CHALLENGES IN MEASURING THE INTERNET FOR THE PUBLIC INTEREST

kc claffy and David Clark

ABSTRACT

The goal of this article is to offer framing for conversations about the role of measurement in informing public policy about the Internet. We review different stakeholders' approaches to measurements and associated challenges, including the activities of U.S. government agencies. We show how taxonomies of existing harms can facilitate the search for clarity along the fraught path from identifying to measuring harms. Looking forward, we identify barriers to advancing our empirical grounding of Internet infrastructure to inform policy, societal challenges that create pressure to overcome these barriers, and steps that could facilitate measurement to support policymaking.

Keywords: Internet, measurement, infrastructure, security, public

Motivation

Society is at a turning point with respect to the Internet. Since the mid-1990's, the private sector has led development of the Internet. Now, as the Internet has become critical infrastructure, public policy concerns are becoming more visible and important. The Internet today exposes users to a range of harms, including those arising from limitations of its network architecture, poor security practices, performance impairment, consolidating industry structure, and the digital divide, many of which are serious enough to create a public interest in mitigating them. Such mitigation

kc claffy: University of California San Diego, US
David Clark: Massachusetts Institute of Technology, US
https://doi.org/10.5325/jinfopoli.12.2022.0003

This material is based on research sponsored by the National Science Foundation (NSF) grants OAC-1724853 and OAC-2131987. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF.

is best initiated with a thorough understanding of the harms, including their scope, extent, and operational contexts in which they arise. The rising influence of adversarial actors is increasing the urgency for a more rigorous understanding of the Internet than today's ecosystem allows.

Government oversight plays an important role in other areas critical to society, such as health care, transportation, ocean and atmosphere, food and drugs, and traditional telecommunications. An important part of this oversight is gathering data to understand how each system is working. However, because there is no agency (or coordinated set of agencies) that have responsibility for oversight of the Internet, the governmental role in data collection is fragmented, and data collection efforts have not achieved even the limited goals that government has established for them.

While private sector actors, including network operators, collect data on the part of the Internet for which they are responsible, the lack of unbiased, neutral data deprives operators, policy makers, scientists, and citizens of a consensus view of the Internet to drive decision-making, or understand the implications of current or new policies or technologies. The quality and quantity of independent research on Internet infrastructure is deeply impaired by a lack of access to relevant data. This situation will get worse, and the lack of rigorous scientific research on properties of the Internet related to public policy objectives will grow even more problematic, as the Internet continues to be ever more deeply embedded in our lives. Examples of such properties include: broadband availability and cost; metrics of resilience; traffic patterns and associated performance impairments; industry structure; privacy breaches; cybercrime; malicious activity that is not (yet) illegal; and associated harms. The goal of this paper is to offer framing for conversations about the role of measurement in informing public policy about the Internet, the barriers to gathering measurements, public policy challenges that are creating pressure for reform in this space, and recommended actions that could facilitate gathering of measurements to support policymaking. We emphasize that there is no simple answer to the questions, "What data should be gathered about the Internet, how should it be gathered, and what policy frameworks should govern its management and sharing?" Our point is that different stakeholders are increasingly asking a form of this question, and we propose a more systematic framework to support, discover, and/or enable consistency across these conversations.

Existing Sources of Measurement

In this section, we review various stakeholders who participate in or induce production of data about the Internet infrastructure, and the challenges they face. We devote the section "The Quality of Currently Available U.S. Governmental Data" to a more in-depth look at various U.S. government activities that involve gathering and publishing measurements of Internet infrastructure.

Private Sector Data Collection

In most of the world, the Internet infrastructure is the product of the private sector. Economic considerations that drive the private sector shape the Internet's structure and operational capabilities, and the incentives of the private sector do not facilitate, or sometimes allow, independent scientific study of the infrastructure. Network operators collect substantial data on their own networks, but typically with a narrow focus and almost always limited availability and corporate interest in the messaging, which leads to concerns about bias in the reported data. Since any such collected data may reveal aspects of business practices, private sector actors do not voluntarily share such data unless it is in their interest to do so. Sometimes a one-time data-sharing agreement with a commercial firm allows a researcher access to specific data to perform collaborative research and report important findings, if approved by the firm. But if the researcher cannot further share the data for replication of results, these arrangements trigger concerns regarding scientific objectivity.

Data Collection by Independent Third Parties

Independent third parties, typically government-funded academic researchers, can measure the Internet from the edge, analyze the measurements, draw their own conclusions, subject these to comparison and peer review, and present results as empirical grounding for policy debates.

Traditionally, the network measurement community has used two methods to gather data: active probing of networks, and passive observation of traffic, which requires deploying monitoring instrumentation at a point in a network where it can observe traffic. Both methods have significant limitations. Active probing from the edge allows for only *inference* of properties or behavior, not direct assessments. Also, while academics can

design international measurement campaigns, the size and diversity of the Internet is a daunting challenge. At what point can a researcher conclude that they have a representative view of the Internet or even a *region* of the Internet?

The set of networks in which a researcher can acquire even temporary vantage points is a small fraction of the networks in the Internet, and not where most traffic originates. Modern applications typically depend on platform elements in addition to the public Internet, such as cloud facilities and content distribution networks (CDNs). CDNs operate rich networks of servers, often using anycast or DNS-based network traffic redirection. Such servers are often hosted in third-party networks, partially masking the CDN's presence from observation. It is difficult and sometimes impossible to attach probing points onto those elements. Probing may be prohibited by terms of service, cost, or technical factors. The result is a growing scope of assets on which application designers (and users of those apps) depend that are beyond the scope of independent measurement.

Passive monitoring raises serious concerns about privacy, both of individuals communicating across the network, as well as of the operator of the network. There is no incentive for a commercial network operator to let any unaffiliated party gather data from its network. Sometimes it is illegal to do so, but even if legal barriers are overcome, there is always a risk that data related to a provider's service offering can shed light on aspects of that service that might reflect poorly on that operator or otherwise reveal something the provider may want to keep secret.

If the privacy barriers can be overcome, a further challenge for passive monitoring is cost: the increasing bandwidth of today's network links has rendered it technically challenging and prohibitively expensive to collect and store network traffic traces for scientific research.

Another approach to independent data collection is to develop and operate platforms that gather data about the infrastructure by acting as a network operator, for example, a route collection effort such as the Route Views project,[1] which peers with hundreds of networks to capture global routing topology information, and store and share longitudinal data. For the Domain Name System, some independent researchers have systematically downloaded Top Level Domain (TLD, e.g., .com) files and stored

1. http://www.routeviews.org/routeviews/

them in a database for web-based querying.[2] The challenge of these efforts is a consistent source and level of funding to sustain them. To gather measurements for scientific study, academics must obtain funding for data collection from organizations such as the National Science Foundation, which have research as their mission, not data collection. Maintaining measurement infrastructure over time is not possible with the typical approach of funding research projects in 3-year increments.

The daunting barriers to collecting data needed for scientific research on Internet infrastructure motivated the National Science Foundation to sponsor two Workshops on Overcoming Measurement Barriers to Internet Research (WOMBIR) in 2021.[3] Barriers discussed at that workshop include evolution of industry structure, lack of capital and incentive for (especially longitudinal) data collection, privacy implications, and limitations of current ethical review mechanisms. The final report (which we co-authored) has more details on these and other topics discussed at the workshop.[4]

Another approach to gathering data on the Internet—and in particular on the human experience of using it—is by surveying a sample of the population. Aside from the U.S. Census, which does survey a sample of the population (see the section "The Quality of Currently Available U.S. Governmental Data"), the most significant survey data on Internet users (and non-users) and their usage is from the Pew Research Center.[5]

Government Data Collection

Government data collection is fraught with difficulties. The purpose and cost must be well-understood by all stakeholders. The government must ensure proper design of the network measurement methodology so that it produces evidence responsive to the inquiry, and proper evolution of a measurement program to retain relevance in the face of technological and societal change, or sunset as appropriate. This design process must also

2. See, for example, http://dns.coffee, which provides archives of DNS zone files. This site is now hosted at CAIDA as https://dzdb.caida.org/, because the student who started the project could not sustain the effort to collect the data.

3. CAIDA. NSF-sponsored Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021)—Parts I and II.

4. Claffy et al. We have included some of our thoughts from that report into this study.

5. https://www.pewresearch.org/topic/internet-technology/. A recent report: Andrew Perrin, Mobile Technology and Home Broadband 2021, Pew Research Center, https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/. URLs accessed March 12, 2022.

consider incentives of network providers to game the system in ways that distort the results in their favor or provide data in a form that resists sound interpretation. In the case of the Internet infrastructure, measurement is a highly technical undertaking, and most government agencies do not have the skills or capacity for the measurement itself much less interpretation of the resulting data. The transnational reach of much Internet infrastructure also complicates the government's role. Most U.S. government data collection about the Internet occurs in classified contexts, either intelligence gathering or law enforcement, and in neither case is associated data available for scientific research. In the United States, no governmental organization has sustaining Internet measurement for scientific research as part of its mission. This lack of government engagement contrasts with other critical infrastructures or substrates critical to society (oceans, atmosphere, and highways), where government agencies have divisions chartered and authorized to undertake measurement and data collection as part of obligations to support the public interest mission of the agency, for example, NOAA, Department of Transportation, National Institutes of Health, Center for Disease Control, the FAA, or the Bureau of Labor Statistics (BLS). In the United States, the lack of any government agency with oversight of the Internet has led to a fragmented and ineffective approach to measurement, which we detail in the section "The Quality of Currently Available U.S. Governmental Data."

