Unknown Is Not False: On Negative Results and the Limits of Inference

Fabián E. Bustamante*
Northwestern University
USA

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The author takes full responsibility for this article's technical content. Comments can be posted through CCR Online.

Abstract

Negative results often conflate two fundamentally different scientific outcomes: evidence that a hypothesis is false, and insufficient evidence to decide. I argue that the latter, Unknown, can itself be a meaningful scientific result when we can characterize where and why the available evidence ceases to support an inference. Using traceroute geolocation as a worked example, I show how making this boundary explicit can prevent unsupported conclusions, make scarce validation more useful, and direct new measurement. Unknown is not a substitute for independent validation: internally consistent evidence can still be systematically wrong. But neither should Unknown be treated as failure. Sometimes establishing where our ability to tell True from False ends is itself the result.

CCS Concepts

• **Networks** → **Network measurement**; • **General and reference** → *Empirical studies.*

Keywords

Negative results, Internet measurement, uncertainty, inference, validation.

1 Introduction

Scientific research rarely proceeds as cleanly as the papers we write about it. Research is an investigation: we pose questions, collect measurements, discover that some of them do not mean what we thought they meant, encounter explanations we cannot distinguish, and sometimes end without being able to answer the question we originally asked, or find ourselves asking a rather different one.

This is particularly relevant to what we commonly call *negative results*. The term sounds simple enough: an experiment did not produce the expected or desired result, but it hides an important distinction. Consider an experiment intended to evaluate some hypothesis H. At the risk of oversimplifying, the investigation can leave us in at least three states: the evidence supports H; the evidence allows us to reject H; or the evidence does not allow us to decide. Call these **True**, **False**, and **Unknown**. The last two are often filed together as negative results, as variations of *it didn't work*. But they are fundamentally different scientific outcomes.

False tells us something about the world: rejecting a hypothesis or ruling out an explanation reduces the space of possibilities, and that is knowledge.

Unknown tells us something different. The experiment may be underpowered, or the instrument unable to separate the alternatives; either way it has not established that the hypothesis is false, only a limit on what can be inferred. That limit need not be a failure of the research: if we can characterize *why* the experiment cannot distinguish among them, we have learned something about the measurement process itself, what it can observe and what it cannot. Failure to establish H is not evidence that H is false.

This issue is especially acute in much of Internet measurement. We study systems we do not control, from a sparse and heterogeneous collection of vantage points, using signals that are often indirect and incomplete. What is observable, and consequently what we infer, can depend strongly on the measurement process and vantage points themselves [1, 17, 22]. In such an environment, *we did not observe it*, *our method cannot determine it*, and *it is not there* are dangerously easy conclusions to conflate.

This article argues that we should make Unknown an explicit scientific outcome, not merely a residual between success and failure. Doing so requires more than attaching an uncertainty score to every answer. A useful Unknown must be characterized: we should be able to say *why* the available evidence is insufficient and, where possible, validate that boundary. I use mapping logical Internet paths onto physical geography as the worked example: a setting where the boundary can be made operational, and where identifying it tells us where scarce validation is worth most.

Before getting there, however, it is worth separating the two outcomes that the label *negative result* so casually puts back together. The easier one is False.

2 False Is Knowledge

A well-supported False should hardly need defending as a scientific result. Refuting a hypothesis eliminates an explanation; showing that a mechanism cannot account for an observation eliminates a causal story; demonstrating that a design cannot satisfy a set of requirements eliminates part of the design space. In each case, the result is negative only in the sense that it rules something out. Scientifically, it has reduced the space of possibilities.

Computer networking and distributed systems have some particularly celebrated examples. CAP establishes that, in the presence of a network partition, a distributed system cannot simultaneously guarantee consistency and availability [12]. FLP establishes that

*This article grew out of an invited keynote at the first ACM SIGCOMM Workshop on Negative Results in Network Measurement (NetNeg 2026).

deterministic consensus cannot be guaranteed in a fully asynchronous system in the presence of even one faulty process [10]. These are "negative results" in a rather literal sense: they tell us what cannot be done. Yet we do not file them away as approaches that failed. We treat the boundary they establish as the contribution.

