Measuring Internet Activity: a (Selective) Review of Methods and Metrics

Robert Faris
Rebekah Heacock

This paper can be downloaded without charge at:

The Berkman Center for Internet & Society Research Publication Series:
http://cyber.law.harvard.edu/publications/2013/measuring_internet_activity

The Social Science Research Network Electronic Paper Collection:
Available at SSRN: http://ssrn.com/abstract=2353457
MEASURING INTERNET ACTIVITY

A (Selective) Review of Methods and Metrics
INTERNET MONITOR is a research project to evaluate, describe, and summarize the means, mechanisms, and extent of Internet content controls and Internet activity around the world.

thenetmonitor.org

COVER IMAGE
“World map of Flickr and Twitter locations”
Eric Fischer
Used under a CC BY 2.0 license
http://www.flickr.com/photos/walkingsf/5912169471
Measuring Internet Activity: A (Selective) Review of Methods and Metrics

Robert Faris
Rebekah Heacock
The authors wish to extend their special thanks to Urs Gasser, co-principal investigator of the Internet Monitor project and executive director at the Berkman Center, for many fruitful conversations and excellent input, support, and guidance in producing this paper. Many thanks to Bruce Etling, Hal Roberts, Fernando Bermejo, and Colin Maclay for their generous help and astute suggestions; Jonathan Zittrain provided additional helpful support and guidance as co-principal investigator. We are grateful for the support and inspiration from the broader Berkman community, whose perspectives, scholarship, and work informed and influenced this paper.
TWO DECADES after the birth of the World Wide Web, more than two billion people around the world are Internet users. The digital landscape is littered with hints that the affordances of digital communications are being leveraged to transform life in profound and important ways. The reach and influence of digitally mediated activity grow by the day and touch upon all aspects of life, from health, education, and commerce to religion and governance. This trend demands that we seek answers to the biggest questions about how digitally mediated communication changes society and the role of different policies in helping or hindering the beneficial aspects of these changes. Yet despite the profusion of data the digital age has brought upon us—we now have access to a flood of information about the movements, relationships, purchasing decisions, interests, and intimate thoughts of people around the world—the distance between the great questions of the digital age and our understanding of the impact of digital communications on society remains large. A number of ongoing policy questions have emerged that beg for better empirical data and analyses upon which to base wider and more insightful perspectives on the mechanics of social, economic, and political life online.

This paper seeks to describe the conceptual and practical impediments to measuring and understanding digital activity and highlights a sample of the many efforts to fill the gap between our incomplete understanding of digital life and the formidable policy questions related to developing a vibrant and healthy Internet that serves the public interest and contributes to human wellbeing. Our primary focus is on efforts to measure Internet activity, as we believe obtaining robust, accurate data is a necessary and valuable first step that will lead us closer to answering the vitally important questions of the digital realm. Even this step is challenging: the Internet is difficult to measure and monitor, and there is no simple aggregate measure of Internet activity—no GDP, no HDI. In the following section we present a framework for assessing efforts to document digital activity. The next three sections offer a summary and description of many of the ongoing projects that document digital activity, with two final sections devoted to discussion and conclusions.

A FRAMEWORK FOR ANALYZING EXISTING EFFORTS

Strictly speaking, the Internet is a network of smaller computer networks that use a specific protocol to communicate, but the range of policies and behaviors that influence and are influenced by the Internet is much broader. These include physical infrastructure and broadband policies, content creation and distribution mechanisms, copyright regimes, international law, social and professional communication, citizen-government interactions, and much more. In order to facilitate structured analysis of the Internet’s many moving parts so as to better approach policy and regulatory questions, scholars have developed several methods of categorizing Internet activity. One of the most prominent of these is the division of the Internet into layers.

