale than in small-scale evaluations. This might be a sign of some bug or unexpected behavior in the algorithm, and might suggest it increases congestion or interferes with other traffic on the internet. These are things that are important to know about, so they can be addressed.

Primarily for safety reasons, engineers have developed sophisticated techniques for deploying new algorithms. Engineers gradually deploy changes by slowly increasing the allocation fraction. They continually monitor the system, and stop the deployment if performance degrades.

While engineers typically use gradual deployments to safeguard against failure, they could also be used to conveniently measure the performance of a new algorithm and look for congestion interference. A gradual deployment is effectively a series of A/B tests with treatment allocations ranging from 0% to 100%. At each allocation (p_1, p_2, \dots) we can observe the outcomes for treatment and control. This gives us points on the graph of Figure 1, and we can use these values to estimate the average treatment effect $\tau(p_i)$, the spillover $s(p_i)$, and a *partial* treatment effect $\rho(p_i) = \mu_T(p_i) - \mu_C(0)$. Once the deployment is finished, we can compare 100% allocation to 0% allocation and estimate TTE. If there is no interference, for all allocations i and j , the average treatment effects are the same $\tau(p_i) = \tau(p_j)$, the partial treatment effects are the same as the average treatment effects $\rho(p_i) = \tau(p_i)$, and there is no spillover $s(p_i) = 0$. We can use statistical tests to check each of these relationships. If they do not hold, it could be a sign of congestion interference.

This is a type of observational design called an *event study* or an interrupted time series [54, Ch. 11]. In an event study, we introduce some change, and compare the state of the system before

and after. This can be contrasted with a naïve A/B test, where we simultaneously compare units with and without the change. In the gradual deployment setting, the change is the increase of treatment allocation from p_i to p_{i+1} .

A major flaw with event studies is that it can be difficult to attribute observed behavior to a particular change. This is especially true because of seasonality: holidays, weekends, and political events all tend to have different traffic patterns than other times. Other teams or organizations regularly make changes and deploy software which can affect similar metrics. In the bitrate capping example, we had data from before and after deployment, but chose to run a more controlled experiment to rule out the possibility of other causes for the behavior we observed.

Another flaw is that this process works well for safely deploying new algorithms, but it is heavily biased towards rejecting new algorithms. As an example, suppose we were testing a new algorithm which behaved like the pacing lab experiment in Section 3.2. In a small allocation A/B test, this algorithm would look worse: throughput would be down and loss would be unaffected. Seeing this, we might invest our time in other, more promising algorithms. We could slightly increase the size of the allocation to look for interference, but throughput increased quite slowly with allocation size. Even if we were able to detect this interference, it would look small. At this point, we might stop the deployment before the algorithm is able to clearly improve performance.

Despite these flaws, event studies are quick and easy ways to get estimates of TTE and spillovers. Large organizations continually deploy changes. When a deployment happens, it is easy to look at the already-collected metrics and use these metrics to estimate TTE and spillovers. Doing so will help build intuition around which algorithms could be affected by congestion interference.

5.2 Measure algorithms in development with targeted switchbacks

Running an event study when deploying a new algorithm is a good way to measure congestion interference and build intuition, but it is a bad way to experiment with new algorithms. We do not want to deploy marginal algorithms to all traffic, and so we may not invest in algorithms that perform poorly in an A/B test. We may miss out on algorithms that have very different effects when widely deployed, like bitrate capping, pacing, or changing the number of TCP connections.

Because of this, we recommend running small targeted experiments in addition to small A/B tests. A targeted experiment allocates a large fraction of traffic within a specific network. The network needs to be structured in such a way that the allocated traffic does not interact with non-allocated traffic. In the paired link experiment in Section 4, we targeted an experiment to two congested links. Using the results from the large fraction allocation, we can get a good estimate of TTE and spillover in this network.

Targeting an experiment allows us to estimate TTE and spillover within a network, without needing to run an algorithm on 100% of traffic globally. It is standard practice in online platforms [52, 68]. While we estimate TTE and spillover for a specific network instead of globally, this helps give additional context to A/B test results and improves our understanding of how a new algorithm behaves.

When running these targeted experiments, we recommend using *switchback designs*. A switchback design divides time into intervals; a given interval is randomly assigned to be either treatment or control. In a treatment interval, we treat almost all of the traffic with the new algorithm. In a control interval, almost all traffic runs the old algorithm.

At a high level, switchback experiments are analyzed by comparing the treatment and control intervals. While we could do 100% allocations in these intervals to get a good TTE estimate, we recommend a smaller allocation (e.g. 90-99%) as in the paired link experiment. Doing so allows us to additionally estimate spillover and the bias of A/B tests, which gives valuable insight into algorithm behavior. The allocation size should be large enough to give statistically significant results, and can be determined by a power calculation.

Like event studies, switchback experiments rely on the change between treatment and control intervals being due to the treatment. However, the assumption is weaker: instead of needing no other events to impact the outcome, a switchback requires that another event does not line up with the treatment intervals.

A switchback experiment can also be vulnerable to carryover effects [14, 32]. The presence of the treatment algorithm can influence the initial conditions of the control algorithm and vice versa. This can cause bias: imagine if we were to switch sessions between one and two parallel connections. Until all sessions that used two parallel connections had completed, the sessions using one would have lower throughput than necessary. If the system reacts poorly to switching between treatment and control, this could also cause problems.