Mandatory Data Reporting

In many sectors, governments collect data by requiring that private sector parties provide it. However, compelled reporting of data, and its analysis, comes at considerable cost to all parties. Such efforts thus require much stronger justification than satisfying scientific exploration. The Paperwork Reduction Act of 1995 further raised the bar for mandating data reporting by the U.S. government, requiring a justification of its need and cost/benefit analysis to the U.S. Office of Management and Budget (OMB). The section "The Quality of Currently Available U.S. Governmental Data" provides a case study that faced this requirement.

Even if justified and executed, government-mandated data is often encumbered, its collection and use limited to purposes established by authorizing legislation, and thus not available for scientific use by independent third parties. This difference in activation thresholds between regulatory oversight and scientific inquiry poses a formidable challenge,

because the 25-year gap in regulatory oversight has led to calls for objective scientific input to public policy debates, which is typically a role for (at least) academic researchers. But if the data collected by government is restricted to use by the government agency that collects it, there may be no way to share the data with academics.

If a government compels the collection of data, the results will normally be a view of the domestic Internet in that country. To get a larger view, some international entity will have to attempt the aggregation of this data and will encounter challenges of inconsistent collection practices and requirements, misalignment of the resulting data, and variable attention in different countries to the validity and accuracy of the data. Entities like the International Telecommunication Union and the Organization for Economic Co-operation and Development face such issues when they attempt to distill disparate data sources from many countries into a coherent aggregated report.

Challenges Spanning Most Internet Measurement Activities

Regardless of who is undertaking measurements, technical, cost, policy, and ethical challenges pose formidable barriers. One additional factor that distinguishes this critical infrastructure from others is the enormous data volumes required for any measurement that examines actual data flows. Any such measurements must also navigate regulatory and ethical guidelines governing data privacy, which require both legal and technical expertise to interpret. The guidelines and laws vary by country, which makes any global Internet measurement science program a daunting proposition.

The Murky Path from Identifying Harms to Measuring Them

To argue that there is a public interest that requires measurement, we start with the premise that the public interest is not being well-served—that there is *harm* to the public interest, either material actual harm or demonstrable reason to believe that significant harm might occur. To explore a persistently turbid space, we start with a high-level discussion of harm, and then explore the path from that high-level view to a focus on specific actions that can mitigate the harms. This path is complex and hard to quantify, especially when private sector stakeholders face so many counter-incentives to voluntary transparency.

Our perspective derives from our background in designing, operating, and studying the network infrastructure layers of the Internet. The metrics it is possible to capture at these layers are only part of the total picture of Internet-related harms. A broader view is critical in motivating a plan of action to mitigate the harms, but we start with a smaller scope: what are the important metrics about the Internet infrastructure itself, and what are the barriers to obtaining measurements to derive those metrics? If we cannot make progress at these layers, which underlie all activity on the Internet, optimism for progress against a broader set of Internet harms is unjustified.

We start with this high-level definition of harm offered in 1978 by criminal justice philosopher John Kleinig in the *American Philosophical Quarterly*: "*an impairment – either with respect to an individual, a firm or society – to an entity's welfare interests, relative to the normal expectations of the time and context.*"[6] What constitutes a harm will evolve over time. For example, if we consider that lack of adequate access to the Internet is a harm, then what constitutes "adequate" access (e.g., what access speeds) will change over time. The goal of this definition of *harm* is to provide a framework to illuminate interactions and tradeoffs—conflicting articulations of welfare interests—in attempting to mitigate any specific harm. The challenge in mitigating many harms in today's Internet is establishing the logic that moves from the abstract concept of harm to more concrete behaviors that can be measured. The difficulty constructing and quantifying that logic leads to persistent disputes about what operational changes will mitigate the actual harms.

An Overview of Harms

We briefly review the broad sweep of harms identified in the Internet, based on several independent taxonomies of harms. We then focus on two harms that are attracting a lot of attention today: lack of (or poor) broadband access; and inability to systematically assess infrastructure resilience. These examples map reasonably well to network-layer phenomena, and thus network-layer measurements, so we can leverage our own expertise.

6. Kleinig, "Crime and the Concept of Harm," American Philosophical Quarterly 15, no. 1 (1978).

Harms to a firm: One of Kleinig's categories was harms to a firm. Many harms to firms today involve cybersecurity failures. In 2018, Agrafiotis et al.[7] identified five classes of harms to firms:

- Physical or digital harm (e.g., damage to intellectual or real property)
- Economic harm (e.g., disrupted operations, profit, growth, investment)
- Psychological harm (e.g., confusion, anxiety, distress, grief, fear)
- Reputational harm (e.g., reduced public perception, loss of goodwill, relationships, staff)
- Social and Societal harm (e.g., negative impact on society)

Economic harm is potentially the easiest to quantify, and thus a natural object of focus, but is only one of several classes of harms.

McAfee and CSIS[8] in 2020 looked at cybersecurity related harms to the firm—harms that they classified as *cybercrime.* Based on interviews with 1,500 businesses, selected to give them a representative sample, they concluded that the global cost of cybercrime in 2020 was approximately $945B. Adding the spending on cybersecurity, which they estimated at $145B, the total costs were well over a trillion dollars. Their survey asked not only about the direct financial costs, but opportunity costs, costs induced by downtime, reduced efficiency, brand damage, loss of intellectual property (IP), morale, and incidental costs related to mitigation, including costs of reporting and insurance. Although this study lacked a rigorous measurement framework, these numbers seem large enough to warrant attention.

Harms to the individual: Clark and Claffy[9] identified a broad sweep of harms to individuals. One that has received much attention is harm from lack of any (or of adequate) access to the Internet, combined with lack of skills and financial barriers to using it. The response to COVID, with the need for work or schooling at home, amplified the negative consequences of the inability to access or use the Internet.

Another set of harms to the individual are financial, and fall into the diffuse "cybersecurity" bucket. In the United States, the Federal Bureau of Investigation (FBI) and the Federal Trade Commission (FTC) report the number and magnitude of reports they receive from citizens that have suffered financial loss. In 2020, the FBI received 791,790 complaints, with reported losses exceeding $4.1 billion.[10] The FTC similarly provides an

7. Agrafiotis et al.
8. Smith and Lostri.
9. Clark and Claffy.
10. Federal Bureau of Investigation.

annual report summarizing the types of complaints they receive.[11] In 2020, they received 1.39M reports of identity theft, and almost 500k reports of imposter scams. They received 2.18M complaints that involved fraud, including categories such as imposter scams, online shopping, Internet services, prizes, sweepstakes, and lotteries. The total losses reported were over $3.3B.

There are a number of challenges in interpreting these numbers. They are much smaller than the $1T McAfee estimate for firms, but the FTC reports are just for the United States, and are actual reported values. The FBI and FTC make no claim that these samples are representative of the total population, and do not attempt any extrapolation to total losses. Finally, as the detailed reports make clear, the Internet (and issues of Internet (in)security) facilitate these losses to different degrees.

Individuals suffer losses beyond those easily quantified economically. The McAfee report noted that one cyber-attack on the U.K. health system induced the cancellation of 19,000 appointments. Perhaps no one died as a result, but for many the consequence was likely not just frustration.

Internet users may be exposed to a variety of content that is disturbing, offensive, misleading, disruptive or illegal. A UK government study of harms that arise in the context of social media provided these categories:[12]

- Child sexual exploitation and abuse.
- Terrorist propaganda and recruitment.
- Glamorizing gang life.
- Content illegally uploaded from prisons.
- Sale of opioids and other illegal drugs.
- Anonymous abuse.
- Cyberbullying.
- Facilitating self-harm and suicide.
- Underage sharing of sexual imagery.
- Online disinformation.
- Online manipulation.
- Online abuse of public figures.

Harms to society: Some harms in the previous list result in collective harms to society. In Clark and Claffy,[13] we catalogued a range of societal

11. Federal Trade Commission.
12. UK government.
13. Clark and Claffy.

harms, including those related to market power and the consequences of the surveillance-driven ecosystem, journalism, and the political process. These harms do not always arise because of faults in security, but from the structure of the ecosystem and the profit-motivated design of important Internet applications. Such harms are much harder to quantify, and do not easily map to economic measures. Importantly, many of these harms do not map well to measurements of the Internet infrastructure itself.

Harms to the Internet: A final category of harm is to the Internet itself—loss of utility (and utilization) of the Internet because it is seen as a vector of these other harms. The loss of trust in the applications and services on the Internet that results from fear of phishing, malware, and fraud have a chilling effect on innovation, potential efficiency, and economic gains from the use of the Internet. Interventions to prevent or mitigate the effects malicious activities such as phishing will likely cause collateral damage to legitimate activities. Lack of universal access to the Internet is a harm not only to the excluded users, but the Internet itself as a global platform. These harms are yet harder to quantify, but they are important.