There is a reason why the same conclusion is harder to draw from an experiment. A proof establishes impossibility relative to explicit assumptions. When an experiment fails to produce an expected result, there is another possibility: perhaps the hypothesis is not wrong; perhaps the experiment was unable to tell. What looks like a refutation can instead be a limitation of the experiment.

This makes experimental False more demanding, not less valuable. To claim False, we need evidence not merely that we failed to establish the hypothesis, but that the evidence actually discriminates against it. Otherwise we have quietly crossed a boundary: from *the evidence shows this is wrong* to *the evidence does not let us determine whether it is right*. The latter is not False. It is **Unknown**.

3 Unknown Is Knowledge, Too

Unknown begins where the evidence stops discriminating. Suppose we ping a computer and, after waiting for some set time, receive no reply. We have learned very little about whether the computer is actually down: the host may be unreachable, a firewall may be dropping ICMP, the reply may have been lost, or the computer may indeed have failed. At the end of that period, concluding that the host is down would turn an inability to distinguish among these possibilities into a claim that one of them is true. The experiment has not established False. It has established that, from this observation alone, the alternatives are indistinguishable.

But we can say more than that. A retransmission would separate transient loss from the rest; a probe to a known-open TCP port would help distinguish ICMP filtering from host failure; a probe from a second vantage point would help distinguish a routing problem from a host problem. Even this small example has the structure the rest of this section is about: not merely that we cannot tell, but which additional observation might let us.

This distinction has a familiar echo in statistics: failure to reject a null hypothesis does not, by itself, establish that the null is true [4]. A nonsignificant result may simply mean that the available evidence cannot distinguish a meaningful effect from its absence. Methods such as equivalence testing make this distinction explicit: establishing that an effect is negligibly small requires different evidence from merely failing to detect it [16]. In our terminology, the latter is Unknown, not False.

But the point here is broader than statistical significance, and the reasons an experiment reaches Unknown are not all of a kind. It may lack data, in which case more of the same measurement would settle the question. It may be unable to discriminate no matter how much data it collects, because the alternatives induce the same observations under the measurement used; statistics calls this a failure of *identifiability*, and the binding constraint is the instrument rather than the sample. Or the quantity being inferred may not be well enough defined for any observation to bear on it, in which case what is needed is conceptual work rather than measurement.

Only the second kind yields a boundary worth reporting. Insufficient data is a fact about a particular study, and more of the same fixes it. Non-identifiability is a fact about what the measurement can see: no amount of further observation of the same kind will decide the question, while a different observation might decide it immediately. That is why it can be characterized, and why characterizing it is informative: it tells us not that we failed, but what kind of evidence the question actually requires.

Not every inconclusive experiment becomes a contribution by calling its outcome Unknown; discovering after the fact that a study was poorly designed is not a boundary, it is a mistake. The interesting case is when we can *characterize* the inability to decide.

This idea has a long history in Internet measurement. Paxson argued that sound measurement requires a "solid understanding of the strengths and limitations of the measurement process" [22]. Measurement methodology has consequently devoted considerable attention to bias, calibration, vantage-point selection, and the limitations of the signals from which we infer properties of the Internet [21, 23]. The point has a counterpart in the philosophy of experiment: Mayo's error-statistical account treats the capacity of a test to have detected an error (its *severity*) as a precondition for reading anything into its outcome, rather than as a caveat attached afterwards [20]. The step I want to emphasize is to treat a sufficiently characterized limitation not merely as a caveat on a result, but sometimes as the result itself.

This changes the meaning of Unknown. It is not a generic statement that "we do not know." It is a more specific claim:

> Given these observations, assumptions, and this method, the available evidence does not distinguish sufficiently among the relevant alternatives.