Carryover effects can be mitigated with sufficiently long intervals. However, typically switchback experiments make the worst-case assumption that all sessions in an interval are dependent (see Appendix B for more details), which essentially means that each interval gives us one data point. Increasing the length of intervals effectively lowers the sample size of the experiment. For networking algorithms, we believe a switch interval of one day is a reasonably conservative place to start. Depending on the setting and the algorithm, it may be appropriate to use a shorter interval on the order of hours or minutes.

5.3 Evaluating alternate designs

Our paired link experiment gives us the results of simultaneous, comparable experiments. We previously analyzed that data to estimate TTE and spillovers. We now use it to evaluate event studies and switchback designs, and show that these designs also accurately estimate TTE.

Having two simultaneous experiments allows us to ask what *would* have happened if we ran only one experiment at a time. Our experiment in Section 4 ran from Wednesday through Sunday, giving us five possible days of data. We can emulate an event study by using data from the 5% link for a few days and then switching to data from the 95% link, representing a deployment of bitrate capping to 95% of traffic. We can emulate a switchback experiment by switching between treatment days and control days more frequently.

We first used baseline data to calibrate a switchback experiment. We ran an A/A test [54, Ch. 19] on the paired links in the week following our main experiment: we applied the control to both links and looked for underlying differences. Using the data from the A/A test, we checked that there would have been no false positives with any switchback design. This increases confidence that there isn't a reliable difference between days in a way that would bias the experiment, and we would recommend doing this in most cases.

We also used baseline data to calibrate an event study. We observed that there were false positives in the majority of metrics with any event study in this experiment. We believe this is because weekends tend to have different traffic patterns than weekdays, and an event study must either treat all the weekend days or all the weekdays together. This is an advantage of using a switchback design.

For the event study, we switched to 95% bitrate capping between Thursday and Friday as shown in Figure 11. For the switchback, we alternated between treatment and control, and randomly started with treatment. This assignment is shown in Figure 12. All other ways of assigning treatment to days yielded similar results, provided at least one day was in treatment and at least one day was in control.

Figure 12 shows the average throughput for this example switchback design, which can then be compared with the throughput in the paired link experiment in Figure 6. Note that because we are switching between experiments, the clear difference in throughput in the paired link time series is much harder to see in the switchback. This highlights the power of running statistical analyses on switchback data.

Our goal with this approach was to use the clean results from our paired link experiment to demonstrate the power of switchback experiments and event studies. If we had actually run these experiments, the results may have been slightly different. For instance, traffic from both links likely shares some bottlenecks in the provider network during offpeak hours, so it is possible that our results during offpeak hours are biased by congestion interference. However the congestion interference we detect is largely because of the behavior during congested hours on isolated congested links.

5.4 Results

The analysis approach for these experiments is identical to the paired link experiment, with the caveat that we only use the subset of the data corresponding to each experiment. We describe the details in Appendix B.

Figure 10 shows the values of TTE estimated by the switchback experiment, event study, and paired link experiment. Both alternate experiments give reasonably good estimates of TTE. The switchback experiment results are very close, and the confidence intervals for its estimates include every TTE from the paired link experiment. It has larger confidence intervals because it includes half as much data. We expect that running the experiment for longer would have reduced the size of the confidence intervals.

The event study gives reasonably accurate estimates of TTE for most metrics, but is biased for throughput, cancelled starts, and % retransmitted bytes. As we observed in analyzing the baseline data, we believe this is because of seasonality issues: weekends tend to have different behavior than weekdays, and so it is more difficult to

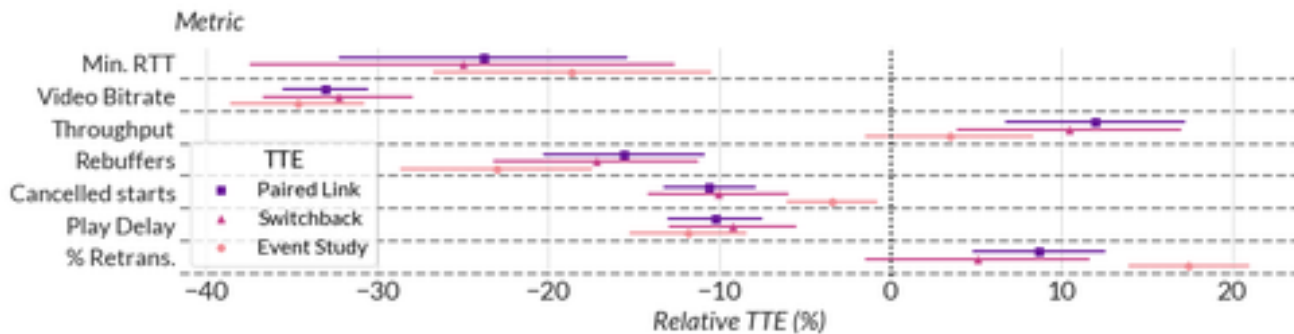


Figure 10: TTE as estimated by the paired link experiment, a switchback experiment, and an event study.

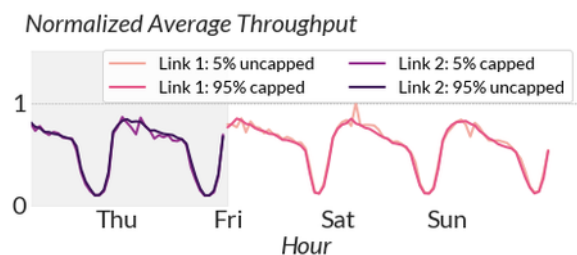


Figure 11: Throughput in a bitrate capping event study. Between Thurs. and Fri., we apply 95% bitrate capping.

