The Internet as Sisyphus: Repeating Measurements, Missing Causes

Loqman Salamatian
Columbia University
New York, USA
loqman@cs.columbia.edu

Abstract

Internet measurement has prioritized what can be observed—latency spikes, packet loss, route changes—over why those observations occur. When performance degrades or routes shift, we often lack the tools to distinguish causes like a congested link from coincidental correlations driven by varying load, measurement bias, or background churn. As a result, explanations remain speculative, and operators struggle to decide whether and how to intervene. This paper explores how causal inference can help fill that gap. We show how classical measurement questions can be framed and analyzed using tools like instrumental variables, causal graphs, and synthetic controls. Finally, we propose design changes for measurement platforms to make causal analysis more feasible.

CCS Concepts

• **Networks** → **Network measurement**; *Network performance modeling*; • **Mathematics of computing** → *Probability and statistics.*

Keywords

Internet measurement, Causal inference

ACM Reference Format:
Loqman Salamatian. 2025. The Internet as Sisyphus: Repeating Measurements, Missing Causes. In *The 24th ACM Workshop on Hot Topics in Networks (HotNets '25), November 17–18, 2025, College Park, MD, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3772356.3772417

HotNets '25, College Park, MD, USA

ACM ISBN 979-8-4007-2280-6/25/11
https://doi.org/10.1145/3772356.3772417

1 Introduction

The Internet's layered design enables modularity by separating concerns: physical infrastructure is abstracted from link protocols, which are abstracted from routing, which in turn is abstracted from transport and applications. This clean separation comes at a cost: it obscures the dependencies and relationships that span across layers. As a result, many of the properties we care most about—performance, resilience, fairness, and compliance—do not originate from any single layer, but from interactions across them, which our current tools and methodologies are poorly equipped to analyze. The abstraction that once enabled clean design now inhibits our ability to trace causes, assign responsibility, or predict the effects of interventions.

This limitation is compounded by other structural challenges: protocols that were not designed to expose their state and a fragmented ecosystem of independently operated networks. The Border Gateway Protocol (BGP)—the de facto standard for interdomain routing—selects paths based on local policies shaped by business agreements and traffic engineering goals. These policies are not publicly visible; observers see the chosen path, but not the rejected alternatives or the reasoning behind the decision. When a route changes, it is often unclear whether the cause was a failure, a configuration update, or a policy shift. As the Internet grows in scale and complexity, even basic questions—Why did this path change? What caused this spike in latency?—require reasoning across layers, networks, and protocols, often with limited visibility into any one of them.

Lacking the visibility and structure needed to support such reasoning, Internet measurement has largely become a *phenomenological* practice. We observe events, describe correlations, and infer plausible explanations. Without causal models, Internet measurement risks becoming Sisyphean: repeating the same observations without uncovering the underlying causes. Fortunately, we are not the first field to grapple with these challenges. A well-established discipline—causal inference— provides a framework for attributing causes, tools to estimate the effects of interventions, and to reason about what would have happened under different conditions than the ones we observed.

Consider the 2021 Facebook outage, where an internal BGP misconfiguration effectively removed Facebook's DNS servers from the Internet [15]. The first observable symptom was DNS failure (i.e., domain names stopped resolving), leading external observers to assume a problem with DNS. But the actual root cause resides layers beneath: a mistaken routing update that severed data center reachability [16]. Similarly, during the 2022 Rogers Communications outage in Canada [34], engineers spent critical hours chasing misleading signals due to missing logs and simultaneous configuration changes. Their postmortem admitted that the true cause was initially misdiagnosed for over half a day [14]. In both cases, surface-level symptoms masked the real failure mechanism.

The disconnect between measurement and meaning: Because of the Internet's increased complexity and opacity, traditional measurement approaches have struggled to keep up. The consequence is a growing disconnect between what we measure and what matters. Too often, this disconnect results in descriptive work that fails to inform design, policy, or operations. A latency spike, for instance, may be harmless—or it may degrade video quality for thousands of users and reflect a traffic engineering decision made by a CDN. In the latter case, identifying the cause points to an actor whose decision can be examined, challenged, or changed. Spikes become meaningful only when placed in a causal chain—something causal models help make explicit.

Many measurement studies are guided by intuitions about routing, topology, or user behavior (e.g., [13, 22, 47, 49, 51] to cite a few). These intuitions implicitly guide choices about what to measure, how to design experiments, and how to interpret findings. Yet while these causal beliefs shape the research, they are rarely articulated explicitly or subjected to formal testing. As a result, conclusions often rest on unverified structural premises, leaving room for confounding, misattribution, or overgeneralizations.

A causal lens for Internet measurement This paper argues for integrating causal inference into Internet measurement, not as a replacement for existing methods, but as a shift in perspective. Our aim is not to discredit prior work: many valuable insights have come from observational studies. Instead, we seek to make their assumptions more explicit, surface potential sources of bias, and show how causal framing can improve interpretation. In particular, we make two contributions:

(1) We demonstrate how causal inference tools such as DAGs, confounding adjustment, instrumental variables, and synthetic controls clarify what questions to ask, what to measure, and how to interpret results (§3). We use a case study testing the common belief that connecting to a local IXP reduces user latency to show how causal tools can turn these claims into testable hypotheses.

(2) We propose concrete design changes to make causal analysis feasible in practice, including DAG-based planning, intent-tagged measurements, and support for exogenous (i.e., externally triggered and system-independent) interventions via existing measurement platforms (§4).