Unknown is also relative to the question being asked. Evidence that cannot distinguish among several alternatives may still support a coarser inference that does not require distinguishing them. If we are trying to chart which submarine cables a traceroute crosses [27], for example, evidence that leaves a path able to traverse either of two cables makes the particular cable Unknown, while its dependence on a landing point shared by both may still be established. More generally, uncertainty in an intermediate inference need not make a downstream conclusion Unknown if that conclusion holds across the alternatives that remain.

Such a claim is auditable. Another researcher can ask why the method abstained, whether the alternatives really are indistinguishable under the available evidence, and what additional observation would separate them.

Observability is heterogeneous [24, 25]: measurements from different vantage-point populations can expose different paths, performance, and behavior, and a method may consequently have strong discriminating power for one part of the Internet and almost none for another. Reporting only the cases in which it produces an answer can make both the method and the Internet look more uniformly measurable than they really are.

The distinction between False and Unknown is therefore not semantic. False warrants a conclusion about the proposition being tested; a characterized Unknown warrants a conclusion about the experiment's ability to decide that proposition. Seen this way, a characterized Unknown is not really a third truth value about H at all. It is a True, one level up: a positive and defensible claim about what the instrument can and cannot resolve.

And this makes Unknown surprisingly constructive. False tells us where to stop: an explanation has been ruled out. Unknown can tell us where to look next. It can identify the observation, vantage point, ground truth, or experimental capability that would make the alternatives distinguishable.

4 The Damage from Getting It Wrong

Why insist on this distinction? Because confusing Unknown with False does more than mislabel an outcome. It changes what we believe is worth pursuing.

An Unknown dressed as False closes a question that the evidence left open. If an experiment cannot distinguish among several explanations, concluding that one of them is false turns a limitation of the evidence into a claim about the world. We lose twice: we make a claim stronger than the evidence supports, and we discard the information contained in the failure to distinguish—information that might have told us what to measure next.

Internet measurement provides many opportunities for this mistake. Consider traceroute. Its limitations are well known. Routers may not respond to probes [14]; load balancing can produce artifacts in the inferred path [5]; different probing methods can expose different paths and topology [19]; and routing asymmetry complicates the interpretation of both paths and RTTs [11].

These limitations can motivate a tempting conclusion: if traceroute does not faithfully reveal the physical path, then trying to map traceroutes onto physical infrastructure is fundamentally unreliable and perhaps not worth doing.

But that conclusion moves too quickly from Unknown to False. The fact that *some* traceroutes do not contain enough information to determine their physical realization does not establish that *no* traceroutes do. Nor does it establish that we cannot determine which is which.

The useful question is therefore not *can traceroute reveal the physical Internet?* but *for which paths, under what observations, and with which supporting metadata does traceroute contain enough evidence to support a particular physical inference?* The first invites a binary answer: the method works or it does not. The second makes the boundary itself an object of study.

Some paths may strongly constrain the possible physical interpretations; others may admit many explanations that the available measurements cannot distinguish. And the same path may fall on either side of that boundary depending on the metadata it is interpreted against. If we can tell those cases apart, Unknown has not ended the investigation. It has told us where the investigation must change.

This is also why simply listing limitations at the end of a paper is not enough. A limitations section says that there are circumstances under which our conclusions may not hold. What we would like, when possible, is stronger: for the inference itself to expose when the available evidence has ceased to be discriminating. Unknown then becomes an output of the measurement process, rather than a qualification added after the answers have already been produced.

The question is whether we can actually build a measurement method that does this.

5 A Worked Example: When Can a Path Tell?

Consider a problem we have been working on: mapping Internet paths onto physical infrastructure. A traceroute gives us a sequence of router interfaces and round-trip times (RTTs), but many questions we would like to ask—which cities a path traverses, which facilities or submarine cables carry it, or which physical failures could affect it—require placing those routers in geographic space.