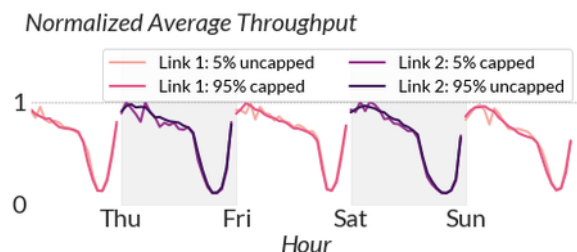


Figure 12: Average throughput over time in a bitrate capping switchback experiment. 95% of traffic is capped on the first and third and fifth day.

attribute the change to the treatment. This is one of the advantages of switchback experiments: randomly choosing intervals over many days helps avoid certain seasonality effects. Despite this, given that event studies are so easy to incorporate into existing workflows, we still recommend cautiously using them to estimate TTE and spillovers when deploying new algorithms.

6 RELATED WORK

A/B tests are heavily used in industry research. There recently have been a number of published A/B tests comparing congestion control algorithms, including BBR [17–19, 46, 72], COPA [63], and Swift

[55]. There have also been many other published A/B tests for other networking algorithms. These include work on initial congestion windows [25], TCP’s loss recovery [29], PRR [24], QUIC [49, 58], failure recovery [57], and ABR algorithms [42, 60, 87]. We do not know how congestion interference affected these results.

We are aware of a few published results that include event studies: Dropbox and Verizon both used them to evaluate BBRv1 [46, 72], and Google reported one for Timely in [61]. In Section 5, we show how to design and analyze these event studies to measure TTE and spillover, and describe how switchback experiments give more reliable results.

Experiments on router performance, especially those related to buffer sizing [11, 12, 74], naturally must treat all traffic using the router. Because of this, they tend to have good estimates of total treatment effects.

Recent studies of social network and marketplace platforms have led to improved understanding of causal inference under interference (e.g., [4, 6, 9, 13, 59]), both through novel experimental design (e.g., [7, 14, 20, 32, 40, 47, 68, 83]) and improved inferential methodology (e.g., [6, 9, 10, 77]). We believe our work is the first to show that these issues affect networking experiments and bias their results at scale.

Switchback designs found recent favor as an approach to testing matching and dispatch policies in ridesharing and food delivery platforms, though they have also been used in applications as varied as agriculture [14, 20, 52, 64, 65]. We are unaware of any prior usage of switchbacks in networking.

We have heard some folklore predictions from the networking community that these sort of issues may exist. The only citeable version of this we know of is in [80].

Finally, our work is informed by the long line of work on fairness in networking. Unfairness between Cubic and BBR, which we describe in Section 3, was previously reported by [16, 43, 44, 71, 81, 82, 84, 85]. Unfairness between parallel connections was first observed by [8]. Unfairness between paced and unpaced Reno flows was shown by [2, 86]. Fairness work is about how algorithms *ought* to share resources, and usually shows that algorithms are unfair in simulations or in a lab [5, 15, 16, 23, 43, 44, 50, 56, 71, 81, 82, 85]. Our work does not address how algorithms *should* share resources, but rather how to avoid experimental bias when they *do*. One way

of interpreting our work is as a way to measure unfairness between treatment and control at scale, in production networks.

7 CONCLUSION

Congestion interference biases the results of networking A/B tests at scale, and it is our responsibility as a community to be aware of this phenomenon. Our results suggest that we should be skeptical when interpreting the results of naïve A/B tests, and consider whether alternate experiment designs should be used instead.

As discussed in Section 5, experimenters can make small changes to existing deployment processes to begin to measure congestion interference, and use targeted switchbacks to further improve these measurements. We should be especially wary of interference when an algorithm changes traffic volumes, tries to control congestion, or is similar to algorithms discussed in the past fairness research in Section 6.

We would love to see more work in networking evaluated with congestion interference in mind, either with published switchback experiments, or at least event studies run during a gradual deployment. This is especially true for high consequence proposals, such as new internet standards.

On the research side, there is much more work to be done on evaluating algorithms at scale in congested networks. We encourage further studies to measure bias, in different networks and with different algorithms. We think it would be valuable to design new experiments and analyses specifically for congested networks. The bias of naïve A/B tests is both a cautionary tale and a significant opportunity for innovation. The internet surely works better thanks to A/B tests of algorithms run in congested networks. We hope that new algorithms tested with better experiments will help improve it even further.

8 ACKNOWLEDGEMENTS

Thank you to Guillaume Basse and Matthew Pawlicki for the very helpful discussions. Thanks also to Neil Perry, Sundararajan Renganathan, Renata Teixeira, the anonymous reviewers, and our shepherd for all their feedback on the paper.