Proxy Harms to Facilitate Estimates or Create a Lever of Intervention

To illustrate the concept of a proxy harm, we use the loss of individual *privacy*, an abstract concept that is (again) hard to taxonomize or quantify in financial terms. Policy-makers address the difficulty of quantifying harms related to privacy by identifying *proxy harms* that stand in for actual harms. Regulations (e.g., the EU General Data Protection Regulation [GDPR]) embed the assumption that unauthorized use of personal data is a harm in and of itself, without requiring a demonstrated harm to those whose data was misused. Of course, not all misuse of personal data is a result of a cybersecurity incident—misuse can arise from deliberate actions of a data holder. The subset of this proxy harm that relates to cybersecurity is the *data breach*, such as when an attacker steals personal information from an enterprise. Classifying data breach as a harm (with concomitant requirements for mandatory reporting, consumer counseling, and free credit reporting) translates the assumed underlying harm to the individual into a financial harm to the firm that suffered the breach, which surveys such as McAfee's attempt to measure.

From Harms to Mitigations

Taxonomizing harms by type of damage, or class of affected stakeholder, can help ensure that we do not miss a major harm. However, to understand how to *mitigate* harms, we need a taxonomy structured around how the harms *arise*. McAfee's report indicates that the most expensive forms of cybercrime are economic espionage, intellectual property (IP) theft, financial crime, and ransomware. They also report that the damage from malware and spyware, closely followed by data breaches, represented the highest cost to organizations surveyed. These two lists are a bit jumbled, in that malware is a vector for many harms including economic espionage, theft of IP, ransomware, and data breach. But it suggests the promise of a taxonomy structured to help bridge between harms and mitigations. Asking how to mitigate malware is a more approachable problem than asking how to mitigate theft of IP, since theft of IP can occur in so many ways.

Such a taxonomy is only the first step. Most attacks today are sophisticated, with many stages in the execution. An attack that leads to data exfiltration may exploit malware, but that malware may have been installed via a social engineering attack that fools an employee into installing it.

Several terms describe attack dynamics—*attack chains, value chains, kill chains*—which all capture the idea that sophisticated attacks today involve many steps, and that a defender can try to select the best step along that chain of steps to intervene and "kill" the attack. To complicate the analysis, some of the same steps may occur in many different attack chains, with different overall structure and targeting different harms.

To illustrate the complexity of the analysis of how to "inform harms via measurement," we use *phishing* as an example.

Illustrative Example: Phishing

Phishing is an attack targeted toward individuals that starts with a false message (e.g., an e-mail) that purports to be from a well-known firm such as a bank or a merchant. It lures the individual to a fake copy of the firm's website, which might capture the login credentials of that individual in order to defraud that user. Phishing was the top category of complaint lodged with the FBI in 2020: about 11% of the 791,790 complaints involved

phishing or a similar attack vector.[14] A Google Safe Browsing report from mid-2020 reported that they had found over 2M active phishing attack sites (distinct sites that were mimicking in some way a legitimate website).[15] These statistics suggest the magnitude of the phishers' efforts, but does not map directly to an estimate of harm.

Given its prevalence, one might prioritize the mitigation of phishing to improve Internet security, even though not all frauds depend on phishing, and mitigating phishing might just chase the malicious actors onto another attack vector. But how could we mitigate phishing? Since phishing starts with a false message, usually embedded in bulk (spam) e-mail messages, one approach is to detect and block spam. Google reports that it blocks over 100M phishing e-mails daily. They claim to block 99.9% of spam.[16] But e-mail is not the only vector by which phishing lures its victims. Social media is now a popular way to initiate a phishing attack. In the third quarter of 2021, Facebook removed 777M instances of "content containing spam," down from a peak of 1.4B in the first quarter of 2020.[17] (There is no data on what fraction of these spam messages were actual or likely phishing attacks.) The LinkedIn platform provides another vector for phishing.[18]

Rather than try to discover and block all the myriad ways an attacker could deliver the first deceptive message, another approach to mitigating phishing is to block access to the fake website to which the phishing message points. Websites generally have an assigned hostname in the Domain Name System, so preventing resolution or use of that name might render the phishing e-mail useless. However, the DNS is a highly decentralized system, and the only way to globally block a name is for the registrar that sold the name or the registry that records the name to disable it. Getting a name disabled is a complex and time-consuming process, and since most phishing attacks are brief, the harm has likely occurred long before the name is actually disabled.

14. FBI.

15. This data from Google is reported in Aaron et al.; The original web is no longer available from Google Transparency Reports.

16. https://cloud.google.com/blog/products/identity-security/protecting-against-cyber-threats-during-covid-19-and-beyond, accessed April 15, 2022.

17. https://www.statista.com/statistics/1013843/facebook-spam-content-removal-quarter/, accessed April 15, 2022.

18. Krebs.

This example illustrates that one can attempt to measure the magnitude of an actual harm (e.g., in economic terms), and one can measure the prevalence of the building blocks that underpin the harm, such as the metrics of phishing (number of phishing e-mails detected or phishing sites detected), but it is hard to robustly link these measurements to actual harm, and thus hard to argue that one or another mitigation based on those lower-level measurements is the best way to reduce the original harm.

Phishing as a Proxy Harm

Reducing the level of phishing might prevent more consequential harms but in many states, phishing in itself is not a criminal act. Countries could establish laws that made certain intermediate steps of phishing criminal acts, such as creating a fake copy of a well-known website for fraudulent or deceptive purposes (which is a form of identity theft), or registering a domain name for fraudulent or deceptive purposes. Ignoring the substantial issues around cross-jurisdictional offenses, such laws would give defenders more tools in their tool box to mitigate harms.

The Quality of Currently Available U.S. Governmental Data

In the United States, there is no single agency (or coordinated set of agencies) with responsibility for data collection related to the Internet. As a result, data collection is fragmentary, and in many cases rudimentary compared to the data collected about other critical infrastructures. We review U.S. federal government Internet infrastructure data collection activities by agency. It is beyond the scope of this paper to analyze the accuracy or utility of all these data sets ourselves, but we provide citations to studies that have tried to use them, when we are aware of them. In most cases, the data is not available for scientific research, constrained by the regulatory framework that authorized its collection.

The Federal Communications Commission

The Federal Communications Commission (FCC) undertakes several data collection activities under different regulations narrowly scoped to questions of service availability, performance, and resilience. We briefly summarize these activities.

Service Availability. As part of its regulatory authority to promote broadband deployment, the FCC plays a central role in compelling providers to report data about service availability.[19] The history is interesting. In 2000, the FCC issued an order to mandate the collection of longitudinal data from providers on fixed broadband deployment (Form 477 data).[20] Their initial notice encountered strong pushback from industry, which the FCC largely accepted:

> *6. …. We rely heavily on the input of commenters, including many service providers that will report pursuant to this program, which has helped us to refine and clarify our data request. Indeed, we believe that the final rules we adopt here will present a significantly lower burden for service providers than the proposal offered in the Notice which initiated this proceeding. Although there may be additional information that could prove useful to our tasks, we understand that we cannot, in one data collection, gather all of the information relevant to every possible future proceeding. Instead, we expect to obtain a baseline of knowledge and understanding about the market for local telephony and broadband services that will both guide us in assessing the overall effectiveness of our actions and will enable us to ask for more specifically targeted information in discrete proceedings before this Commission. We believe that we have distilled our proposal down to that information which is most essential to tracking the development of local competition and the deployment of broadband service to American consumers. Moreover, and most telling about our goals for this proceeding, we also take affirmative steps to ensure that the information collection does not outlive its usefulness by adopting a sunset provision that will terminate the reporting requirement after five years, unless the Commission affirmatively acts to extend it.*[21]

The FCC summarized its decision at that time (year 2000) on the coarse reporting granularity for Form 477 data later in this document:

> *52. To develop this more nuanced understanding of local telephone competition and broadband deployment, we direct providers to compile a list of the Zip Codes in which they offer local telephony and broadband*

19. U.S. Congress.
20. Federal Communications Commission. *In the Matter of Local Competition*.
21. FCC. 00-114, para 6.

services for each state in which they complete Form 477. Such lists should be easily obtainable from companies' provisioning or billing databases and will give the Commission insight into the areas served by particular providers. By knowing whether any or multiple providers are serving a given Zip Code, the Commission will have a much clearer understanding of which customers have—or are likely to have in the near future—a choice among service providers. We conclude that this data, used in conjunction with other publicly-available data, will enable the Commission and others to determine the availability of services and competition in discrete geographic areas, including rural areas and other traditionally underserved areas.

53. We thus decline the suggestions of some commenters that we require providers to complete and file forms at some finer geographic levels, such as the census block level or the Zip Code level. As described above, we conclude that completing forms at these finer levels of geographic granularity would be administratively more difficult for providers. Not only would providers have to identify data at those levels of detail, but we think that a reporting requirement that requires a national service provider to complete over 30,000 zip-code based forms would impose costs far greater than the benefits to be derived.