2 Related Work

While we are not the first to pursue statistical rigor or to ask causal questions in Internet measurement [1, 5, 20, 22, 28, 29, 36–38, 44, 45, 50], our contribution is to explicitly foreground the promise of causal inference using the formal language and tools developed in that literature.

Closest to our efforts are a few studies that apply causal methods in well-scoped domains like adaptive bitrate (ABR) streaming and congestion control [6–8, 51]. These subfields are ahead in integrating causal reasoning, in part because the systems are more tractable: dynamics are relatively well-understood, control inputs (e.g., bitrate selection) are narrowly defined, and ground-truth outcomes (e.g., rebuffering) are observable. These studies often assume exogeneity: that decisions affect outcomes without altering latent system conditions. This assumption simplifies counterfactual reasoning—we can ask what would have happened under a different action without modeling its impact on the environment. In many Internet measurement studies, this assumption fails. Routing changes can trigger congestion, CDN decisions can shift demand, and measurement itself can affect behavior. These kinds of feedback effects, where choices and conditions influence each other, are referred to as endogeneity. They complicate causal analysis because the impact of a decision cannot be separated cleanly from the environment in which it occurs.

Still, recent work has begun to bridge this gap. PoiRoot, for example, models the causal structure of path changes and uses BGP poisoning as an instrumental variable to identify root causes [18]. Reuter et al. [33] show how confounding limits the use of observational data to detect RPKI-based filtering and instead rely on targeted experiments via the PEERING testbed [42]. Both illustrate how causal inference—whether through explicit modeling or carefully designed interventions—can extract insight even in settings with partial observability and endogenous dynamics.

3 Primer on Causal Inference

To accurately interpret measurements from complex systems like the Internet, we need tools to distinguish causation from mere correlation. Routing decisions, for instance, are driven by a complex mix of each network's internal policies, business agreements, and adaptive control mechanisms (e.g.,

load-balancing, SDN). A given route may be preferred because it is cheaper, conforms to traffic-engineering goals, or satisfies regulatory or peering requirements. These choices shift with time of day, user demand, and network state. Therefore, performance differences observed across routes may not reflect the inherent quality of the paths themselves, but rather the conditions under which each path is typically used.

Graphical models for causality. To reason clearly in these confounded observational settings, we need a way to represent and test our assumptions about how variables influence each other. Causal graphical models, introduced by Judea Pearl [30], provide a formal language for encoding these assumptions. These models use directed acyclic graphs (DAGs), where nodes represent variables and edges indicate causal influence—distinguishing cause from effect. Every causal claim rests on a view of how the system works: which factors can be treated as external, which interactions are stable, and which influences can be ruled out. Making these assumptions visible is what allows causal inference to be both transparent and testable.

Running example: routing and latency. To illustrate, we use a simple example: How do routing changes affect user-observed latency? Let R denote whether a route change occurred (e.g., switching transit providers), L the latency, and C the level of network congestion (e.g., from diurnal traffic patterns). Congestion can influence both routing and performance: for example, higher load may trigger SDN systems like EdgeFabric [41] or Espresso [52] to shift to a different route ($C \rightarrow R$), while also increasing latency ($C \rightarrow L$).

The ladder of causation. Pearl's ladder of causation offers a framework for understanding the different types of questions we can ask about cause and effect, organized by the level of reasoning they require. Each step up the ladder demands stronger assumptions and a more explicit model of how the system works. In our routing example, the ladder distinguishes three types of questions:

(1) Association: What latency values are observed under different routes? This is answered by estimating $P(L \mid R)$, which reflects statistical association based on observed data.

(2) Intervention: What would the latency be if we forced the network to take a specific route, regardless of the usual conditions that influence routing? Answering this question requires the ability to control the route selection.

(3) Counterfactual: Given that a route change occurred and a high latency was observed, what would the latency have been if, in that same situation, the route had not changed? This question asks about a hypothetical alternative for a specific event and requires a detailed model of how routing and latency interact, including the influence of all relevant variables, to simulate what would have happened under a different choice.

For these causal questions to be well-defined, we typically rely on the Stable Unit Treatment Value Assumption (SUTVA). SUTVA has two parts: first, that each unit's outcome depends only on its own treatment and not on the treatments applied to others (i.e., there is no interference); and second, that the treatment itself is consistent and well-defined (i.e., there is only one version of it, applied in the same way to all treated units). These assumptions allow us to interpret "the effect" of an intervention unambiguously, ensuring that differences in outcomes reflect the treatment itself rather than variations in how it was applied or interactions between units. For example, if "changing a routing policy" sometimes means adjusting a local-preference value and other times means depeering from a transit provider entirely, then the "treatment" is not well-defined. These actions differ in scope and impact, so we cannot attribute a single causal effect to "changing the routing policy."