In The Wealth of Networks, Yochai Benkler describes the Internet as primarily consisting of three layers, each with various accompanying policy questions.¹

¹ Note: The review below began with an inventory of past and current projects at the Berkman Center, followed by an expanded exploration of research conducted by our partners and within our networks. Consultations with experts in the field provided additional leads. We continue to search for additional data and research in this space, and we expect a number of projects exist that we have not yet discovered. As such, this paper is a work in progress. We welcome any and all feedback; please send your comments to info@thenetmonitor.org.
The **physical layer** is at the lowest level, encompassing both the infrastructure that facilitates communication—wires, fiber optic cables, and other transmission channels—and the devices we use to access the Internet. This layer serves as a necessary platform for all online activity. Policy questions within this layer include investment in and deployment and regulation of broadband infrastructure, the development of more open wireless networks, and the massive growth of handheld devices that can reach the web.

The **logical layer** builds upon the physical layer and comprises the protocols, software, and applications that enable users to transmit information online. A range of policy issues are mediated through this layer, including tensions between the free and open delivery of content, the desire for copyright protection, and questions about intermediary liability. Benkler points to the Digital Millennium Copyright Act and the debates over the legality of peer-to-peer networks as examples.

Finally, the **content layer** represents the most high-level of the three, comprising the actual information and knowledge that is produced and communicated over the web. Within this layer are questions about the intersection of digital media and political reform, the extent to which digital technologies can further economic development, and how best to regulate online activity to prevent criminal activity while protecting free speech, debate, and deliberation.

Others have adapted and expanded upon this framework. Jonathan Zittrain, in *The Future of the Internet and How to Stop It*, adopts different versions of a layered framework (“The exact number of layers varies depending on who is drawing [them] and why, and even by chapter of this book.”), renaming Benkler’s logical layer the “protocol layer” and later combining this with the physical layer into a single “technical layer.” He also append additional layers onto Benkler’s model: immediately below the content layer, an application layer that encompasses the “tasks people might want to perform on the network”; above it, a social layer consisting of behaviors and interactions enabled by the layers below.

In a December 2011 article in *Science*, Zittrain works with John Palfrey to propose a three-layered framework for understanding Internet activity in order to make better policies concerning this activity:

First, we need to know more about the architecture of the network and how it is changing. For example, is the web becoming more or less centralized over time? How much are unrelated content and services converging to common hosting within a comparative handful of cloud providers? This is an area where researchers have collected a great deal of data, but we have yet to connect and translate these data consistently for policymakers and the public. Second, we need to know more about how information flows or stutters across the network. Where are there blockages? From what sources do they arise? And third, we need to know more about human practices in these digitally mediated environments.

This framework maps closely to Benkler and Zittrain’s layers—an infrastructure layer akin to the physical layer, an information flow layer that extends Benkler’s logical layer (Zittrain’s protocol layer) to include examination of not only the technology involved in the transmission of information online but also the policies and activities that may inhibit this transmission, and a human practices layer encompassing both Benkler’s content layer and Zittrain’s application and social layers.

For the purposes of this paper, we have chosen to draw most heavily on Zittrain and Palfrey’s work, dividing data and research on Internet activity into three main categories:
INFRASTRUCTURE AND ACCESS
This category includes questions about the physical architecture and operation of the Internet—how is the network functioning? how is it changing over time?—as well as metrics that describe the ability of individual users to gain physical access to the Internet—how much does access cost? at what speeds? Policy considerations closely tied to this category revolve around how best to expand access to digital networks to more users, such as: which market structures and infrastructure policies are most successful and efficient at promoting the deployment of broadband infrastructure? How much should be invested in this infrastructure, and by which entities?

CONTROL
This category expands upon Palfrey and Zittrain’s second question—how does information flow across the network?—and encompasses various efforts to regulate Internet activity. It includes measurements of both technical and non-technical attempts to control this flow by governments, private companies, and independent actors. Related policy questions center upon the mix of regulation and policy that best serves public interest: what are the costs and benefits of different approaches to regulating online activity to address issues such as cybersecurity, copyright infringement, online defamation, hacking, spam, and criminal activity? What are the public obligations and liabilities of private platforms and intermediaries?