Over the last 20 years since the FCC began this coarse-grained reporting requirement, the resulting data has been widely criticized as consistently over-reporting the availability of broadband.[22] Starting at least in 2016, the FCC acknowledged this limitation in its release of the Form 477 data:

A provider that reports deployment of a particular technology and bandwidth in a census block may not necessarily offer that service everywhere in the block. Accordingly, a list of providers deployed in a census block does not necessarily reflect the number of choices available to any particular household or business location in that block, and the number of such providers in the census block does not purport to measure competition. [23]

22. See, for example, Drew; Jon; Tibken; Lindsay.

23. Federal Communications Commission. "Fixed Broadband Deployment Data from FCC Form 477."

Concerns about the quality of the form 477 data led Congress to mandate in 2020 (20 years later) that the FCC revise its rules for data collection.[24] As a result of this legislation, the FCC is launching a new effort on broadband data collection, to update its current broadband maps with more detailed and precise information on the availability of fixed and mobile broadband services.[25] This new system, which includes requirements for much more detailed reporting, is expected to come on line in the summer of 2022, and is expected to supersede Form 477 filing at some point in the future.[26]

Performance Measurement. Since 2011, the FCC has also administered an active broadband performance measurement program. They contracted with a third party (SamKnows) to implement the Measuring Broadband America (MBA) program, which provides sample-based views of throughput, latency, jitter, and DNS performance of U.S. broadband services, using measurement devices placed in the homes of volunteer consumers.[27] This longitudinal data set has seen substantive but limited use by researchers in their study of U.S. broadband.[28]

Outage Reporting. Per regulatory requirements, since 2004 the FCC has collected data on network outages and restoration of network operations from providers that meet certain criteria via its Network Outage Reporting System (NORS).[29] Under this framework, qualifying communication providers (wireline, cable, satellite, wireless, interconnected VoIP, and Signaling System 7 providers) are required to report network outages that last at least 30 minutes and satisfy other thresholds specified in the regulation.[30] Since 2007, in the wake of the Katrina disaster, the FCC also operates the Disaster Information Reporting System (DIRS) as

24. Kevin; For legislation, see: Broadband Deployment Accuracy and Technological Availability Act. https://energycommerce.house.gov/sites/democrats.energycommerce.house.gov/files/documents/S1822_amdt_01_xml.pdf, accessed April 15, 2022.

25. https://www.fcc.gov/BroadbandData.

26. For a detailed review of the background and timing of this new system, see FCC, "Public Notice DA 22-182."

27. Federal Communications Commission. "Measuring Broadband America."

28. Bischof et al. "Need, Want, Can Afford–Broadband Markets and the Behavior of Users"; Rula; Bischof et al. "The Utility Argument–Making a Case for Broadband SLAs."

29. Federal Communications Commission, "Network Outage Reporting System."

30. 47 CFR § 4.9—Outage reporting requirements—threshold criteria. Specific to each type of communication provider, but for example, for (a) Cable: "(1) Potentially affects at least 900,000 user minutes of telephony service; (2) Affects at least 667 OC3 minutes; (3) Potentially affects any special offices and facilities (in accordance with paragraphs (a) through (d) of § 4.5); or (4) Potentially affects a 911 special facility (as defined in paragraph (e) of § 4.5)...." https://www.law.cornell.edu/cfr/text/47/4.9, accessed April 15, 2022.

a voluntary, web-based reporting system that offers communications providers a single, coordinated process to report their communications infrastructure status information and request help during major disasters and subsequent recovery efforts. The system is only active in anticipation of or in the wake of disasters. Information reported is confidential.[31] The NORS and DIRS data is all confidential, although in a 2016 Report and Order (12 years after the data collection began), the Commission found that state and federal agencies would benefit from direct access to NORS data and that "such a process would serve the public interest if implemented with appropriate and sufficient safeguards."[32] In March 2021, the FCC adopted an Order to implement this sharing with state, federal, and other government agencies, but explicitly rejected a public comment suggesting that "advocates, researchers, and the public" should also qualify for access.[33]

The Commerce Department (Including Census Bureau)

The Bureau of the Census provides limited data on Internet use, derived from various surveys. As part of the Current Population Survey (a

31. Federal Communications Commission, "Disaster Information Reporting System (DIRS)."

32. Federal Communications Commission, "Amendments to Part 4 of the Commissions Rules Concerning Disruptions to Communications."

33. Federal Communications Commission, "In the Matter of Amendments to Part 4 of the Commission's Rules Concerning Disruptions to Communications."; Comment about researchers on page 10: *"We also find unconvincing, the view of one commenter that "advocates, researchers and the public," among others, should be eligible for direct access purportedly "to hold telecommunications providers accountable and monitor the communications rights of impacted communities." This approach fails to address the Commission's findings that have long treated NORS and DIRS filings as presumptively confidential to further national security and protect commercially sensitive information. Federal Communications Commission FCC 21-34 We find that granting such broad access to NORS and DIRS information would effectively render that treatment moot and thereby detract from these objectives."* Later in the report (p. 22–23), the FCC discusses another comment suggesting FCC perform a risk assessment related to releasing data to "researchers and public interest representatives" but the FCC rejected the comment and acknowledged that *"It is perhaps less likely, however, that public interest organizations or researchers would qualify for such sharing under our rules. Insofar as this commenter would have us relax the "need-to know" requirements to allow such expanded sharing, we reject that proposal, as we believe that the balance we have struck between disclosure of some information to facilitate localized responses to emergencies and service outages caused by them, on the one hand, and the protection of sensitive data from unnecessary disclosure, on the other, will best serve the overall public interest. We also note that no commenter has recommended a practical alternative to the Commission's proposal that would enable aggregation at a lower threshold while ensuring that national security and competitive concerns are addressed."*

telephone survey) they have in some years[34] asked questions in the *Internet and Computer Use Supplement* (ICUS).[35] These supplementary questions are asked on behalf of the National Telecommunications and Information Administration (NTIA), and the latest data is from 2019.

Another source of Census Bureau data about the Internet started in 2013, when the Bureau began asking about computer and Internet usage as part of the *American Community Survey (ACS)*.[36] According to the Census Bureau, the data in the ICUS is richer, due to the larger set of questions, but the larger sample from the ACS gives better estimates at finer geographic granularity.[37] The ACS, which targets about 3.5M household each year,[38] asks three multiple-choice questions about computer and Internet usage:

- In this house, apartment or mobile home—do you or any member of this household own or use any of the following types of computers (desktop or laptop, smartphone, tablet or other portable wireless computer, other).
- In this house, apartment or mobile home—do you or any member of this household have access to the Internet (yes by paying, yes without paying, no)
- Do you or any member of this household have access to the Internet using a—(cellular data plan, broadband, satellite, dialup, other)

The ICUS additionally asks where respondents use computers and the Internet outside of the home, as well as attitudes toward these technologies for both users and non-users.[39]

The Census introduced the *Household Pulse* survey on April 23, 2020, with the goal of measuring the social and economic effects of COVID on American households. Originally, the survey posed detailed questions about Internet use, Internet access, costs, and the nature of the household's connection (wired, wireless, cellular, etc.). It appears that the Census removed these questions from the latest survey.

34. The supplemental survey was conducted in 2007, 2009, 2010, 2011, 2012, 2013, 2015, 2017, and 2019.

35. https://www.census.gov/data/datasets/time-series/demo/cps/cps-supp_cps-repwgt/cps-computer.2019.html, accessed April 15, 2022.

36. https://www.census.gov/programs-surveys/acs, accessed April 15, 2022.

37. Marten.

38. https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/sample-size-definitions.html, accessed April 15, 2022.

39. For the actual questions, see attachment 8 of U.S. Census.

The NTIA provides an "*Indicators of Broadband Need Map,*" but they do not collect any data themselves. They depend on data from the FCC's Form 477 reporting, Microsoft's speed test measurements of its users interacting with Microsoft servers, Measurement Lab's NDT speed test measurements, Ookla's speed test measurements, and the ACS from the Census Bureau.[40]

The significant hurdles to government data collection are evident in this list. For example, the NTIA, in order to have the Census Bureau add the ICUS questions to the November 2021 Current Population Survey, had to present evidence to the OMB in accordance with the Paperwork Reduction Act of 1995 that the survey was necessary for the proper function of the department, and would not present an unjustified burden on the survey respondents. This in turn requires that NTIA post a Request for Comments from the public in the Federal Register.[41]

Bureau of Labor Statistics

The BLS computes one economic measure relevant to the Internet, a single Consumer Price Index (CPI) that tracks the cost of Internet access for urban users, with no regional data, and no measures of rural access.[42] A CPI does not tell us what the service costs, but rather how the price has changed over the period of measurement. In 2021, Greenstein provided a critique of this measure and its value.[43] One problem with the BLS price index is the difficulty of folding into the index the change in quality of Internet access since the start of the measure in 1997. During this period, users have switched from dialup to broadband, and moved from one service tier to another. The CPI is composed of a weighted blend of the prices for service contracts with fixed parameters. Because users frequently move from one contract to another one, usually at a higher speed, the computed CPI less reflects the change in cost for a given service, and more the implicit change in quality that users obtain by moving to a new contract.

40. Moyer.

41. Federal Register, Vol. 86, No. 99, Tuesday, May 25, 2021, p. 28083 @@why is a 1995 act having a 2021 copyright?

42. https://data.bls.gov/timeseries/CUUR0000SEEE03?output_view = data, accessed April 15, 2022.