REFERENCES

- [1] Alberto Abadie, Joshua Angrist, and Guido Imbens. 2002. Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica* 70, 1 (Jan. 2002), 27.
- [2] A. Aggarwal, S. Savage, and T. Anderson. 2000. Understanding the Performance of TCP Pacing. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, Vol. 3. IEEE, Tel Aviv, Israel, 1157–1165. <https://doi.org/10.1109/INFCOM.2000.832483>
- [3] Julia Alexander. 2020. Amazon and Apple Are Reducing Streaming Quality to Lessen Broadband Strain in Europe. (March 2020). <https://www.theverge.com/2020/3/20/21188072/amazon-prime-video-reduce-stream-quality-broadband-netflix-youtube-coronavirus>
- [4] Peter M. Aronow and Cyrus Samii. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* 11, 4 (12 2017), 1912–1947. <https://doi.org/10.1214/16-AOAS1005>
- [5] Rukshani Athapathu, Ranysha Ware, Aditya Abraham Philip, Srinivasan Seshan, and Justine Sherry. 2020. Prudentia: Measuring Congestion Control Harm on the Internet. 2. <http://www.justinsherry.com/papers/athapathu-n2women20.pdf>
- [6] Susan Athey, Dean Eckles, and Guido W. Imbens. 2018. Exact p-Values for Network Interference. *J. Amer. Statist. Assoc.* 113, 521 (2018), 230–240. <https://doi.org/10.1080/01621459.2016.1241178> arXiv:<https://doi.org/10.1080/01621459.2016.1241178>
- [7] Pat Bajari, Brian Burdick, Guido Imbens, James McQueen, Thomas Richardson, and Ido Rosen. 2019. Multiple Randomization Designs for Interference. (2019). <https://assets.amazon.science/c1/94/0d6431bf46f7978295d245dd6e06/double-randomized-online-experiments.pdf>
- [8] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, M. Stemm, and R. H. Katz. 1998. TCP Behavior of a Busy Internet Server: Analysis and Improvements. In *Proceedings. IEEE INFOCOM '98, the Conference on Computer Communications. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Gateway to the 21st Century (Cat. No.98, Vol. 1. 252–262 vol.1.* <https://doi.org/10.1109/INFCOM.1998.659661>
- [9] Guillaume Basse, Avi Feller, and Panagiotis Toulis. 2019. Randomization tests of causal effects under interference. *Biometrika* 106, 2 (02 2019), 487–494. <https://doi.org/10.1093/biomet/asy072> arXiv:<https://academic.oup.com/biomet/article-pdf/106/2/487/28575447/asy072.pdf>
- [10] Guillaume W. Basse, Hossein Azari Soufiani, and Diane Lambert. 2016. Randomization and The Pernicious Effects of Limited Budgets on Auction Experiments. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9–11, 2016 (JMLR Workshop and Conference Proceedings)*, Arthur Gretton and Christian C. Robert (Eds.), Vol. 51. JMLR.org, 1412–1420. <http://proceedings.mlr.press/v51/basse16b.html>
- [11] Neda Beheshti, Yashar Ganjali, Monia Ghobadi, Nick McKeown, and Geoff Salmon. 2008. Experimental Study of Router Buffer Sizing. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement (IMC '08)*. ACM, New York, NY, USA, 197–210. <https://doi.org/10.1145/1452520.1452545>
- [12] Neda Beheshti, Petr Lapukhov, and Yashar Ganjali. 2019. Buffer Sizing Experiments at Facebook. In *Proceedings of the 2019 Workshop on Buffer Sizing*. ACM, Palo Alto CA USA, 1–6. <https://doi.org/10.1145/3375235.3375244>
- [13] Thomas Blake and Dominic Coey. 2014. Why Marketplace Experimentation is Harder than It Seems: The Role of Test-Control Interference. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation (EC '14)*. Association for Computing Machinery, New York, NY, USA, 567–582. <https://doi.org/10.1145/2600057.2602837>
- [14] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. 2021. Design and Analysis of Switchback Experiments. *arXiv:2009.00148 [stat]* (Jan. 2021). arXiv:<http://arxiv.org/abs/2009.00148>
- [15] Bob Briscoe. 2007. Flow Rate Fairness: Dismantling a Religion. *ACM SIGCOMM Computer Communication Review* 37, 2 (March 2007), 63–74. <https://doi.org/10.1145/1232919.1232926>
- [16] Yi Cao, Arpit Jain, Kriti Sharma, Aruna Balasubramanian, and Anshul Gandhi. 2019. When to Use and When Not to Use BBR: An Empirical Analysis and Evaluation Study. In *Proceedings of the Internet Measurement Conference*. ACM, Amsterdam Netherlands, 130–136. <https://doi.org/10.1145/3355369.3355579>
- [17] Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2017. BBR: Congestion-Based Congestion Control. *Commun. ACM* 60, 2 (Jan. 2017), 58–66. <https://doi.org/10.1145/3009824>
- [18] Neal Cardwell, Yuchung Cheng, Soheil Hassas Yeganeh, Priyaranjan Jha, Yousuk Seung, Kevin Yang, Ian Swett, Victor Vasiliev, Bin Wu, Luke Hsiao, Matt Mathis, and Van Jacobson. 2019. BBR v2: A Model-Based Congestion Control Performance Optimizations. (Nov. 2019). <https://www.ietf.org/proceedings/106/slides/slides-106-iccg-update-on-bbrv2-00>
- [19] Erik Carlsson and Eirini Kakogianni. 2018. Smoother Streaming with BBR. (Aug. 2018). <https://engineering.atspotify.com/2018/08/31/smooth-streaming-with-bbr/>
- [20] Nicholas Chamandy. 2016. Experimentation in a Ridesharing Marketplace. (Dec 2016). <https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-f75a9c4fcf01>
- [21] Bruno Crépon, Esther Dufló, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment *. *The Quarterly Journal of Economics* 128, 2 (May 2013), 531–580. <https://doi.org/10.1093/qje/qjt001>
- [22] Nikos Diamantopoulos, Jeffrey Wong, David Issa Mattos, Ilias Gerostathopoulos, Matthew Wardrop, Tobias Mao, and Colin McFarland. 2019. Engineering for a Science-Centric Experimentation Platform. *arXiv:1910.03878 [cs]* (Oct. 2019). arXiv:[cs/1910.03878](https://arxiv.org/abs/1910.03878) <http://arxiv.org/abs/1910.03878>
- [23] Mo Dong, Tong Meng, Doron Zarchy, Engin Arslan, Yossi Gilad, P. Brighten Godfrey, and Michael Schapira. 2018. PCC Vivace: Online-Learning Congestion Control. In *NSDI*. 15.
- [24] Nandita Dukkipati, Matt Mathis, Yuchung Cheng, and Monia Ghobadi. 2011. Proportional Rate Reduction for TCP. In *Internet Measurement Conference*. 15.
- [25] Nandita Dukkipati, Tiziana Refice, Yuchung Cheng, Jerry Chu, Tom Herbert, Amit Agarwal, Arvind Jain, and Natalia Sutin. 2010. An Argument for Increasing TCP's Initial Congestion Window. *ACM SIGCOMM Computer Communication Review* 40, 3 (June 2010), 26–33. <https://doi.org/10.1145/1823844.1823848>
- [26] Eric Dumazet. 2013. Pkt_sched: Fq: Fair Queue Packet Scheduler [LWN.Net]. (Aug. 2013). <https://lwn.net/Articles/564825/>
- [27] Eric Dumazet. 2013. Tcp: TSO Packets Automatic Sizing [LWN.Net]. (Aug. 2013). <https://lwn.net/Articles/564979/>
- [28] Dean Eckles, Brian Karrer, and Johan Ugander. 2016. Design and Analysis of Experiments in Networks: Reducing Bias from Interference. *Journal of Causal Inference* 5, 1 (Feb. 2016). <https://doi.org/10.1515/jci-2015-0021>