There is no shortage of candidate answers. Commercial geolocation databases, reverse-DNS names, geofeeds, and other sources can suggest locations for a router. The problem is that giving us a location is not the same as giving us evidence that the location is right. That distinction matters when the sources disagree. If one database says Miami and another says São Paulo, absent ground truth, which should we believe?

The traceroute itself provides a useful constraint. Suppose two consecutive routers are placed in São Paulo and Miami, while their observed RTT increment is only 3 ms. The two cities are some 6,500 km apart; at the effective propagation speed of light in fiber, a round trip over that separation requires more than 60 ms, twenty times what we observed. Whatever the true locations of those routers may be, that particular interpretation cannot be right. We did not need to know the correct location of either router to establish this. The proposed interpretation contained enough evidence to falsify itself.

This asymmetry is useful. Confirming a router location generally requires some form of ground truth; ruling out an impossible location may require only a physical constraint. Falsification is in that sense *cheaper* than confirmation, though not free: an RTT increment constrains propagation only if the forward path traverses both hops in order, the return paths are comparable, and neither measurement is dominated by ICMP generation delay. The São Paulo–Miami case survives because the margin is an order of magnitude; a violation of a few milliseconds is not a refutation but a measurement we cannot interpret. Even the cheap False rests on auxiliaries that can fail.

Can we generalize that idea from one obviously impossible pair of locations to an entire path? Path consistency scoring does this with a simple model: each hop carries a set of candidate locations drawn from metadata, and the observed RTT increments constrain which transitions between consecutive candidates are physically possible. Decoding a path means choosing the sequence of candidates that best satisfies those constraints; the path consistency score (PCS) summarizes how well even that best sequence explains the evidence [15]. The important point here is not the machinery. It is the change in question.

Rather than asking whether each inferred router location is correct, we now ask whether the *path as a whole* is consistent with the evidence already contained in the measurement. A poor PCS means that even the best geographic interpretation we can construct has difficulty explaining what we observed. We have generalized the São Paulo–Miami consistency check from a pair of points to a path.

But there is a catch. RTT is not geographic distance. Queueing, asymmetric routing, ICMP processing, hidden path structure, and other effects can weaken the relationship between an RTT increment and the physical separation of two visible hops. A poor

PCS can therefore have at least two explanations: the proposed geography is inconsistent with informative latency evidence, or latency itself is not sufficiently informative on this path. Only the first supports a geographic refutation. Without distinguishing the two, PCS risks turning Unknown back into False.

This motivates a second quantity, R, which asks whether the latency evidence was doing any work in the first place. For each transition we compare the propagation delay implied by a geographic interpretation with the observed RTT increment, producing a sequence of latency–distance residuals along the path. R measures how closely those residuals agree under two interpretations: the decoded path, and a reference path taken here to be the raw database labels before decoding. When R is high, the two interpretations locate latency–distance tension in the same places, and PCS can be read as a statement about the geography. When R is low, they do not, and PCS should be read instead as a warning about whether the model applies to this path at all.

Two cautions follow from what R compares. It is a property of a path *and* a metadata source, not of the path alone: the same traceroute can score near one against one database and near zero against another, because R partly records how far decoding had to move the labels it started from. A path that scores low against every source is a far better candidate for Unknown than one that scores low against a single vendor. And low R has two explanations of its own—latency that is genuinely uninformative, or a database that was wrong and duly corrected—of which only the first is a limit on what the measurement can tell us. That case splits again along Section 3's lines: noise on a single traceroute is a lack of data, while latency that cannot resolve the alternatives however often we probe is the non-identifiability worth reporting. R separates neither pair. The diagnostic introduced to separate PCS's failure modes has failure modes of its own, which is the argument of this article applied to its own example.