Confounding and collider bias. The first step in estimating the effect of an intervention is identifying and adjusting for *confounders* (i.e., variables that influence both the decision being analyzed and the outcome and can create spurious associations that obscure the true causal effect). In our example, C influences both R and L, making it a confounder. In the DAG, this confounder appears as a backdoor path: $R \leftarrow C \rightarrow L$. To isolate the effect of route on latency, we need to "block" this path. In practice, *blocking the confounding path* requires analyzing many measurements taken under varying conditions, and comparing latencies across routes only when C is similar, e.g., at comparable load levels. By holding C constant, we block the indirect influence of congestion on both route selection and latency, allowing us to attribute observed differences more confidently to the route itself rather than to the conditions under which it was chosen. While adjusting for confounders is essential, not all conditioning helps. A collider is a variable that is influenced by two others. In the context of speed-test analysis, the decision to run a test can act as a collider: both changes in routing (e.g., switching to a new ISP) and poor network performance (e.g., high latency or low throughput) can independently prompt users to run a test. If we analyze only the speed tests that are actually run, we are conditioning on this shared outcome. This can create a spurious association between routing changes and performance degradation, even if no causal link exists—because both make speed tests more likely to occur. As a result, claims about the effect of routing on performance, drawn solely from observed tests, may be biased by this collider.

Confounding and collider bias both reflect a deeper issue: observational data must be interpreted in light of how they

were generated. Confounders bias estimates when left unadjusted; colliders do so when conditioned on. But in both cases, the challenge is the same—causal inference depends not just on which variables are measured, but on understanding the pathways through which the data was produced. Without this, even large, rich datasets can lead to the wrong conclusions.

An example of confounding bias A SIGCOMM'21 paper on cellular reliability [24] finds higher failure rates at the strongest signal levels. To their credit, the authors recognize this anomaly and attribute it to dense deployments in transit hubs, which introduce interference and overhead. However, this makes signal strength a proxy, not a cause: deployment density confounds both signal strength and failure. Without adjusting for this factor, the observed correlation is misleading.

Using randomization and natural experiments. The cleanest way to estimate causal effects is through randomized assignment, where variation in the treatment is entirely exogenous, that is, driven by factors external to the system rather than by its internal state or behavior. Consider how M-Lab assigns users to measurement servers during speed tests, using a load balancer that randomly routes each test to one of several sites in a nearby city [17, 26]. This mechanism introduces controlled, random variation in routing: the user's traffic may traverse entirely different AS paths depending on the assigned site, even when their device and network are the same. Because M-Lab sites in certain metros are standardized (bare-metal servers with identical configurations) and clients are load-balanced evenly across them,[1] differences in performance across sites can be attributed directly to routing rather than user intent or server behavior. This is effectively a randomized experiment, the gold standard for causal inference.

When full randomization is not possible, we can look for natural experiments: external events that induce quasi-random variation in decisions. These events mimic randomized trials by introducing exogenous shocks that are plausibly independent of the outcome. A common method for leveraging natural experiments is the use of an *instrumental variable*. An instrumental variable is a factor that (1) influences the decision being studied and (2) affects the outcome only through that decision—not through any other route in the DAG (*the exclusion restriction*). When these conditions hold, the instrument isolates the portion of variation in the decision that is effectively random with respect to the outcome, enabling unbiased causal estimates.

[1]This observation does not hold for cloud-hosted M-Lab sites, which rely on different infrastructure.

Many sources of routing variation on the Internet such as software updates, policy changes, traffic engineering decisions, or link failures, can, under certain conditions, serve as viable instruments, but doing so requires careful justification. In each case, the key challenge lies in establishing that the event is exogenous: that it affects performance only through its impact on routing and not through correlated factors like congestion, maintenance activity, or recovery from a previous fault. For example, suppose an operator changes its BGP local preference to favor a cheaper transit provider. While this might appear to create a clean intervention, the change can also alter upstream load and trigger adaptive responses in neighboring networks, which in turn affect congestion and path length. In such cases, the exclusion restriction is violated because the intervention influences performance through multiple causal channels, not just the intended route change. Still, not all such events are invalid by design. Scheduled link maintenance or sudden policy shifts imposed by regulators may generate exogenous variation that approximates random assignment when their timing and scope are independent of network conditions. The distinction between a valid and invalid instrument thus hinges often on the strength of the justification: whether researchers can credibly argue that its effect is confined to the targeted mechanism. In reality, none of these events arrive with clean labels saying "instrumental variables." Identifying viable ones requires a mix of domain insight, careful measurement design, and a healthy dose of skepticism.

An example of misinterpreted natural experiment. An IMC'21 paper on user latency sensitivity [47] analyzes variation in page load times and user interactions to infer how latency affects user behavior. The authors describe this variation as a natural experiment, but do not identify an exogenous source of latency independent of user intent or engagement. Instead, they normalize for observable factors to reduce confounding. While this technique improves robustness, the result remains observational. Without a valid instrumental variable or randomized assignment, the variation used does not support formal causal claims.

Building counterfactuals. The most challenging and arguably most important question in causal inference is the counterfactual: What would have happened under a different scenario, given what actually occurred? Counterfactuals go beyond estimating average treatment effects or identifying system-wide trends and ask about specific alternate realities.

For instance, suppose a user's video call experienced degraded quality right after their traffic was rerouted through a new transit provider. The counterfactual question is: Would the call quality have been better had the route change not occurred?

Answering this question requires more than blocking confounders or leveraging instrumental variables—it demands a model of how the system behaves under different conditions. In principle, one could specify a structural model: a DAG that encodes the dependencies between routing decisions, traffic load, queuing behavior, protocol dynamics, and performance outcomes. With such a model, counterfactuals could be computed by simulating interventions. But in practice, this is rarely feasible. Even if the structure were known, estimating the necessary relationships would require specific measurements across layers, networks, and time, an unrealistic expectation given the Internet's decentralized, evolving, and opaque nature. Despite the difficulty, counterfactuals are exactly the kind of reasoning that operators implicitly rely on. When something goes wrong and multiple variables change (e.g., routing updates, traffic spikes, policy adjustments), the question they want answered is rarely "Is routing correlated with latency?". It is more often: "Was this degradation caused by the routing change, or would it have happened anyway?" Counterfactuals are the only way to formally pose and answer such questions.