- [29] Tobias Flach, Nandita Dukkkipati, Andreas Terzis, Barath Raghavan, Neal Cardwell, Yuchung Cheng, Ankur Jain, Shuai Hao, Ethan Katz-Bassett, and Ramesh Govindan. 2013. Reducing Web Latency: The Virtue of Gentle Aggression. In *SIGCOMM*. 12.
- [30] Ken Florance. 2020. Reducing Netflix Traffic Where It's Needed While Maintaining the Member Experience. (March 2020). <https://about.netflix.com/en/news/reducing-netflix-traffic-where-its-needed>
- [31] Andrew Gelman and Jennifer Hill. 2006. Causal Inference Using Regression on the Treatment Variable. In *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, Cambridge, 167–198. <https://doi.org/10.1017/CBO9780511790942.012>
- [32] Peter Glynn, Ramesh Johari, and Mohammad Rasouli. 2020. Adaptive experimental design with temporal interference: A maximum likelihood approach. *arXiv preprint arXiv:2006.05591* (2020).
- [33] Hadas Gold. 2020. Netflix and YouTube Are Slowing down in Europe to Keep the Internet from Breaking. (March 2020). <https://www.cnn.com/2020/03/19/tech/netflix-internet-overload-eu/index.html>
- [34] Nirmal Govind. 2018. A/B Testing and Beyond: Improving the Netflix Streaming Experience with Experimentation and Data... (June 2018). <https://netflixtechblog.com/a-b-testing-and-beyond-improving-the-netflix-streaming-experience-with-experimentation-and-data-5b0ae9295bdf>
- [35] Ilya Grigorik. 2013. HTTP: HTTP/1.X - High Performance Browser Networking (O'Reilly). (2013). <https://hpbn.co/http1x/#using-multiple-tcp-connections>
- [36] Ilya Grigorik and Surma. 2019. Introduction to HTTP/2 | Web Fundamentals. (Sept. 2019). https://developers.google.com/web/fundamentals/performance/http2#request_and_response_multiplexing
- [37] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network A/B Testing: From Sampling to Estimation. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Florence Italy, 399–409. <https://doi.org/10.1145/2736277.2741081>
- [38] M. E. Halloran and C. J. Struchiner. 1995. Causal Inference in Infectious Diseases. *Epidemiology (Cambridge, Mass.)* 6, 2 (March 1995), 142–151. <https://doi.org/10.1097/0001648-199503000-00010>
- [39] David Holtz, Ruben Lobel, Inessa Liskovich, and Sinan Aral. 2020. Reducing Interference Bias in Online Marketplace Pricing Experiments. *arXiv:2004.12489 [econ, stat]* (April 2020). <http://arxiv.org/abs/2004.12489>
- [40] David Holtz, Ruben Lobel, Inessa Liskovich, and Sinan Aral. 2020. Reducing Interference Bias in Online Marketplace Pricing Experiments. (2020). [arXiv:stat.ME/2004.12489](http://arxiv.org/abs/2004.12489)
- [41] Guanglei Hong and Stephen W. Raudenbush. 2006. Evaluating Kindergarten Retention Policy. *J. Amer. Statist. Assoc.* 101, 475 (Sept. 2006), 901–910. <https://doi.org/10.1198/016214506000000447>
- [42] Te-Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. 2014. A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service. In *Proceedings of the 2014 ACM Conference on SIGCOMM*. ACM, Chicago Illinois USA, 187–198. <https://doi.org/10.1145/2619239.2626296>
- [43] Per Hurtig, Habtegebireil Haile, Karl-Johan Grinnemo, Anna Brunstrom, Eneko Atxutegi, Fidel Liberal, and Ake Arvidsson. 2018. Impact of TCP BBR on CUBIC Traffic: A Mixed Workload Evaluation. In *2018 30th International Teletraffic Congress (ITC 30)*. IEEE, Vienna, 218–226. <https://doi.org/10.1109/ITC30.2018.00040>
- [44] Geoff Huston. 2018. TCP and BBR. (May 2018). <https://ripe76.ripe.net/presentations/10-2018-05-15-bbr.pdf>
- [45] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, USA.
- [46] Alexey Ivanov. 2020. Evaluating BBRv2 on the Dropbox Edge Network. *arXiv:2008.07699 [cs]* (Aug. 2020). [arXiv:cs/2008.07699](http://arxiv.org/abs/2008.07699)
- [47] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Weintraub. 2020. Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670* (2020).
- [48] Matt Joras and Yang Chi. 2020. How Facebook Is Bringing QUIC to Billions. (Oct. 2020). <https://engineering.fb.com/2020/10/21/networking-traffic/how-facebook-is-bringing-quick-to-billions/>
- [49] Matt Joras and Yang Chi. 2020. How Facebook Is Bringing QUIC to Billions. (Oct. 2020). <https://engineering.fb.com/2020/10/21/networking-traffic/how-facebook-is-bringing-quick-to-billions/>
- [50] Arash Molavi Kakhki, Samuel Jero, David Choffnes, Cristina Nita-Rotaru, and Alan Mislove. 2017. Taking a Long Look at QUIC: An Approach for Rigorous Evaluation of Rapidly Evolving Transport Protocols. In *Proceedings of the 2017 Internet Measurement Conference*. ACM, London United Kingdom, 290–303. <https://doi.org/10.1145/3131365.3131368>
- [51] Brian Karrer, Liang Shi, Monica Bhole, Matt Goldman, Tyrone Palmer, Charlie Gelman, Mikael Konutgan, and Feng Sun. 2020. Network Experimentation at Scale. *arXiv:2012.08591 [cs, stat]* (Dec. 2020). [arXiv:cs, stat/2012.08591](http://arxiv.org/abs/2012.08591)