43. Greenstein.

Federal Bureau of Investigation and the FTC

Both the FBI and the FTC provide data on consumer reports of online harms, in particular economic losses.[44] Since the FTC data includes the FBI data, we describe the FTC data here. The FTC provides an extensive analysis of harms from various perspectives: the category of report, with details on reports of fraud and identity theft, amounts lost, contact method, and demographic factors. The FTC does not attempt to extrapolate from these actual reports to an estimate of total losses, but the detail in these reports provides some insights about possible methods of mitigation.

National Institute of Standards and Technology

National Institute of Standards and Technology (NIST) provides a dashboard that reports the deployment of a security enhancement (Resource Public Key Infrastructure [RPKI]) to the global routing protocol of the Internet (Border Gateway Protocol [BGP]).[45] We are not aware of any other Internet infrastructure measurement undertaken by NIST.

Data.gov

The U.S. government provides a website, data.gov, where all data collected by the government is intended to be cataloged. However, the website is not well-organized to find data relevant to the Internet. Neither "Internet" nor "broadband" are provided as topics or tags. One can do a full-text search for "Internet," which yields any entry containing the word "Internet." Limiting the search to the Department of Commerce (which covers NTIA and the Census) turned up data on such topics as chemical contaminants, mussels, market data for sweeteners, pink salmon, fish catch by recreational anglers, and ionospheric properties.

Searching entries from the Department of Homeland Security (DHS) similarly results in many extraneous entries, but the data relevant to the Internet is operational security data about the government's infrastructure.[46] DHS collects data from the EINSTEIN system to understand

44. Federal Bureau of Investigation; Federal Trade Commission.
45. https://rpki-monitor.antd.nist.gov/, accessed April 15, 2022.
46. https://catalog.data.gov/dataset/cyber-security-information-3dd4a, accessed April 15 2022.

threats to government networks.[47] It also purchases threat intelligence feeds from private security firms, as do many enterprises, and presumably for the same purpose: protection of operational networked infrastructure.

Our summary is that government data collection is fragmented and inadequate for systematic analysis of harms. In our survey of government data collection efforts, we have found piecemeal efforts of limited utility even in the limited contexts in which the government collected or procures a given data set.

Barriers to Improving Empirical Grounding for Public Policies Related to Internet Infrastructure

We identify three immediate barriers to progress: lack of capital and incentive for longitudinal data collection and sharing by the research community; immaturity (relative to other disciplines) of frameworks to safeguard privacy and assess ethical concerns in Internet research; and lack of technical innovation to navigate privacy concerns, impeded by the first two barriers.

Lack of Capital and Incentive for Longitudinal Data Collection and Sharing

Organized collection and curation of data is expensive. Researchers may collect data in support of a specific undertaking, such as a PhD thesis, but sustaining such an effort is typically cost-prohibitive after the student graduates or moves on to another publishable topic. Sustainability costs arise from the size of the Internet and the resulting data sets, and the effort to make the data usefully accessible.

Incentives for publication, funding, and graduation/promotion also favor one-off snapshots that may become stale. Program committees favor novelty over repeated analysis of previous results. Publishing replication studies can be quite challenging, especially if those results have not changed. Evaluation of scientific promotion does not always value artifacts such as data sets or infrastructure.

Structural limitations of funding agencies reinforce these practices. Most funding sources fund short (3-year) research projects, with no mechanism

47. https://catalog.data.gov/dataset/national-cybersecurity-protection-system-einstein-ff834, accessed April 15, 2022.

for extending the budget by a small amount to enable sustained measurement. Moreover, funding agencies do not yet have a way to evaluate longitudinal Internet measurement research, nor an explicit program to review and renew longitudinal Internet measurement activities.

Existing attempts to encourage public data and revisiting of results have included community awards, reproducibility badges, and reproducibility tracks at conferences. These have had only partial success due to their low professional impact, relative to promotion, publications, and degrees. What the academic research community can do on its own does not provide sufficient incentive for meaningful change.[48]

Privacy Implications of Infrastructure Measurements

Measurement data spans a spectrum of identifiability, from personally identifiable information (PII) that includes, for example, an e-mail address, to information aggregated such that it cannot be related to an identifiable person. Most Internet measurement data lies between these extremes. Common examples are data sets that include source and destination IP addresses, location, and/or portions of packet payloads. The growing importance of safeguarding privacy in the personal-data-driven ecosystem has triggered three trends that increase the complexity of measurement efforts.

1. **Application traffic is increasingly encrypted.** Thus, even when passive collection of application-layer information is feasible, its utility is limited. Observers must infer what they can from data flows without being able to see content.
2. **Evolution of complex privacy law.** Many researchers do not understand the implications of privacy laws and regulations to academic research. In the case of health care in the United States, the Health Insurance Portability and Accountability Act (HIPAA) has specific provisions that govern research practices that use medical data. Internet infrastructure researchers need similar procedures to protect misuse of personally identifiable information (PII), and governments must affirm that those procedures are consistent with their laws and regulations. The most pertinent regulations today (2022) are the European GDPR and the California Consumer Privacy Act (CCPA). Although companies

48. NSF-sponsored Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021)—Final Report.

may be subject to the GDPR and/or the CCPA, it is unclear that university researchers are.[49] Both the GDPR and the CCPA encourage forms of data minimization such as pseudonymization and de-identification; these developments are triggering increased interest in disclosure control technologies that can perform such data minimization.

3. **Technological frameworks to support work with PII have emerged,[50] but their utility for Internet infrastructure data science is not yet clear.** For example, with differential privacy, a researcher does not obtain direct access to a data set but may submit queries; the amount of distortion in the result of the query is calibrated to ensure that a metric of privacy leakage remains below a specified threshold. A critical gap remains: *identifying how to apply these privacy preserving technologies to networking problems,* and where networking questions do not fit—for example, when a few queries would consume the entire privacy budget. Measurement researchers face a steep learning curve in order to leverage these advanced privacy-preserving frameworks. It is generally easier just to collaborate with someone else who has measurement data, even if that person is at a company and the data must remain secret and any results must have approval from company lawyers before publishing.

Limitations of Ethical Review Institutions

Another barrier to undertaking and sustaining measurement activities is the lack of familiarity (and consistent treatment) of privacy concerns by Institutional Review Boards (IRBs). IRBs are tasked with ethical and regulatory oversight of measurement research that involves the collection, use, or sharing of PII, consistent with ethical principles, and more recently with privacy laws and regulations.[51] In this rapidly evolving research ecosystem, IRB decisions are surprisingly variable across institutions and the community would benefit from more uniformity.[52]

---

49. The GDPR applies to entities in the European Union, to data processing related to the offering of goods or services to European subjects, and to the monitoring of the behavior of European subjects; see GPDR Recitals 22-24. The CCPA applies to for-profit businesses; see CCPA Section 1798.140(d).

50. Cynthia Dwork et al.; CACM; Evans et al.; Lindell; Francis et al.; Corrigan-Gibbs and Boneh.

51. Erin and David; Dittrich and Kenneally, and Bailey.

52. NSF-sponsored Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021)—Final Report.

The IRB model presents another limitation for today's Internet research environment. "Curiosity-driven" research is a valuable component of evolving our understanding of the Internet ecosystem. A long-standing aphorism in computer science is that when you perform a new measurement you find a new bug—in some network configuration, protocol implementation or the measurement tool itself.

But the idea of exploratory, curiosity-driven research is at odds with the constraints of an IRB. IRBs require a well-defined question and experimental protocol before approving a project. Exploratory research does not always start with a well-formed research question. Important insights have resulted from using data to explore questions that were not contemplated when the data was collected. Similarly, regulations such as the GDPR require that the researcher identify and state the purpose for which they are collecting data containing PII, before the collection begins, and must commit that they will use the data only for that purpose. These are important and necessary safeguards against the misuse of data, but they inhibit exploratory research with sensitive data sets.

Public Policy Problems That Create Pressure for Change

The scientific research community cannot by itself solve the problems we describe. If they are worth solving, it will require higher-level attention, and it is not clear who has that responsibility. In this section, we revisit two public policy issues that evidence suggests are rising to the level where the public interest calls for orchestrated attention to the problem and its mitigation: Internet access, and security and resilience of the infrastructure. Both issues are receiving increasing attention from governments, suggesting an impending point of transition. We suggest the policy and scientific research community should recognize this inflection point and help shape it in constructive ways.

Understanding Deployment and Uptake of Internet Access

Understanding the state of Internet access is a grand challenge because it inherently comes with enormous scale, and deep societal importance. Technologies such as 5G and low-earth-orbit satellite expand the range of options for access, as well as the range of performance and affordability of these options. Access properties of interest include deployment coverage,

availability, adoption, throughput, latency, reliability, and usage. Some of this data is notoriously hard to acquire, and public debate on the accuracy of these data sets for quantifying differences in these properties across the country (the digital divide) has continued for decades.

We discussed in the section "The Quality of Currently Available U.S. Governmental Data" the existing efforts in the FCC and NTIA related to Internet deployment, and their limitations. Understanding longitudinal trends in access properties requires creative and technically sound methods to use all forms of data collection, even those that contain inaccuracies.

As the need to ensure access to a specified level of broadband service increases, maximizing the utility of existing data sets will require federation of data collection, standardization of reporting, methods to overcome measurement bias (e.g., from crowdsourced measurements), multilevel spatial analysis and representation, and support for local contributions to national data sets that preserve privacy. Moreover, effectively mapping broadband access over time requires measurements that go beyond basic access, to capture metrics of quality of service, reliability, and affordability.