A more practical alternative to building a full causal model is the synthetic control method [4, 53]. It estimates what would have happened on a path that changed (e.g., due to a routing shift) by constructing a weighted combination of similar paths that did not. The weights are selected so that the combined pre-change performance of these paths closely tracks that of the original. The intuition is that, while no single comparison path may be a perfect match, their weighted average can smooth out individual noise and approximate the underlying factors driving performance. This synthetic trajectory then serves as a stand-in to estimate how the original path would have performed had the change not occurred. This method relies on a few key conditions [3]: no interference between units (the routing change on one path should not affect others in the donor pool), a good pre-change fit (the synthetic path must closely track the actual path before the change), and no other major shocks coinciding with the change (e.g., infrastructure upgrades that could independently impact performance).

In our context, suppose a user's path is rerouted through a new transit provider, and video call quality drops. To estimate what would have happened without the routing change, we build a synthetic version of the path by combining others that (a) did not reroute and (b) showed similar pre-change trends. Rather than matching on static features, we align based on temporal performance patterns (e.g., latency, traffic, time-of-day usage), which controls for external factors like congestion or regional demand. The synthetic path's post-change performance then serves as an estimate of the counterfactual, i.e., what the user would have experienced had the route stayed the same.

ASN / City	RTT Δ (ms)	RMSE Ratio	p
3741 / East London	+3.40	236	0.053
3741 / Johannesburg	+1.50	17	0.857
37053 / Cape Town	−0.12	23	0.862
37611 / Edenvale	−0.91	16	0.406
37680 / Durban	−2.20	199	0.086
327966 / Polokwane	−7.28	85	0.333
328622 / eMuziwezinto	−1.30	18	0.143
328745 / Johannesburg	+0.30	18	0.857

Table 1: Estimated RTT change for paths that begin crossing NAPAfrica-JNB. All cities are in South Africa.

This approach has several advantages: it is data-driven and does not require strong assumptions about the functional form of the underlying system. Importantly, synthetic control takes into account how things were changing over time, not just what they looked like at a single moment. By building the comparison on shared performance trajectories before the change, synthetic control helps isolate the impact of a single event from the broader churn and gives us a more realistic estimate of what would have happened if the change hadn't occurred. In doing so, it provides a powerful and realistic framework for counterfactual reasoning in settings where randomized experiments are impossible and full structural models are infeasible.

An example of incorrect counterfactual reasoning A SIGMETRICS'24 paper on Internet resilience, Xaminer [32], simulates physical-layer failures (e.g., cable cuts) and traces their downstream effects on network-layer connectivity. While this provides a valuable map of potential exposure, it falls short of modeling how the network would *respond* to such failures. True resilience analysis requires counterfactual reasoning: not just asking what infrastructure is at risk, but how routing, connectivity, and performance would change if a specific failure occurred. Without modeling these dynamic adaptations, the analysis risks conflating exposure with impact and cannot quantify the actual robustness of routing in the face of real-world events.

Case study: Does joining an IXP reduce latency? [2]

A common belief in network operations is that once an access ISP connects to a domestic IXP, users behind it will benefit from lower latency to access local content instead of being tromboned through distant transit providers. This case study illustrates how one can use causal inference to transform that operational belief into a formal, testable hypothesis: instead of asking simply whether RTT decreases, we ask whether IXP membership causes latency decreases once confounders are accounted for. Answering this question requires more than observing a drop in latency. While we omit the DAG due to space constraints, we identify key factors that

[2] All code and data used are available at our public repository [39].

may confound the observed relationship: IXP deployment can trigger topological and routing changes, which are themselves influenced by independent variables such as traffic load, business policy, and infrastructure upgrades. Without a causal model to account for these dependencies, we risk misattributing the cause of observed RTT shifts. We frame the addition of an IXP as an intervention: the "treatment" is the first appearance of the IXP in a path. If the belief is true, this intervention should cause a drop in RTT. Before performing the analysis, we first assess whether the conditions for causal estimation are reasonably satisfied. The "no interference" assumption may not hold perfectly: adding an IXP not only introduces a new path but also reshapes the local routing topology. Traffic shifts toward the new link can alter path preferences and congestion for neighboring networks, making the treatment's effect partially dependent on its surroundings. Similarly, overlapping infrastructure changes (e.g., the three new PoPs deployed since June 2025 according to PeeringDB [31]) could confound observed latency changes. As with all analyses based on user-initiated speed tests, our results inherit potential sampling biases, since measurements are not uniformly distributed across users or time. We leave a formal assessment of how these factors influence our results to future work, but assume their impact is limited, making synthetic control an appropriate framework for estimating the local impact of IXP adoption.

With these caveats in mind, we test: *For a network that previously did not use an IXP, how does median RTT change after the IXP first appears in the path?*

For this case study, we use M-Lab speed tests together with the traceroutes automatically triggered after each test. We determine whether a path crosses the NAPAfrica IXP [27] by matching hop IP addresses against addresses announced by the IXP [2][3]. We analyze performance at the ⟨ASN, city⟩ level: users within the same ASN and city are likely to share routing policies, last-mile conditions, and local peering options, while still allowing us to distinguish geographically distinct regions within the same provider.