The combination gives us something qualitatively different from a confidence score attached to every answer. When the evidence is informative, PCS can support a verdict about the geographic interpretation. When it is not, the method can abstain. It can say, in effect: *from this traceroute, with these observations, I cannot distinguish the relevant geographic alternatives.*

That is Unknown made operational. The system does not infer a location and then append a generic caveat that traceroute is imperfect. It attempts to identify, path by path, whether the evidence has earned the right to support the inference at all. The boundary that was abstract in Section 3 has become something we can measure.

6 Unknown Can Be Useful

Making Unknown explicit does more than prevent unsupported conclusions. Once we can characterize where a method's evidence ceases to be informative, that boundary can help us decide how to use scarce validation resources.

Returning to the geolocation example, suppose we have a large corpus of inferred paths but ground truth for only a small subset, a common situation in Internet measurement. The obvious use of a ground-truth label is to verify one inference: compare the inferred path with the known answer and determine whether it is correct. With a budget of B labels, this gives us B directly verified cases.

But those same labels can do more. For every ground-truthed path, we can also observe R and determine whether the PCS-based judgment was correct. Across the labeled sample, we can therefore estimate how the reliability of the judgment varies with R. Ground truth still verifies individual paths, but the verification outcomes now also *calibrate the boundary*: they tell us when the method's own reliability signal is actually predictive of error. Because R compares the decoded path against the database labels rather than against a verified one, it can be computed where no ground truth exists, and that calibration can then be applied beyond the labeled sample.

This creates what I will call a *validation dividend*. Direct validation uses a label to justify a verdict about the object that was labeled. Calibration uses a collection of labels to justify a rule about when the method's evidence is sufficient, and that rule can cover many unlabeled objects.

The premise can already be checked. Computed against the raw database labels—the form available without ground truth—R separates error regimes on our validated corpus: across four commercial databases, paths with $R \geq 0.8$ have a median per-path error of 3.4–5.7 km, while the $R < 0.2$ bin carries a far heavier tail [15]. That is what makes R a candidate reliability signal rather than a decoration. The same separation appears in the cleanest setting we can construct, where every candidate comes from public metadata and the comparison is made against an actively validated path rather than a database (median per-path error of 4.0 km at $R \geq 0.8$ against 244.7 km at $R < 0.2$). But that variant consumes the ground truth it would otherwise calibrate: it shows that the metric is sensitive to the right thing, not that it can be used where labels are absent.

Two caveats travel with the usable form. R is high in part when the database was already right, so the signal is confounded with source quality; and diagnosing *why* R is low is not something the metric itself does. The dividend (how much of an unlabeled corpus a calibrated rule can cover as the label budget shrinks) follows from the premise, but we have not measured it, and how large it is remains an open question.

The basic idea is not new: use labeled data to determine when an inference is sufficiently reliable to report. Selective prediction, learning with a reject option, and conformal prediction all embody versions of this idea [8, 26]. What makes the measurement setting harder is the assumption needed to transfer what is learned from labeled cases to the unlabeled population. Conformal coverage, for example, relies on exchangeability between the calibration data and the population to which the guarantee is applied [26]. In Internet measurement, that assumption is particularly suspect. Ground truth comes disproportionately from responsive routers in well-covered regions, precisely where we may need it least. A calibration learned there has not been shown to hold in poorly observed regions, where the boundary matters most.

Calibration does not manufacture ground truth. The boundary itself has to be tested, not asserted, and Unknown shows where that testing is still owed.

This suggests a second use that we are exploring: let Unknown direct the measurement process itself. Measured broadly from the existing vantage points of a platform such as RIPE Atlas, paths whose R stays low against every metadata source mark places where latency cannot adjudicate among the geographic alternatives—and

where a poor PCS therefore tells us nothing we can act on. Rather than simply reporting those paths as uncertain, a measurement system could ask a more useful question: *what observation would most reduce this uncertainty?* The answer depends on which branch of the Unknown we are in. If latency is genuinely uninformative on a path, the remedy is a different vantage point: probes retasked or deployed not according to geographic or AS coverage, but where they are expected to make currently indistinguishable alternatives distinguishable [6]. If instead the metadata was wrong, no additional probe helps, and what is needed is better evidence about the locations themselves—operator-published geofeeds, for instance. Separating the two is not something R currently does, which makes it the first thing the boundary tells us to work on.