- [52] David Kastelman and Raghav Ramesh. 2018. Switchback Tests and Randomized Experimentation Under Network Effects at DoorDash. (Feb. 2018). <https://medium.com/@DoorDash/switchback-tests-and-randomized-experimentation-under-network-effects-at-doordash-f1d938ab7c2a>
- [53] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Chicago Illinois USA, 1168–1176. <https://doi.org/10.1145/2487575.2488217>
- [54] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (first ed.). Cambridge University Press. <https://doi.org/10.1017/9781108653985>
- [55] Gautam Kumar, Nandita Dukkkipati, Keon Jang, Hassan M. G. Wassel, Xian Wu, Behnam Montazeri, Yaogong Wang, Kevin Springborn, Christopher Alfeld, Michael Ryan, David Wetherall, and Amin Vahdat. 2020. Swift: Delay Is Simple and Effective for Congestion Control in the Datascenter. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*. ACM, Virtual Event USA, 514–528. <https://doi.org/10.1145/3387514.3406591>
- [56] Ike Kunze, Jan Ruth, and Oliver Hohlfeld. 2020. Congestion Control in the Wild—Investigating Content Provider Fairness. *IEEE Transactions on Network and Service Management* 17, 2 (June 2020), 1224–1238. <https://doi.org/10.1109/TNSM.2019.2962607>
- [57] Raul Landa, Lorenzo Saino, Lennert Buytenhek, and João Taveira Araújo. 2021. Staying Alive: Connection Path Reselection at the Edge. In *NSDI 2021*. 20.
- [58] Adam Langley, Alistair Ridloch, Alyssa Wilk, Antonio Vicente, Charles Krasick, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, Jeff Bailey, Jeremy Dorfman, Jim Roskind, Joanna Kulik, Patrik Westin, Raman Tenneti, Robbie Shade, Ryan Hamilton, Victor Vasilev, Wan-Teh Chang, and Zhongyi Shi. 2017. The QUIC Transport Protocol: Design and Internet-Scale Deployment. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 183–196. <https://doi.org/10.1145/3098822.3098842>
- [59] Charles F. Manski. 2013. Identification of treatment response with social interactions. *The Econometrics Journal* 16, 1 (2013), S1–S23. <https://doi.org/10.1111/j.1368-423X.2012.00368.x> [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1368-423X.2012.00368.x](https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1368-423X.2012.00368.x)
- [60] Hongzi Mao, Shannon Chen, Drew Dimmery, Shaun Singh, Drew Blaisdell, Yuandong Tian, Mohammad Alizadeh, and Eytan Bakshy. 2020. Real-World Video Adaptation with Reinforcement Learning. *arXiv:2008.12858 [cs]* (Aug. 2020). [arXiv:cs/2008.12858](http://arxiv.org/abs/2008.12858)
- [61] Radhika Mittal, Vinh The Lam, Nandita Dukkkipati, Emily Blem, Hassan Wassel, Monia Ghobadi, Amin Vahdat, Yaogong Wang, David Wetherall, and David Zats. 2015. TIMELY: RTT-Based Congestion Control for the Datascenter. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. ACM, London United Kingdom, 537–550. <https://doi.org/10.1145/2785956.2787510>
- [62] Whitney K. Newey and Kenneth D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55, 3 (1987), 703–708. <https://doi.org/10.2307/1913610>
- [63] Garg Nitin. 2019. COPA Congestion Control for Video Performance. (Nov. 2019). <https://engineering.fb.com/2019/11/17/video-engineering/copa/>
- [64] Samuel D. Oman and Esther Seiden. 1988. Switch-Back Designs. *Biometrika* 75, 1 (March 1988), 81–89. <https://doi.org/10.1093/biomet/75.1.81>
- [65] James Robins. 1986. A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling* 7, 9–12 (Jan. 1986), 1393–1512. [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6)
- [66] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
- [67] Ahmed Saeed, Nandita Dukkkipati, Vytautas Valancius, Vinh The Lam, Carlo Contavalli, and Amin Vahdat. 2017. Carousel: Scalable Traffic Shaping at End Hosts. In *The Conference of the ACM Special Interest Group*. ACM Press, 404–417. <https://doi.org/10.1145/3098822.3098852>
- [68] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M. Airoldi. 2017. Detecting Network Effects: Randomizing Over Randomized Experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Halifax NS Canada, 1027–1035. <https://doi.org/10.1145/3097983.3098192>
- [69] Robert Sayre. 2008. Change Max-Persistent-Connections-per-Server to 6. (March 2008). https://bugzilla.mozilla.org/show_bug.cgi?id=423377
- [70] Nate Schloss and Ben Maurer. 2017. This Browser Tweak Saved 60% of Requests to Facebook. (Jan. 2017). <https://engineering.fb.com/2017/01/26/web/this-browser-tweak-saved-60-of-requests-to-facebook/>
- [71] Dominik Scholz, Benedikt Jaeger, Lukas Schwaighofer, Daniel Raumer, Fabien Geyer, and Georg Carle. 2018. Towards a Deeper Understanding of TCP BBR Congestion Control. In *2018 IFIP Networking Conference (IFIP Networking) and Workshops*. IEEE, Zurich, Switzerland, 1–9. <https://doi.org/10.23919/IFIPNetworking.2018.8696830>
- [72] Anant Shah. 2019. BBR Evaluation at a Large CDN. (Nov. 2019). <https://blog.apnic.net/2019/11/01/bbr-evaluation-at-a-large-cdn/>