To estimate the impact of the IXP crossing, we apply the robust synthetic control method [9]: for each ⟨ASN, city⟩ that starts crossing the IXP from June 2025, we construct a weighted combination of ⟨ASN, city⟩ from the donor pool, matching the RTT trends before the IXP appeared. This approach controls for broader performance shifts that may be unrelated to the IXP itself (e.g., regional congestion, diurnal effects) by ensuring that the synthetic path closely matches the treated path's behavior prior to the IXP. While we do not model all possible confounders explicitly, this pre-change alignment and the use of a donor pool that does not route via the IXP provide a pragmatic first-order control for temporal and structural biases. We then compare the observed RTT after the change to this synthetic baseline

Table 1 shows the estimated RTT change and two diagnostic statistics: the RMSE ratio and a placebo-based p-value. The RMSE ratio compares the synthetic control model's fit error after the IXP appears in the path to its fit error beforehand. A large increase may indicate that the path's behavior diverged from the donor pool after the change. The p-value is computed by comparing this RMSE ratio to those from placebo models applied to paths that did not cross the IXP; it quantifies how likely such a shift could arise from model noise alone. Most paths show small changes or high p-values. Two paths (ASN3741 / East London and ASN37680 / Durban) show moderate RTT changes (3.4 and −2.2 ms) with marginal p-values (<0.10). The largest observed drop (−7.3 ms in Polokwane) is not statistically significant ($p = 0.33$).

While RTT occasionally decreases after traversing the IXP, the effect is neither consistent nor robust. This result demonstrates how causal inference tools can convert a widely repeated operational claim into a testable hypothesis and show when the data fail to support it.

4 Measurement Design for Causal Analysis

The first step in any causal analysis is to articulate assumptions about how the system works by constructing a causal graph (i.e., a DAG). We recommend that measurement studies build such a DAG a priori to make structural assumptions explicit—what variables matter, how they interact, and where interventions may take effect. DAGs are not learned from data alone; they require domain insight, protocol knowledge, and operational experience. They clarify which effects are identifiable and what measurements are needed to isolate them.

Whether causal effects are identifiable hinges on which variables are observed, how much variation exists across conditions, and whether measurements capture the right parts of the system—those where interventions produce meaningful differences in outcomes. A central lesson from causal inference is that more data does not necessarily mean better insight. The value of a measurement lies in whether it helps resolve causal ambiguity, for example, by blocking backdoor paths or inducing variation along a hypothesized causal link. This perspective does not dismiss earlier measurement efforts—many studies have yielded valuable insights by analyzing the available public data (e.g., [12, 19]) or focusing their measurements on a clearly defined topic of interest (e.g., [10, 23]). The key distinction is that causal reasoning imposes stricter demands: coverage, volume, and regularity are useful only insofar as they help answer a causal question.

[3]We chose NAPAfrica-Johannesburg because, during the month we analyzed, it had the largest number of new ASN–city pairs that began crossing the IXP compared to other locations.

The strategic selection of measurements—what is measured, when, where, and under what conditions—becomes a central design problem. Recent work has begun to use these ideas in the context of topology discovery [40].

To generalize this approach, researchers need better tools for reasoning about measurement design. Before collecting data, one should be able to define a causal question (e.g., the impact of a routing change on latency), specify the relevant variables (e.g., RTT, path, time), and assess whether the planned setup can identify the desired effect under plausible assumptions. Existing libraries like Dagitty [48], DoWhy [43], and EconML [11] already support this kind of reasoning in other domains. Bringing such tools to networking could help researchers validate assumptions, avoid invalid inferences, and guide the design of both passive and active measurement campaigns. We envision future measurement studies adopting a causal protocol: specify the causal graph, identify confounders and instruments, validate assumptions, and report uncertainty in causal estimates.

Of course, enabling this kind of planning requires infrastructure support. While platforms like RIPE Atlas [35] and Archipelago [46] have advanced the scale and reach of Internet measurement, they were built with conventional goals in mind—broad coverage, fixed-interval sampling, and minimal interference—rather than causal inference. To make causal analysis more feasible, we propose a set of enhancements aligned with a measurement-for-causality mindset:

(1) Platforms should support conditional measurement activation triggered by external signals (e.g., BGP changes), scheduled maintenance windows, or IXP outage notifications. These time-bounded disruptions provide natural experiments where routing or availability changes can be cleanly linked to shifts in performance. Certain platforms used to operate that way (e.g., Hubble [21]), and Arkipelago's recent deployment is allowing this logic as well [25].

(2) Each measurement could be tagged with its purpose or trigger context (e.g., baseline monitoring or anomaly reaction), enabling downstream analysts to properly account for selection bias (e.g., conditioning on colliders).

(3) Platforms could expose APIs that allow researchers or clients to induce exogenous variation—providing the knobs needed for causal inference. PEERING [42] offers a nice template: it lets researchers control BGP announcements from a real AS, enabling experiments that selectively influence routing decisions. Bringing similar flexibility to platforms like RIPE Atlas and Ark would expand the space of causal experiments. Examples include toggling IPv4 vs. IPv6 to alter AS paths, rotating DNS resolvers to shift CDN edge selection. These mechanisms act as instrumental variables, isolating specific effects (e.g., of routing on performance) while minimizing confounding.