This admits a clean evaluation: withhold probes from a well-covered platform, then ask whether restoring them according to the resulting Unknowns recovers inferential coverage faster than restoring them at random or by conventional coverage criteria. The test is not whether the signal rediscovers the probes that were removed, but whether it identifies the measurements whose return most reduces the uncertainty their absence created.

The broader point is that Unknown need not mean *stop*. Once the boundary of an experiment is observable, it can become an input to the next experiment.

7 Unknown Has Limits, Too

There is a fundamental limit to this argument. A method can expose its uncertainty only when something in the available evidence reveals that the inference should be questioned. Internal consistency is not correctness.

Geolocation shows it. R measures agreement between two interpretations, so by construction it cannot detect cases where both are consistently wrong [15]. A router placed wrongly but consistently with its observed latencies and neighbors yields a high PCS and an R saying latency is informative. The system can be calibrated and wrong at once.

This is *coherent error*: observations, assumptions, and inference agree with one another and are wrong together. It is structural rather than accidental. A hypothesis is never tested alone but together with auxiliary assumptions about the instrument, the sample, and the model, and evidence bears on that conjunction—the Duhem–Quine problem, and the same bundle that made falsification cheap rather than free in Section 5. So more observations from the same process need not help: if they share the source of the error, they reinforce it.

Internet measurement offers an instance. In 1999, Faloutsos et al. reported power-law relationships in measured Internet topologies [9], which became part of a "scale-free" interpretation in which a few highly connected hubs left the network robust to random failure but vulnerable to attack [3]. The observation was partly an artifact: traceroute does not sample the Internet uniformly, and traceroute-like sampling can produce an apparent power law where the underlying graph has none [1, 17]. Because that bias was shared by every traceroute-derived topology, more measurement made the pattern look more solid rather than less.

The episode went further, and this is the harder limit. The inference carried a second defect independent of the data: a degree distribution does not determine a graph's structure, so an architectural claim was drawn from a statistic that cannot support one [7, 18]. That is not evidence agreeing with itself and being wrong; it is a question posed in terms no available evidence could settle. The coherence held only inside a frame that treated the traceroute graph as the network and the degree sequence as its structure—a frame built, used, and eventually dismantled inside the measurement community itself, and dismantled not by more data but by a different account of what a router-level topology is. Nothing looked Unknown from within it, and that is the more troubling observation.

This places a boundary around Unknown itself. A consistency check can only see the alternatives its model represents; a frame is a candidate set. PCS can tell us that a proposed geography is inconsistent with the latency evidence, but not that a router location is the wrong quantity to infer in the first place. An MPLS tunnel may hide its interior hops entirely, or have replies for them generated at the tunnel egress, so that several consecutive hops show RTT increments near zero and appear to sit in one place. The check scores that as a geography, and may score it well. It has no candidate for "these hops are not separate locations at all" because that alternative is not in its set. This is the third case from Section 3—a quantity not well enough defined for observation to bear on it—in a harder form, hidden now by the method's own candidate set. This does not make a characterized Unknown worthless; it says what kind of claim it is. Naming that set, and the evidence that failed to separate it, puts the frame on the page where others can contest it. **Characterizing where evidence stops discriminating protects us only inside a frame. Nothing inside a frame can tell us the frame is wrong.**

8 Conclusions

Research is an investigation; a paper is necessarily a compression of it. The problem is not only that negative results disappear in that compression, but that False and Unknown are compressed into the same thing. False eliminates possibilities. A well-characterized Unknown establishes where the available evidence ceases to discriminate among them, and that boundary can prevent unwarranted conclusions, identify where better evidence is needed, and, as the geolocation example illustrates, help direct scarce validation.