Assessment of Resilience of Infrastructure Components

The Internet's status as critical infrastructure underlying (and in many cases controlling) other critical infrastructure means that *infrastructure resilience* is a growing societal concern. Abstractly, the resilience of a system is the ability to provide and maintain an acceptable level of service in the face of various faults and challenges to normal operation, whether due to malice, operator error, hardware or software faults, or any other reason. But this definition says nothing about how to achieve resilience. The real measure of resilience is whether the user experience is disrupted, but user-level impairment is sufficiently hard to measure that some researchers use proxy measures, such as "bit-risk miles," a technical measure of how much capacity and connectivity is lost with a given failure.[53] The relationship between these proxy measures and impairment is unclear.

It also bears noting that the centralization of services will have an uncertain impact on network resilience. Providers of large-scale services with many customers may invest more in mechanisms that improve resilience, but a failure of such a service can disrupt many dependent services.

53. Eriksson, Durairajan, and Barford.

Another framing for resilience focuses on the three key systems that must function for the Internet to provide services: the routing system, the naming system, and the Certificate Authority system. Each system has its own approach to providing resilience, its own challenges to measurement of resilience, and interdependencies with the other systems. Thus, the overall resilience of the Internet is a complex, multidimensional space that is difficult to assess.

Routing System. The routing system must be working in order for the routers to forward packets. Both the routing protocols used internally to ISPs and the global routing protocol (BGP) compute routes dynamically, and will fall back to alternative routes if they exist when one route fails. Researchers have run simulations to predict the loss of connectivity from a given failure. However, such research can only explore the first tier of resilience: the alternative routes that are announced via BGP. In response to failures, network operators can change their routing policies and enable new paths, and they can add new physical connectivity to change the network topology. Network operators can rapidly perform such physical changes if they are local to a data center. This agility makes it difficult to measure or assess the resilience of the Internet to failures of links and routers.

Domain Name System. The Domain Name System must be working so that users can translate domain names into Internet addresses. Again, there are tiers of mechanisms that add resilience to the DNS. Local name resolvers cache the results of queries to avoid having to query authoritative parts of the system. Large DNS resolvers that answer queries about the top-level names are highly replicated, sometimes with heterogeneous implementations, and in many cases exploiting *anycast* routing (where one IP address is assigned to multiple distributed nodes), which enables resilience in case of resolver failure. These tiers of mechanisms make it difficult to estimate how a failure of a resolver will affect the user experience.

Certificate Authority system. The Certificate Authority system must be available so that users can verify that they have reached the intended website (or other service). The CA system is complex, and less mature than the DNS. Many advisories describe how to make a given CA resilient, but we find less discussion about how the overall system can enhance its resilience.

The mechanisms that ensure resilience of these systems are often not active when the Internet is operating normally, which renders elusive the capability to assess resilience of these systems. Operators can do

chaos engineering[54] to assess how their systems respond to failures, but third-party researchers cannot take that approach. In this circumstance, a research agenda must be opportunistic—leveraging sources of data that shed light on aspects of resilience. Case studies of outages (which can represent failures of resilience) illustrate what did *not* work. It is harder to measure the "near misses" as the airlines do—when did mechanisms kick in to preserve the quality of the user experience? A first step could be to review existing literature of attempts to measure various aspects of resilience of Internet infrastructure—what aspect of resilience they studied, using what data, and what methods. The goal would be to generalize from existing studies and develop a conceptual overview of aspects of resilience.

Concrete recommendations and next steps to close empirical gap.

In this section we consider recently proposed concrete steps to make progress on the gap in empirical data to inform policymaking, and the role of academics in doing so. As we develop and evolve taxonomies of Internet harms that map to measurements, we will have to reason carefully about useful ways to measure harms, how to better identify effective remedies to these harms, and how to overcome the counterincentives to data sharing in order to inform these taxonomies.

Systematically analyze risks of sharing proprietary data

Sharing of data by the private sector carries risk beyond concerns about inappropriate release of PII. Release of certain data could lead to adverse commentary on some stakeholder, or policies adverse to the stakeholder's interests, and these concerns trigger understandable hesitation. If we can understand more about the structure of proprietary data, we may be able to improve the options for controlled access for research purposes. Even if sharing such data requires regulatory support, understanding these risks is

54. Informally, *chaos engineering* is the intentional disabling of system elements to confirm what happens. Dependencies among elements are sometime so difficult to track down by inspection that only by causing a fault in the system is it possible to determine what the consequences would be.

prerequisite to developing reasonable regulations. We identify three sorts of proprietary data, with different barriers to sharing.

- **Functional data.** Functional data arises from the business practices of the enterprise. In general, functional data is unique to a given firm. Examples include the network of friend relationships in Facebook or the retweet structure of Twitter. Such data is valuable to researchers attempting to understand the propagation of disinformation. Other examples include a DNS registrar's database of metadata regarding ownership of domain names, business and technical aspects of interconnection agreements among Internet service providers (ISPs), which relate to how traffic is routed and measures of resilience, and packet flow data.

  There are obvious counter-incentives to sharing functional data, including the operational complexity of making it externally available, and the risk that it reveals proprietary information. Misuse of the data may violate the terms under which the firm acquired it. This latter applies particularly to PII. Since release of such data represents a risk to the firm, the firm will need to control data disclosure.
- **Event data.** Event data relates to things that happen to firms, including penetrations, exfiltration, financial losses, and so on. Event data is in general not unique to an enterprise. All enterprises are attacked, or may suffer a breach or a loss. Data of this sort can inform a range of research, including evolving patterns of attack or losses. Firms may be reluctant to reveal firm-specific event data, but may benefit from industry-wide aggregation of such data. For this class of data, it may be possible to define general practices that are sufficient to reduce risk (see section "From Harms to Mitigations"). Geoff Huston recently praised Akamai[55] for their unusually full and careful reporting of a recent outage they suffered, an outage visible enough to get press coverage.
- **Observations.** Observations are data explicitly gathered to inform properties of security or performance, for example, blocklists, collections of malware, or speed tests. When proprietary, the owner will likely sell such data to firms that want to protect themselves or their customers. Researchers can often use historic forms of such data with little commercial value (customers want real-time threat data in order to react to it), but high research value. An untapped opportunity is to

55. Huston.

archive historical data points of certain commercial data sets, in cases where archives can support noncommercial longitudinal research on security and resilience, without interfering with the revenue model for the data.

These categories are not rigid or exclusive. Some firms may translate some of their functional data into observations they sell, which may further inhibit their interest in sharing data with researchers. With increasing interest in the security and resilience of the Internet, we can expect increasing calls for disclosure of event data, along the lines of the data breach laws. Firms will be more willing to accept a requirement for disclosure if they understand the terms under which the disclosed data will be used, and the steps taken to reduce risk. Observations today are sometimes released to researchers on a no-cost basis, under the assumption that the results may benefit the firm as well as society, and the commercial risk is low. But these arrangements are typically one-time events, between trusted colleagues, which makes replication of the research difficult. The government could purchase the data with the understanding that only vetted researchers will use it in approved ways, and then make it available to those vetted researchers.

It is important to learn how other fields of science and critical infrastructure research have addressed their data challenges.

Formalize Roles for Academics in Approaches to Empirical Study of Harms

Other parts of the globe are moving to regularize cybersecurity data, and they have explicitly recognized the importance of engaging and sustaining the academic research establishment in this effort. A recent announcement from the European Union illustrates that this area is receiving substantial attention. In their proposed regulation for Digital Services,[56] they discuss the importance of ensuring access to proprietary data by the academic research community. The report states:

> *"Investigations by researchers on the evolution and severity of online systemic risks are particularly important for bridging information asymmetries and establishing a resilient system of risk mitigation, informing online platforms, Digital Services Coordinators, other competent*

56. European Commission.

> *authorities, the Commission and the public. This Regulation therefore provides a framework for compelling access to data from very large online platforms to vetted researchers.*
> *... In order to be vetted, researchers shall be affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request."*

This regulation emphasizes the need to enable and encourage the academic community to work with proprietary data, sending an important signal that they intend to make their academic research establishment a recognized part of shaping the future of the Internet in the European Union.

The United States has not yet taken such a proactive stance. The best evidence of U.S. consideration of this issue is in the U.S. Cyber Solarium Commission report from 2019. That report set out a strategic plan to improve the security of cyberspace.[57] Among its many recommendations is that the government establish a U.S. Federal Bureau of Cyber Statistics (BCS), to provide the government with the information that it needs for informed planning and action. A recent report from the Aspen Institute echoed this call.[58] Legal academics and lobbyists have already started to consider its structure.[59] There have thus far been no technical, scientific, or engineering voices in this conversation. The Solarium report provides an opportunity to consider the relationship academics could or should have with such a government function.[60]

In this context, the Solarium report makes two concerning observations. First, the report charges the Bureau with the task of "purchasing private or proprietary data repositories," implying that much of the data will not be public. Second, the report proposes that the Bureau "host academics as well as private-sector and independent security researchers as a part of extended exchanges." There is no consideration of engagement of on-campus research groups including graduate students, or release of datasets that can stimulate independent development of advanced analytical

57. Cyberspace Solarium Commission report.
58. The Aspen Institute.
59. Kissick and Rosenzweig.
60. National Academy of Sciences.

techniques such as machine learning. Creation of a BCS could have the unintended consequence of sidelining the academic research community that has been measuring and analyzing the Internet, and which can provide peer-reviewed scientific research on the societal impacts of the Internet, as well as its technical attributes.