(4) Unlike traditional causal inference settings where treatments and observations are exogenous, Internet measurement is reflexive: who measures and when depends on the system's internal state. Measurement is therefore an action, shaped by cost, user intent, and system dynamics. Speed tests, for instance, are triggered by users experiencing poor performance. Recognizing this endogeneity as an asset is an important dimension: who measures and when reflects underlying network conditions. Rather than discarding this bias, we should strive to treat it as signal, using it to guide adaptive measurement and causal attribution.

Challenges, Limitations, and Feasibility. Establishing causality on the Internet is inherently difficult. We cannot observe every relevant variable across layers and networks; user-initiated measurements sample non-randomly from the true population; conditions evolve rapidly, making stable baselines elusive; and interventions on one part of the Internet can affect others on a much larger scale than in other disciplines, often violating the assumption that an intervention has "no interference" beyond its target. A routing change or configuration tweak in one network can ripple across continents within seconds, altering congestion, reachability, and performance far beyond its origin. In most domains where causal inference is applied, interference is geographically or socially bounded—for instance, the effects of a new teaching method may influence other students in the same classroom but not schools across the country. Ethical constraints compound these limits: researchers cannot deliberately degrade performance or trigger outages simply to test their hypotheses. Yet acknowledging these challenges is not an argument for inaction. Even when perfect isolation is unattainable and fully random variation is unavailable, we can still design analyses around natural experiments where partial variation exists, revealing causal relationships without harming users, and providing a structured way to articulate what can, and cannot, be inferred from the data.

5 Conclusion

Internet measurement has long focused on observation rather than explanation. Integrating causal inference offers a principled way to connect what we see to why it happens. This shift enables diagnosis and informed intervention in an increasingly complex Internet. By treating measurement as a causal instrument, we move beyond repeating observations toward understanding the mechanisms that drive them.

Acknowledgements

I would like to acknowledge Tom Koch, Ethan Katz-Bassett, Martin Devaux, and the anonymous HotNets reviewers and Zili Meng, the shepherd, for their feedback on the earlier versions of this paper.

References

[1] 2004. Strategies for sound Internet measurement. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement* (Taormina, Sicily, Italy) *(IMC '04)*. Association for Computing Machinery, New York, NY, USA, 263–271. doi:10.1145/1028788.1028824

[2] 2025. NAPAfrica IX Johannesburg – PeeringDB. https://www.peeringdb.com/ix/592. Accessed: 2025-10-20.

[3] Alberto Abadie. 2021. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature* 59, 2 (2021), 391–425.

[4] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* 59, 2 (2015), 495–510.

[5] Muhammad Abdullah, Zafar Ayyub Qazi, and Ihsan Ayyub Qazi. 2022. Causal impact of Android go on mobile web performance. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) *(IMC '22)*. Association for Computing Machinery, New York, NY, USA, 113–129. doi:10.1145/3517745.3561456

[6] Neil Agarwal, Rui Pan, Francis Y Yan, and Ravi Netravali. 2025. Mowgli: Passively Learned Rate Control for {Real-Time} Video. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*. 579–594.

[7] Abdullah Alomar, Pouya Hamadanian, Arash Nasr-Esfahany, Anish Agarwal, Mohammad Alizadeh, and Devavrat Shah. 2023. CausalSim: A Causal Framework for Unbiased Trace-Driven Simulation. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, Boston, MA, 1115–1147. https://www.usenix.org/conference/nsdi23/presentation/alomar

[8] Abdullah Alomar, Pouya Hamadanian, Arash Nasr-Esfahany, Anish Agarwal, Mohammad Alizadeh, and Devavrat Shah. 2023. CausalSim: A Causal Framework for Unbiased Trace-Driven Simulation. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, Boston, MA, 1115–1147. https://www.usenix.org/conference/nsdi23/presentation/alomar

[9] Muhammad Amjad, Devavrat Shah, and Dennis Shen. 2018. Robust synthetic control. *Journal of Machine Learning Research* 19, 22 (2018), 1–51.

[10] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. 2020. Cloud Provider Connectivity in the Flat Internet. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) *(IMC '20)*. Association for Computing Machinery, New York, NY, USA, 230–246. doi:10.1145/3419394.3423613

[11] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: A Python package for ML-Based heterogeneous treatment effects estimation. *Version 0. x* (2019).

[12] Zachary S. Bischof, Kennedy Pitcher, Esteban Carisimo, Amanda Meng, Rafael Bezerra Nunes, Ramakrishna Padmanabhan, Margaret E. Roberts, Alex C. Snoeren, and Alberto Dainotti. 2023. Destination unreachable: Characterizing Internet outages and shutdowns. In *Proceedings of the ACM SIGCOMM 2023 Conference* (New York, NY, USA) *(ACM SIGCOMM '23)*. Association for Computing Machinery, New York, NY, USA, 608–621. doi:10.1145/3603269.3604883

[13] Randy Bush, Olaf Maennel, Matthew Roughan, and Steve Uhlig. 2009. Internet optometry: assessing the broken glasses in Internet reachability. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement* (Chicago, Illinois, USA) *(IMC '09)*. Association for Computing Machinery, New York, NY, USA, 242–253. doi:10.1145/1644893.1644923

[14] Canadian Radio-television and Telecommunications Commission (CRTC). 2024. CRTC Report: Rogers Communications July 2022 Network Outage. https://crtc.gc.ca/eng/publications/reports/xona2024.htm Accessed: 2025-06-23.