Taking Unknown seriously asks something of both authors and reviewers. We should not expect every investigation to end with a binary answer, nor treat an inability to reach one as evidence that the approach "didn't work." A paper may make a scientific contribution by establishing that a question cannot be answered from particular observations or under particular assumptions, provided that boundary is itself supported by evidence.

Publication norms can help make such outcomes visible. One suggestion from discussions following the talk that motivated this article was a dedicated *Boundaries and Limitations* appendix outside the main page limit: not another generic limitations section, but a place to state what the evidence can and cannot support, which alternatives remain indistinguishable, and which conclusions depend critically on particular assumptions.[1] SIGCOMM's current Proposals for Change initiative [2] provides a natural opportunity to

[1]The idea was proposed by Loqman Salamatian.

experiment with mechanisms like this. Other empirical disciplines already go further: the CONSORT statement, which governs the reporting of randomized trials in medicine, makes trial limitations, including the generalizability of findings, a required reporting item rather than an optional courtesy [13].

None of this eliminates uncertainty. Coherent error reminds us that evidence can be internally consistent and still systematically misleading; independent evidence remains essential. Making Unknown explicit does something more modest: it stops us from turning limits we *can* identify into conclusions the evidence never established. A paper need not answer every question it asks. Sometimes the result is True. Sometimes it is False. And sometimes what we have learned is where our ability to tell the difference ends.

That boundary is the result.

Acknowledgments

I thank Ioana Livadariu and Mattijs Jonker for the invitation, and workshop participants, Marwan Fayed, Walter Willinger, David Choffnes, and Theo Benson for discussions that sharpened the argument. As always, I am grateful to my students, whose questions and skepticism continually shape how I think about research.

References

[1] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. 2009. On the Bias of Traceroute Sampling; or, Power-Law Degree Distributions in Regular Graphs. *J. ACM* 56, 4 (2009). https://doi.org/10.1145/1538902.1538905

[2] ACM SIGCOMM. 2026. Proposals for Change. https://www.sigcomm.org/initiatives/changes. Accessed: 2026-08-24.

[3] Réka Albert, Hawoong Jeong, and Albert-László Barabási. 2000. Error and Attack Tolerance of Complex Networks. *Nature* 406, 6794 (2000), 378–382. https://doi.org/10.1038/35019019

[4] Douglas G. Altman and J. Martin Bland. 1995. Absence of Evidence is not Evidence of Absence. *BMJ* 311, 7003 (1995), 485. https://doi.org/10.1136/bmj.311.7003.485

[5] Brice Augustin, Xavier Cuvellier, Benjamin Orgogozo, Fabien Viger, Timur Friedman, Matthieu Latapy, Clémence Magnien, and Renata Teixeira. 2006. Avoiding Traceroute Anomalies with Paris Traceroute. In *Proc. of IMC*. https://doi.org/10.1145/1177080.1177100

[6] Paul Barford, Azer Bestavros, John Byers, and Mark Crovella. 2001. On the Marginal Utility of Network Topology Measurements. In *Proc. ACM IMW*. https://doi.org/10.1145/505202.505204

[7] John C. Doyle, David L. Alderson, Lun Li, Steven Low, Matthew Roughan, Stanislav Shalunov, Reiko Tanaka, and Walter Willinger. 2005. The "Robust Yet Fragile" Nature of the Internet. *Proc. of the National Academy of Sciences (PNAS)* 102, 41 (2005), 14497–14502. https://doi.org/10.1073/pnas.0501426102

[8] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-Free Selective Classification. *Journal of Machine Learning Research* 11 (2010).