Evaluate Potential Roles for Research Funding Agencies in Proposed Approaches

In addition to the European Union's role above, governments could undertake other steps that would narrow the empirical gap that Internet policymakers face.

1. The government could help **navigate misalignment of incentives that impede data sharing to support scientific research**. This includes shepherding data use agreements with providers to facilitate industry contribution of large, shareable data sets for research and STEM work force training. As an example, for over 10 years, DHS supported the PREDICT and IMPACT programs to promote sharing of data for cybersecurity research,[61] and published the legal agreements governing data sharing so others could benefit.[62] Another output of these programs was the Menlo Report,[63] which proposed concrete risk assessment methods and tools for sharing information security incident and threat data. Inspired by the 1979 Belmont Report[64] supporting ethical research in medical and behavioral sciences, the Menlo Report emphasized the following principles: identification of stakeholders and informed consent; balancing risks and benefits; fairness and equity; and compliance, transparency, and accountability.
2. To navigate the entire data management lifecycle, funding agencies could also **facilitate standardization of data practices** by promoting (and funding) the creation of working groups to standardize rules of data set generation and sharing of data artifacts, and to create common application platforms and tooling for maintaining and sharing best-practice pipelines for issuing, processing, and publishing measurements.

61. Department of Homeland Security.

62. Ibid.

63. Kenneally and Dittrich; Dittrich and Kenneally and Bailey.

64. Office of the Secretary, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "Ethical Principles and Guidelines for the Protection of Human Subjects of Research."

3. The government could **fund demonstrations of compensating benefits to the private sector in a program of data sharing**. Each actor in the Internet ecosystem may have an accurate view of their part of the system, but not about the state of their competitors, or the larger ecosystem. Allowing neutral third parties to obtain data from multiple actors can give the private sector, as well as governments and society, a global view of the state of the Internet. But the government will have to find ways to limit liability as a result of responsible sharing of data for documented scientific research.
4. The government could also **encourage/promote specific sharing of data to support academic training of STEM professionals** to work with large realistic Internet data sets. While synthetic network data can be used for classroom exercises, serious research of the sort that leads to professional development requires real data, with the genuine potential for new discovery.
5. The U.S. government could send a strong **signal** to the private sector that builds and operates the Internet: **data sharing is a necessary aspect of sustaining critical infrastructure**, the Internet has now reached this level of maturation, and (as is true in other aspects of society) responsible data sharing needs to be part of normal practice. Developing this model now is a worthwhile activity before some future Internet catastrophe forces an ad-hoc approach to Internet data sharing that would be less beneficial to operators, policymakers, and citizens.
6. The federal government can **contribute to advancing the application of privacy-preserving techniques to Internet infrastructure data**, by promoting cross-fertilization among the fields of Internet measurement, privacy-preserving algorithms, and privacy laws and regulations. *Codes of conduct such as this can enable responsible sharing of data in ways that protect stakeholders while allowing research. To keep science and engineering communities competitive, governments will need to encourage and incentivize sharing data under such standard usage agreements. Funding agencies, journals, conferences, and professional societies should incentivize research conducted under these conditions.*
7. Because technical privacy-preserving tools will not resolve all concerns about the sharing of sensitive data, the government could also **promote well-understood practices**, used in this and other sectors, **to responsibly share data** with qualified independent scholars (Table 1) to allow replication or extension of previous work.

TABLE 1: Elements of a Practical Data Sharing Agreement

Data is made available in curated repositories, or otherwise provided in ways that allows adequate access for legitimate scientific research: • Access requires registration with data source and legitimate research need • Standard anonymization methods are used where needed • Recipients agree to not repost corpus • Recipients agree that they will not deanonymize data • Recipients can publish analysis and data examples necessary to review research • Recipients agree to use accepted protocols when revealing sensitive data, such as security vulnerabilities or data on human subjects • Recipients agree to cite the repository and provide publications back to repository • Repository can curate enriched products developed by researchers

8. Governments could promote the creation and operation of an **oversight function** (a kind of meta-IRB) to oversee community measurement platforms, develop **best practices** around data anonymization, and **support matchmaking** between researchers and data providers. In the United States, the National Science Foundation may be best positioned to undertake such an effort. In the United States, university IRBs exist as part of a set of requirements to receive federal funding, so the government could use this rubric to enrich IRBs understanding of best practices associated with Internet research, of privacy preserving techniques, and of approaches articulated in privacy laws and regulations sufficiently to evaluate privacy risks, even if university research is not subject to those regulations.

Final Thoughts

Advances in measurement in the public interest will have to address these challenges: objectivity of measurements and associated inferences; legitimate business interests in secrecy; respect for privacy, the role of the research community, and sustainability. There are limits to what any given community or even set of stakeholders can do to overcome barriers to Internet measurement in the public interest. There are many actors in the ecosystem—researchers and the academic context within which they sit (with its priorities for publication, funding, advancement, and tenure), service providers, governments, advocates for various objectives ranging from privacy to improved access, and funding agencies. Changing the landscape of network measurement would require adjustment in many

parts of this ecosystem. We recognize the argument that the benefit does not justify the cost. But we also recognize that the Internet is the only critical infrastructure without dedicated government oversight, including a data collection function. It would surprise us if this lack of oversight persisted for another decade. In our view, the question is not whether there needs to be measurement of the Internet in the public interest. The question is how to achieve it sustainably and constructively.

BIBLIOGRAPHY

Aaron, Greg, Lyman Chapin, David Piscitello, and Colin Strutt. "Phishing Landscape 2020." 2020. http://www.interisle.net/PhishingLandscape2020.pdf, accessed April 15, 2022..

Agrafiotis, Ioannis, Jason Nurse, Micheal Goldsmith, Sadie Creese, and David Upton. "A Taxonomy of Cyber-Harms: Defining the Impacts of Cyber-Attacks and Understanding How They Propagate." *Journal of Cybersecurity* 4 (2018): tyy006.

Anderson, Ross, Chris Barton, Rainer Bohme, Richard Clayton, Michel van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. "Measuring the Cost of Cybercrime." *Workshop on Economics and Information Science*, 2012. https://www.econinfosec.org/archive/weis2012/papers/ Anderson_WEIS2012.pdf, accessed April 15, 2022.

The Aspen Institute. "A National Cybersecurity Agenda for Resilient Digital Infrastructure." 2020. https://www.aspeninstitute.org/publications/a-national-cybersecurity-agenda-for-resilient-digital-infrastructure/, accessed April 15, 2022.

Bischof, Zachary S., Fabián E. Bustamante, and Rade Stanojevic. "Need, Want, Can Afford–Broadband Markets and the Behavior of Users." In *Proceedings of the ACM Internet Measurement Conference*, 2014. Vancouver, BC, Canada, Association for Computing Machinery. https://doi.org/10.1145/2663716.2663753.

Bischof, Zachary S., Fabian E. Bustamante, and Rade Stanojevic. "The Utility Argument–Making a Case for Broadband SLAs [in English (US)]." In *Passive and Active Measurement-18th International Conference, PAM 2017, Sidney, Austrailia, Proceedings*, edited by Steve Uhlig, Johanna Amann, and Mohamed Ali Kaafar, 156–169. Springer Verlag, 2017. doi: 10.1007/978-3319-54328-4_12, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

Brian, Cute. "A Conversation about Evolving the Effectiveness of Our Multistakeholder Model." March 2019. https://www.icann.org/en/system/files/files/draft-evolving-multistakeholder-model-issues-list-25apr19-en.pdf, accessed April 15, 2022.

Brodkin, Jon. "AT&T Gave FCC False Broadband-Coverage Data in Parts of 20 States." *Ars Technica,* April 17, 2020. https://arstechnica.com/tech-policy/2020/04/att-gave-fcc-false-broadband-coverage-data-in-parts-of-20-states/, accessed April 15, 2022.

CACM. "Differential Privacy: The Pursuit of Protections by Default." *Communications of the ACM (New York, NY)* 64, no. 2 (January 2021). doi:10.1145/3434228. https://doi.org/10.1145/3434228.

CAIDA. "NSF-Sponsored Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021)–Parts I and II." 2021. https://www.caida.org/workshops/wombir/2101/, accessed April 15, 2022.

Claffy, K. C., David Clark, John Heidemann, Fabian Bustamante, Mattijs Jonker, Aaron Schulman, and Ellen Zegura. "NSF-Sponsored Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021)–Final Report." *Computer Communications Review,* 2021. https://www.caida.org/catalog/papers/2021_wombir2021_report/wombir2021_report.pdf, accessed April 15, 2022.

Claffy, K., G. Polyzos, and H. Braun. "Application of Sampling Methodologies to Wide-Area Network Traffic Characterization. " In *ACM SIGCOMM Computer Communication Review*, Vol. 23. October 1993.