[15] Cloudflare. 2021. *Understanding How Facebook Disappeared from the Internet.* https://blog.cloudflare.com/october-2021-facebook-outage/ Accessed: 2025-06-23.

[16] Facebook Engineering. 2021. *More details about the October 4 outage.* https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/ Accessed: 2025-06-23.

[17] Phillipa Gill, Christophe Diot, Lai Yi Ohlsen, Matt Mathis, and Stephen Soltesz. 2022. M-Lab: User-Initiated Internet Data for the Research Community. *ACM SIGCOMM Computer Communication Review* 52, 1 (2022), 34–37.

[18] Umar Javed, Italo Cunha, David Choffnes, Ethan Katz-Bassett, Thomas Anderson, and Arvind Krishnamurthy. 2013. PoiRoot: Investigating the root cause of interdomain path changes. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 183–194.

[19] Weifan Jiang, Tao Luo, Thomas Koch, Yunfan Zhang, Ethan Katz-Bassett, and Matt Calder. 2021. Towards identifying networks with Internet clients using public data. In *Proceedings of the 21st ACM Internet Measurement Conference* (Virtual Event) *(IMC '21)*. Association for Computing Machinery, New York, NY, USA, 753–762.

[20] Srikanth Kandula and Ratul Mahajan. 2009. Sampling biases in network path measurements and what to do about it. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement* (Chicago, Illinois, USA) *(IMC '09)*. Association for Computing Machinery, New York, NY, USA, 156–169. doi:10.1145/1644893.1644912

[21] Ethan Katz-Bassett, Harsha V. Madhyastha, John P. John, Arvind Krishnamurthy, David Wetherall, and Thomas Anderson. 2008. Studying black holes in the Internet with Hubble. In *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation* (San Francisco, California) *(NSDI'08)*. USENIX Association, USA, 247–262.

[22] S. Shunmuga Krishnan and Ramesh K. Sitaraman. 2012. Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. In *Proceedings of the 12th ACM Internet Measurement Conference* (Boston, Massachusetts, USA) *(IMC '12)*. Association for Computing Machinery, New York, NY, USA, 211–224. doi:10.1145/2398776.2398799

[23] Fangfan Li, Arian Akhavan Niaki, David Choffnes, Phillipa Gill, and Alan Mislove. 2019. A large-scale analysis of deployed traffic differentiation practices. In *Proceedings of the ACM Special Interest Group on Data Communication* (Beijing, China) *(SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 130–144. doi:10.1145/3341302.3342092

[24] Yang Li, Hao Lin, Zhenhua Li, Yunhao Liu, Feng Qian, Liangyi Gong, Xianlong Xin, and Tianyin Xu. 2021. A nationwide study on cellular reliability: measurement, analysis, and enhancements. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference* (Virtual Event, USA) *(SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 597–609. doi:10.1145/3452296.3472908

[25] Matthew Luckie, Shivani Hariprasad, Raffaele Sommese, Brendon Jones, Ken Keys, Ricky Mok, and et al. 2025. An Integrated Active Measurement Programming Environment. In *Proceedings of the 26th International Conference on Passive and Active Network Measurement (PAM 2025) (Lecture Notes in Computer Science, Vol. 15567)*. Springer, 137–152. doi:10.1007/978-3-031-85960-1_12

[26] Measurement Lab. 2025. Measurement Lab – an open, distributed Internet measurement platform. https://www.measurementlab.net/. A consortium providing open-source tools and the largest publicly available Internet performance data collections.

[27] NAPAfrica. 2025. NAPAfrica: Africa's most active Internet Exchange Point. https://www.napafrica.net/. Accessed July 2025.

[28] Hung X Nguyen and Matthew Roughan. 2012. Rigorous statistical analysis of internet loss measurements. *IEEE/ACM Transactions on Networking* 21, 3 (2012), 734–745.

[29] Hung X. Nguyen and Patrick Thiran. 2007. Network loss inference with second order statistics of end-to-end flows. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (San Diego, California, USA) *(IMC '07)*. Association for Computing Machinery, New York, NY, USA, 227–240. doi:10.1145/1298306.1298339

[30] Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment* (2010), 39–58.

[31] PeeringDB. 2025. PeeringDB: The Interconnection Database. https://www.peeringdb.com/. Accessed: June 2025.

[32] Alagappan Ramanathan, Rishika Sankaran, and Sangeetha Abdu Jyothi. 2024. Xaminer: An Internet Cross-Layer Resilience Analysis Tool. In *Abstracts of the 2024 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems* (Venice, Italy) *(SIGMETRICS/PERFORMANCE '24)*. Association for Computing Machinery, New York, NY, USA, 99–100. doi:10.1145/3652963.3655091

[33] Andreas Reuter, Randy Bush, Italo Cunha, Ethan Katz-Bassett, Thomas C Schmidt, and Matthias Wählisch. 2018. Towards a rigorous methodology for measuring adoption of RPKI route validation and filtering. *ACM SIGCOMM Computer Communication Review* 48, 1 (2018), 19–27.

[34] Reuters. 2022. *Rogers Communications services down for thousands of users - Downdetector*. https://www.reuters.com/business/media-telecom/rogers-communications-services-down-thousands-users-downdetector-2022-07-08/ Accessed: 2025-06-23.