[9] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On Power-Law Relationships of the Internet Topology. In *Proc. of ACM SIGCOMM*. https://doi.org/10.1145/316188.316229

[10] Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. 1985. Impossibility of Distributed Consensus with One Faulty Process. *J. ACM* 32, 2 (1985), 374–382. https://doi.org/10.1145/3149.214121

[11] Romain Fontugne, Cristel Pelsser, Emile Aben, and Randy Bush. 2017. Pinpointing Delay and Forwarding Anomalies using Large-Scale Traceroute Measurements. In *Proc. of IMC*. https://doi.org/10.1145/3131365.3131384

[12] Seth Gilbert and Nancy Lynch. 2002. Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services. *SIGACT News* 33, 2 (2002), 51–59. https://doi.org/10.1145/564585.564601

[13] Sally Hopewell, An-Wen Chan, Gary S. Collins, et al. 2025. CONSORT 2025 Statement: Updated Guideline for Reporting Randomised Trials. *BMJ* 389 (2025), e081123. https://doi.org/10.1136/bmj-2024-081123

[14] Hakan Kardes, Mehmet Hadi Gunes, and Kamil Sarac. 2015. Graph Based Induction of Unresponsive Routers in Internet Topologies. *Computer Networks* 81 (2015), 178–200. https://doi.org/10.1016/j.comnet.2015.02.012

[15] Santiago Klein, Caleb J. Wang, and Fabián E. Bustamante. 2026. Overconfident Coordinates: Quantifying Confidence in Traceroute Geolocation. arXiv:2606.24027 [cs.NI] https://arxiv.org/abs/2606.24027

[16] Daniël Lakens, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 1, 2 (2018), 259–269. https://doi.org/10.1177/2515245918770963

[17] Anukool Lakhina, John W. Byers, Mark Crovella, and Peng Xie. 2003. Sampling Biases in IP Topology Measurements. In *Proc. of IEEE INFOCOM*. https://doi.org/10.1109/INFCOM.2003.1208685

[18] Lun Li, David Alderson, Walter Willinger, and John Doyle. 2004. A First-Principles Approach to Understanding the Internet's Router-Level Topology. In *Proc. of ACM SIGCOMM*. https://doi.org/10.1145/1015467.1015470

[19] Matthew Luckie, Young Hyun, and Bradley Huffaker. 2008. Traceroute Probe Method and Forward IP Path Inference. In *Proc. of IMC*. https://doi.org/10.1145/1452520.1452557

[20] Deborah G. Mayo. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

[21] Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. 2008. In Search of the Elusive Ground Truth: The Internet's AS-level Connectivity Structure. In *Proc. of ACM SIGMETRICS*. https://doi.org/10.1145/1384529.1375482

[22] Vern Paxson. 2004. Strategies for Sound Internet Measurement. In *Proc. of IMC*. https://doi.org/10.1145/1028788.1028824

[23] Matthew Roughan, Walter Willinger, Olaf Maennel, Debbie Perouli, and Randy Bush. 2011. 10 Lessons from 10 Years of Measuring and Modeling the Internet's Autonomous Systems. *IEEE Journal on Selected Areas in Communications* 29, 9 (2011).

[24] Mario A. Sánchez, John S. Otto, Zachary S. Bischof, David R. Choffnes, Fabián E. Bustamante, Balachander Krishnamurthy, and Walter Willinger. 2013. Dasu: Pushing Experiments to the Internet's Edge. In *Proc. of USENIX NSDI*.

[25] Pavlos Sermpezis, Lars Prehn, Sofia Kostoglou, Marcel Flores, Athena Vakali, and Emile Aben. 2023. Bias in Internet Measurement Platforms. In *Proc. of TMA*. https://doi.org/10.23919/TMA58422.2023.10198985

[26] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. 2019. Conformal Prediction Under Covariate Shift. In *Proc. of NeurIPS*.

[27] Caleb J. Wang, Ying Zhang, Qianli Dong, Esteban Carisimo, Ramakrishnan Durairajan, and Fabián E. Bustamante. 2026. Calypso's Voyage: Charting Traceroute Paths Through Submarine Cables. In *Proc. of IMC*.