- [73] Steve Souders. 2008. Roundup on Parallel Connections. (March 2008). <https://www.stevesouders.com/blog/2008/03/20/roundup-on-parallel-connections/>
- [74] Bruce Spang, Brady Walsh, Te-Yuan Huang, Tom Rusnock, Joe Lawrence, and Nick McKeown. 2019. Buffer Sizing and Video QoE Measurements at Netflix. In *Proceedings of the 2019 Workshop on Buffer Sizing*. ACM, Palo Alto CA USA. <https://doi.org/10.1145/3375235.3375241>
- [75] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* (1990), 465–472.
- [76] Diane Tang, Ashish Agarwal, Deirdre O’Brien, and Mike Meyer. 2010. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. In *KDD’10*. 10.
- [77] Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. 2012. On causal inference in the presence of interference. *Statistical Methods in Medical Research* 21 (2012), 55 – 75.
- [78] Martin Tingley. 2018. Streaming Video Experimentation at Netflix: Visualizing Practical and Statistical Significance. (Sept. 2018). <https://netflixtechblog.com/streaming-video-experimentation-at-netflix-visualizing-practical-and-statistical-significance-7117420f4e9a>
- [79] Linus Torvalds. [n. d.]. `Tcp_input.c` - Linux (v5.11-Rc5). ([n. d.]). https://github.com/torvalds/linux/blob/2ab38c17aac10bf55ab3efde4c4db3893d8691d2/net/ipv4/tcp_input.c#L873
- [80] Donald F Towsley. 2015. TCP, Congestion Control. (Nov. 2015). <http://gaia.cs.umass.edu/cs653/slides/tcp.pdf>
- [81] Belma Turkovic, Fernando A. Kuipers, and Steve Uhlig. 2019. Fifty Shades of Congestion Control: A Performance and Interactions Evaluation. *arXiv:1903.03852 [cs]* (March 2019). [arXiv:1903.03852](https://arxiv.org/abs/1903.03852)
- [82] Belma Turkovic, Fernando A. Kuipers, and Steve Uhlig. 2019. Interactions between Congestion Control Algorithms. In *2019 Network Traffic Measurement and Analysis Conference (TMA)*. 161–168. <https://doi.org/10.23919/TMA.2019.8784674>
- [83] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. Graph Cluster Randomization: Network Exposure to Multiple Universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’13)*. Association for Computing Machinery, New York, NY, USA, 329–337. <https://doi.org/10.1145/2487575.2487695>
- [84] Ranysha Ware, Matthew K. Mukerjee, Srinivasan Seshan, and Justine Sherry. 2019. Beyond Jain’s Fairness Index: Setting the Bar For The Deployment of Congestion Control Algorithms. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks*. ACM, Princeton NJ USA, 17–24. <https://doi.org/10.1145/3365609.3365855>
- [85] Ranysha Ware, Matthew K. Mukerjee, Srinivasan Seshan, and Justine Sherry. 2019. Modeling BBR’s Interactions with Loss-Based Congestion Control. In *Proceedings of the Internet Measurement Conference*. ACM, Amsterdam Netherlands, 137–143. <https://doi.org/10.1145/3355369.3355604>
- [86] David X. Wei, Pei Cao, and Steven H. Low. 2006. TCP Pacing Revisited. (2006). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.2658&rep=rep1&type=pdf>
- [87] Francis Y Yan, Hudson Ayers, Chenzhi Zhu, Sadjad Fouladi, James Hong, Keyi Zhang, Philip Levis, and Keith Winstein. 2020. Learning in Situ: A Randomized Experiment in Video Streaming. In *NSDI*. Santa Clara, CA, USA, 16. <https://www.usenix.org/system/files/nsdi20-paper-yan.pdf>