[35] RIPE NCC. 2025. RIPE Atlas. https://atlas.ripe.net/.

[36] Matthew Roughan. 2005. Fundamental bounds on the accuracy of network performance measurements. *ACM SIGMETRICS Performance Evaluation Review* 33, 1 (2005), 253–264.

[37] Matthew Roughan. 2016. Lies, Damn Lies, and Internet Measurements: Statistics and Network Measurements. In *IFIP International Workshop on Traffic Monitoring and Analysis (TMA)*. Louvain-la-Neuve, Belgium. https://roughan.info/talks/tma_2015.pdf Keynote talk.

[38] Kavé Salamatian and Serge Fdida. 2003. A framework for interpreting measurement over Internet. In *Proceedings of the ACM SIGCOMM Workshop on Models, Methods and Tools for Reproducible Network Research* (Karlsruhe, Germany) *(MoMeTools '03)*. Association for Computing Machinery, New York, NY, USA, 87–94. doi:10.1145/944773.944788

[39] Loqman Salamatian. 2025. The Internet as Sisyphus: Repeating Measurements, Missing Causes (Code and Data Repository). https://github.com/Burdantes/Internet-As-Sisyphus. Accessed October 2025.

[40] Loqman Salamatian, Kevin Vermeulen, Italo Cunha, Vasilis Giotsas, and Ethan Katz-Bassett. 2024. metAScritic: Reframing AS-Level Topology Discovery as a Recommendation System. In *Proceedings of the 2024 ACM on Internet Measurement Conference* (Madrid, Spain) *(IMC '24)*. Association for Computing Machinery, New York, NY, USA, 337–364. doi:10.1145/3646547.3688429

[41] Brandon Schlinker, Hyojeong Kim, Timothy Cui, Ethan Katz-Bassett, Harsha V. Madhyastha, Italo Cunha, James Quinn, Saif Hasan, Petr Lapukhov, and Hongyi Zeng. 2017. Engineering Egress with Edge Fabric: Steering Oceans of Content to the World. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (Los Angeles, CA, USA) *(SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 418–431. doi:10.1145/3098822.3098853

[42] Brandon Schlinker, Kyriakos Zarifis, Italo Cunha, Nick Feamster, and Ethan Katz-Bassett. 2014. Peering: An as for us. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*. 1–7.

[43] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. (11 2020). doi:10.48550/arXiv.2011.04216

[44] Christopher A. Small, Narendra Ghosh, Hany Saleeb, Margo I. Seltzer, and Keith Smith. 1997. *Does Systems Research Measure Up?* Technical Report TR-16-97. Harvard University, Computer Science Department. https://dash.harvard.edu/bitstreams/7312037d-caf7-6bd4-e053-0100007fdf3b/download Accessed: July 2025.

[45] Bruce Spang, Veronica Hannan, Shravya Kunamalla, Te-Yuan Huang, Nick McKeown, and Ramesh Johari. 2021. Unbiased experiments in congested networks. In *Proceedings of the 21st ACM Internet Measurement Conference*. ACM New York, NY, USA, New York, NY, USA, 80–95.

[46] CAIDA Archipelago (Ark) Team. 2025. Archipelago (Ark) Measurement Infrastructure. https://catalog.caida.org/collection/archipelago. Accessed: 2025-07.

[47] Parth Thakkar, Rohan Saxena, and Venkata N. Padmanabhan. 2021. AutoSens: inferring latency sensitivity of user activity through natural experiments. In *Proceedings of the 21st ACM Internet Measurement Conference* (Virtual Event) *(IMC '21)*. Association for Computing Machinery, New York, NY, USA, 15–21. doi:10.1145/3487552.3487839

[48] Benito Van Der Zander, Johannes Textor, and Maciej Liskiewicz. 2015. Efficiently finding conditional instruments for causal inference. In *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 3243–3249.

[49] Feng Wang, Zhuoqing Morley Mao, Jia Wang, Lixin Gao, and Randy Bush. 2006. A measurement study on the impact of routing events on end-to-end Internet path performance. *ACM SIGCOMM Computer Communication Review* 36, 4 (2006), 375–386.

[50] Walter Willinger and Vern Paxson. 1998. Where mathematics meets the Internet. *Notices of the AMS* 45, 8 (1998), 961–970.

[51] 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 *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. USENIX Association, Santa Clara, CA, 495–511. https://www.usenix.org/conference/nsdi20/presentation/yan

[52] Kok-Kiong Yap, Murtaza Motiwala, Jeremy Rahe, Steve Padgett, Matthew Holliman, Gary Baldus, Marcus Hines, Taeeun Kim, Ashok Narayanan, Ankur Jain, Victor Lin, Colin Rice, Brian Rogan, Arjun Singh, Bert Tanaka, Manish Verma, Puneet Sood, Mukarram Tariq, Matt Tierney, Dzevad Trumic, Vytautas Valancius, Calvin Ying, Mahesh Kallahalla, Bikash Koley, and Amin Vahdat. 2017. Taking the Edge off with Espresso: Scale, Reliability and Programmability for Global Internet Peering. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (Los Angeles, CA, USA) *(SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 432–445. doi:10.1145/3098822.3098854

[53] Jakob Zeitler, Athanasios Vlontzos, and Ciarán Mark Gilligan-Lee. 2023. Non-parametric identifiability and sensitivity analysis of synthetic control models. In *Conference on Causal Learning and Reasoning*. PMLR, PMLR, 850–865.