Clark, Drew. "Broadband Mapping Is a Mess. No One Knows What To Do About It." *Broadband Communities Magazine,* July 2019. Accessed February 23, 2022. http://www.bbcmag.com/law-and-policy/broadband-mapping-is-a-mess-no-one-knows-what-to-do-about-it.

Clark, D., and K. Claffy. "Toward a Theory of Harms in the Internet Ecosystem." In *Telecommunications Policy Research Conference (TPRC)*, September 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3443341, accessed April 15, 2022.

Corrigan-Gibbs, Henry, and Dan Boneh. "Prio: Private, Robust, and Scalable Computation of Aggregate Statistics." In *USENIX Conference on Networked Systems Design and Implementation*. Boston, MA: NSDI'17, 2017.

Cyberspace Solarium Commission report. 2020. https://www.solarium.gov/report.

Department of Homeland Security. "IMPACT Cybertrust Data Sharing Project." https://www.impactcyber trust.org/, accessed April 15, 2022.

———. "IMPACT Cybertrust Legal Tools." https://www.impactcybertrust.org/tools_legal, accessed April 15, 2022.

Dittrich, David, Erin Kenneally, and Michael Bailey. "Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Menlo Report." 2013. http://ssrn.com/abstract = 2342036, accessed April 15, 2022.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating Noise to Sensitivity in Private Data Analysis. " In *Theory of Cryptography*, edited by Shai Halevi and Tal Rabin, 265–284. Berlin, Heidelberg: Springer, 2006.

Eriksson, Brian, Ramakrishnan Durairajan, and Paul Barford. "Riskroute: A Framework for Mitigating Network Outage Threats. " In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, 405–416. Santa Barbara, CA: Association for Computing Machinery, 2013.

European, Commission."Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services." 2020. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN, accessed April 15, 2022.

Evans, David, Vladimir Kolesnikov, and Mike Rosulek. "A Pragmatic Introduction to Secure Multi-Party Computation." *Foundations and Trends in Privacy and Security* 2 (2018). http://doi.org/10.1561/3300000019, http://securecomputation.org/.

Federal Bureau of Investigation. "2020 Internet Crime Report." 2020. https://www.ic3.gov/Media/PDF/AnnualReport/2020IC3Report.pdf, accessed April 15, 2022.

Federal Communications Commission. "00-114, In the Matter of Local Competition and Broadband Reporting CC Docket No. 99-301." 2000. https://transition.fcc.gov/Bureaus/Common_Carrier/Orders/2000/fcc00114.pdf, accessed March 12, 2022.

———. "Amendments to Part 4 of the Commissions Rules Concerning Disruptions to Communications, et al., PS Docket No, Further Notice of Proposed Rulemaking, and Order on Reconsideration, 31 FCC Rcd 5817, 5853, para. 88." 2016 Report and Order and Further Notice. 2016. https://www.fcc.gov/document/part-4-ro-fnprm-and-order-reconsideration, accessed March 12, 2022.

———. "Disaster Information Reporting System (DIRS)." https://www.fcc.gov/general/disaster-information-reporting-system-dirs-0, accessed March 12, 2022.

———. "'Fixed Broadband' Deployment Data from FCC Form 477." https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477, accessed March 12, 2022.

———."Measuring Broadband America." https://www.fcc.gov/general/measuring-broadband-america, accessed March 12, 2022.

———. "The National Broadband Plan: Connecting America." 2010. http://downlo ad.broadband.gov/plan/national-broadband-plan.pdf, accessed March 12, 2022.

———. "Network Outage Reporting System." https://www.fcc.gov/network-outage-reporting-system-nors, accessed March 12, 2022.

———. "Public Notice DA 22-182, Broadband Data Task Force And Office Of Economics And Analytics Announce Inaugural Broadband Data Collection Filing Dates." February 22, 2022. https://www.fcc.gov/document/fcc-announces-inaugural-broadband-data-collection-filing-dates, accessed March 12, 2022.

Federal Trade Commission."Consumer Sentinel Network 2020 Data Book." 2021. https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2020/csn-annual data book 2020.pdf, accessed April 15, 2022.

Feld, Harold.*The Case for the Digital Platform Act: Market Structure and Regulation of Digital Platforms*, New York: Roosevelt Institute/Public Knowledge, May 2019.

Francis, Paul, Sebastian Eide, and Reinhard Munz. "Diffix: High-Utility Database Anonymization." In *Annual Privacy Forum*, edited by E. Schweighofer, H. Leitold, A. Mitrakas and K. Rannenberg, 141–158. June 2017. doi:10.1007/978-3-319-67280-9_8.

Greenstein, Shane."Digital Infrastructure. " In *Economics Analysis and Infrastructure Investment*, edited by Edward Glaeser and James Poterba, 423–426. Chicago: NBER Book, University of Chicago Press, 2021.

Huston, Geoff. "Opinion: Why is this Unusual?" July 2021. https://blog.apnic.net/2021/07/27/opinion-why-isthis-unusual/, accessed April 15, 2022.

Internet Corporation for Assigned Names and Numbers. "Board Action on Competition, Consumer Trust, and Consumer Choice Review." March 2019. https://www.icann.org/en/blogs/details/board-action-on-competition-consumer-trust-and-consumer-choice-review-5-3-2019-en, accessed April 15, 2022.

———. "Board Action on Security, Stability, and Resilience 2 Review." July 2021. https://www.icann.org/en/system/ files/bm/rationale-ssr2-22jul21-en.pdf, accessed April 15, 2022.

———. "ICANN Articles of Incorporation." https://www.icann.org/resources/pages/articles-2012-02-25-en, accessed April 15, 2022.

———."ICANN Bylaws." https://www.icann.org/resources/pages/governance/bylaws-en/, accessed April 15, 2022.

Kenneally, Erin, and David Dittrich. "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research." 2012. http://ssrn.com/abstract = 2445102..

Kissick, Chas, and Paul Rosenzweig. "Considerations for the Structure of the Bureau of Cyber Statistics." *Lawfare,* October 2020. https://www.lawfareblog.com/considerations-structure-bureau-cyber-statistics, accessed April 15, 2022.

Kleinig, ."Crime and the Concept of Harm. "*American Philosophical Quarterly*. 15,no. 1 (1978): 27–36.

Krebs, Brian."How Phishers are Slinking Their Links Into LinkedIn." February 2, 2022. https://krebsonsecurity.com/2022/02/how-phishers-are-slinking-their-links-into-linkedin/, accessed April 15, 2022.

Lindell, Yehuda."Secure Multiparty Computation. "*Communications of the ACM*. 64, no. 1 (December 2020): 86–96. doi:10.1145/3387108..

Marten, Michael."Computer and Internet Use in the United States: 2018, ACS-49, U.S. Census. " In *Principles and Practices for a Federal Statistical Agency*, edited by Brian A. Harris-Kojetinand Constance F. Citro, 7th ed.Washington, DC: National Academy of Sciences, 2021. https://www.nationalacademies.org/our-work/7th-edition-of-principles-and-practices-for-a-federal-statistical-agency, accessed April 15, 2022.

Moyer, Tim. "Indicators of Broadband Need Map." 2020. https://broadbandusa.ntia.doc.gov/sites/default/files/2021-07/IBN%20-%20Presentation_0.pdf, accessed April 15, 2022.

Rula, John P., Zachary S. Bischof, and Fabian E. Bustamante. "Second Chance: Understanding Diversity in Broadband Access Network Performance." In *Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data*, 9–14. C2B(1)D '15, London: Association for Computing Machinery, 2015. https://doi.org/10.1145/2787394.2787400.

Smith, Zhanna Malekos, and Eugenia Lostri. "The Hidden Costs of Cybercrime." *MaCafee,* 2020. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf, accessed April 15, 2022.

Stern, Lindsay. "The Consequences of a Broadband Deployment Report With Flawed Data." *Public Knowledge,* April 25, 2019. https://publicknowledge.org/the-consequences-of-a-broadband-deployment-report-with-flawed-data/, accessed April 15, 2022.

Taglang, Kevin. "Congress Tells FCC to Fix Broadband Maps Now." March 2020. https://www.benton.org/blog/congress-tells-fcc-fix-broadband-maps-now, accessed April 15, 2022.

Tibken, Shara. "CNET, Millions of Americans Can't Get Broadband Because of the FCC's Faulty Map. There's a Fix." February 19, 2021. https://www.cnet.com/tech/services-and-software/millions-of-americans-still-cant-get-broadband-heres-a-potential-fix/, accessed April 15, 2022.

U.K government. "Online harms white paper." April 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf, accessed April 15, 2022.

U.S. Census. "Current Population Survey." *Technical Documentation CPS—19,* November 2019. https://www2.census.gov/programs-surveys/cps/techdocs/cpsnov19.pdf, accessed April 15, 2022.

U.S. Congress. "Telecommunications Act of 1996, Pub. L. No. 104-104, 706(a), 110 Stat. 56, 153 (codified as 47 U.S.C. 1302)." 1996.

U.S. Department of Health and Human Services. "Office of the Secretary, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research." *Ethical Principles and Guidelines for the Protection of Human Subjects of Research,* April 1979. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html, accessed April 15, 2022.

U.S. General Services Administration (GSA). "Broadband Data Resources in data.gov." 2021. https://catalog.data.gov/dataset?q = broadband, accessed April 15, 2022.