A ETHICS

While our experiments involve live traffic running on a large video streaming service, our work is not human subjects research, and we have no way to identify the individual users of the platform. We only have access to performance-related data. We ran experiments which improved behavior during congestion, but they did so at the cost of reducing video quality. Netflix’s customers have the ability to opt out of experiments, if they choose to.

B APPENDIX: ANALYSIS OF EXPERIMENTAL DATA

In this appendix we describe our general approach to analysis of data from experiments at scale, and how we apply this approach in the context of the experiments reported in Sections 4 and 5. For the duration of the appendix, we consider data for a fixed representative metric collected on a per-session basis (e.g., average throughput).

In our experiments units are video sessions, and we let A_i denote the treatment condition of session i , where $A_i = 1$ denotes treatment

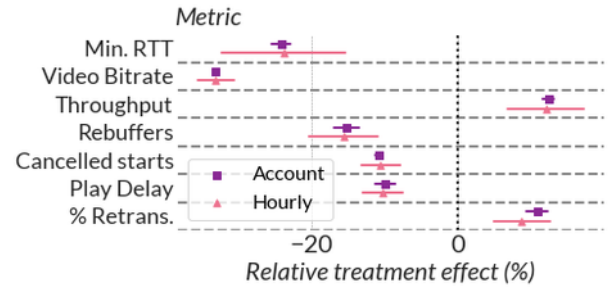


Figure 13: Comparison of treatment effect sizes and confidence intervals when aggregating by hour or by account.

and $A_i = 0$ denotes control. Let Y_i denote the observed outcome on session i . Let $h_i \in \{1, \dots, 24\}$ denote the hour of session i . Our first step in analysis is to aggregate data at the *hourly* level: for each hour $t = 1, \dots, 24$ and each treatment condition $A = 0, 1$, we compute:

$$Z_t(A) = \frac{\sum_i Y_i \mathbb{1}_{h_i=t, A_i=A}}{\sum_i \mathbb{1}_{h_i=t, A_i=A}}.$$

This is the average outcome for sessions in treatment condition A during hour t .

Next, we use a regression approach to estimate the treatment effect [31, Ch 9], using the following model specification:

$$Z_t(A) = c + \beta_0 A + \beta_t + \varepsilon_i, \text{ for all } t, A.$$

Here $t = 1, \dots, 24$ and $A = 0, 1$; β_0 is the coefficient on the treatment indicator; each β_t is a fixed effect to control for hour-of-day heterogeneity; c is an intercept term; and ε_i is the error term. We fit this model using least squares linear regression, and estimate confidence intervals using Newey-West robust standard errors [62] with a lag of two hours. This is a common approach in econometrics to account for autocorrelation between successive hours, and heteroskedasticity in the error terms ε_i . We use hats to denote the corresponding estimates; in particular, $\hat{\beta}_0$ is the estimated coefficient on the treatment indicator, and thus an estimator for the average treatment effect.

We note that the approach we take here—where we aggregate data to the hourly level—essentially makes a worst case assumption that sessions within a given hour and treatment condition are *perfectly correlated* with each other. This is a very conservative assumption, that we feel only strengthens the case in our paper. Though conservative, this is current practice in analysis of switchback experiments in other industries [52]. If we were to analyze the results using the standard account-level standard errors, we would get much tighter confidence intervals as shown in Figure 13. Correcting standard error estimates to properly estimate dependencies between sessions remains an active area of investigation.

We now describe how we apply this approach to our experiments in Sections 4 and 5.

B.1 Application to paired link experiment

In Section 4, sessions on link 1 were randomized 95% to treatment and 5% to control; and sessions on link 2 were randomized 5% to treatment and 95% to control.

We carry out four separate analyses on this data. First, to compute the approximate estimate $\widehat{\text{TTE}}$ for TTE, we consider the 95% of all sessions in the treatment group on link 1 as our treatment sessions ($A_i = 1$); and the 95% of all sessions in the control group on link 2 as our control sessions ($A_i = 0$). We ignore all other sessions. We then follow the analysis workflow above, and set $\widehat{\text{TTE}} = \hat{\beta}_0$ from the resulting fitted regression.

To estimate spillover, we use only the 5% control sessions on link 1 and the 95% control sessions on link 2. We set $A_i = 1$ for the control sessions on link 1, and $A_i = 0$ on link 2. We compute $\widehat{\text{s}}(0.95) = \hat{\beta}_0$ from the resulting fitted regression.

Finally we compute two “naïve” estimates using the difference in means estimator (1) from Section 2. In particular, for $p = 0.95$, we use only the sessions on link 1: we consider all sessions in the treatment group on link 1 as our treatment sessions ($A_i = 1$), and all sessions in the control group on link 1 as our control sessions ($A_i = 0$). All sessions on link 2 are ignored. An analogous approach

is carried out for $p = 0.05$ using the treatment and control sessions on link 2 (ignoring all sessions on link 1), to compute $\hat{\tau}(0.05)$. We aggregate to the account level, not the hour level, as is standard when analyzing A/B tests.

Finally, all reported values are normalized to make them more interpretable. In particular, we divide all estimates by the average across all control sessions on link 2 (where 95% of the traffic was control). This approach ensures all reported values are a relative difference measured against the same global control condition.

B.2 Application to switchback experiments and event studies

In Section 5, we analyzed a switchback experiment and an event study that was emulated using the data from the paired link experiment. This analysis was carried out as follows. For the three days chosen to be treatment intervals, we define all treatment sessions on link 1 to have $A_i = 1$, and ignore all other sessions. For the two days chosen to be control intervals, we define all control sessions on link 2 to have $A_i = 0$, and ignore all other sessions. We then proceed with the analysis workflow above, and report $\hat{\beta}_0$ as our emulated estimate of TTE.