ONLINE ACTIVITY: CONTENT AND COMMUNITIES
This category, based on the human services, content, and social layers described above, covers the ways in which humans create, consume, and share information online, as well as relationships and communities that are formed. Policy considerations in this category concentrate on how, and to what extent, digital technologies might best be used to increase social welfare: how can digital tools best be leveraged to contribute to better governance or a more informed populace? What factors contribute to or detract from online creativity, innovation, and digitally facilitated economic growth, knowledge accumulation, and scientific discovery?

INFRASTRUCTURE AND ACCESS
Physical access to the Internet lies at the base of Internet activity—it is necessary but not sufficient for meaningful participation in digital life. The metrics that researchers and analysts use for assessing infrastructure and access are fairly well developed and, although not without errors and bias, provide a reasonably accurate depiction of online connectivity at the national level, and in some cases at the subnational level as well. Table 1, below, outlines the most common sources of data about infrastructure and access.

Internet penetration levels are measured most commonly using one of two techniques. In the first, subscription data provided by Internet service providers is aggregated to provide estimates of the total number of subscriptions by businesses and individuals for a country or region. When divided by the population this renders an average of Internet subscriptions per capita. A second approach to estimating Internet and broadband penetration levels is surveys carried out on a sample of households and businesses. This method offers a somewhat better measure of household penetration levels as it avoids a number of the problems with industry subscription data, such as the impact of multiple subscription households and less than clear distinctions between business and consumer connections. In places where a significant portion of the population rely on cybercafés or other forms of shared access, surveys of individuals will offer a better
TABLE 1: METRICS AND MEASURES FOR INTERNET INFRASTRUCTURE AND ACCESS

<table>
<thead>
<tr>
<th>Metric</th>
<th>Data Sources</th>
</tr>
</thead>
</table>
| **Penetration** | • Business surveys: total number of substitutions  
• Household and consumer surveys: proportion of houses connected to Internet/broadband |
| **Speed** | • Business surveys and market research: advertised speeds (e.g., OECD, FCC)  
• Content delivery networks and web services: download speeds (e.g., Akamai, Netflix)  
• Distributed client-side hardware: download and upload speeds (e.g., government partnerships with SamKnows)  
• Crowdsourcing: download and upload speeds (e.g., Ookla’s Speedtest, M-Lab) |
| **Price** | • Market research: comparison of offers across different ISPs and countries (e.g., OECD, FCC)  
• Crowdsourcing: user-submitted information on prices (e.g., Ookla’s Net Index) |
| **Infrastructure: location, size, and routing** | • IP address distribution  
• Allocation of domains  
• Number of Internet hosts  
• Number, size, and relationships of autonomous systems (AS)  
• Network bandwidth estimates  
• Internet exchange (IX) location and traffic  
• Route identification and analysis  
• National network status (e.g., Renesys, Arbor Networks)  
• International pipe location, traffic, and dependencies |

Although these various approaches measure different things, they are, as expected, highly correlated.\(^5\) Data on individual Internet use aggregated by the International Telecommunication Union from national household surveys\(^6\) shows that Internet penetration rates are highest in northern Europe and lowest in sub-Saharan Africa, though more Internet users reside in Asia than in any other continent.\(^7\)

Measuring the quality of broadband connectivity is somewhat more complicated. Performance can vary considerably over time depending on how many other users are competing for limited bandwidth, and the quality of Internet connections cannot be fully captured in one metric. We often use speed as shorthand for performance, but even here, what we most often mean is throughput: how many bits per second a user receives or sends. This is particularly relevant when downloading or uploading a large document, music, photos, or a video. Other aspects of performance can be important, such as packet loss, latency (time delays in delivering packets), or jitter (packet delay variation). Losing packets or having some packets delayed is more important when streaming a video or using a voice over IP service, such as Skype, than when viewing text online. These different metrics are often highly correlated, so that Internet speed is a reasonable but unreliable approximation of Internet...
performance quality. Nevertheless, it is commonly used as a proxy for the wider set of performance metrics.

One standard measure of speeds is advertised speeds as reported by ISPs. Although advertised speeds inform the decisions of consumers when selecting broadband providers and service options, these ‘up to’ speeds are not necessarily a good measure of actual speeds and may be substantially higher than the speeds that households experience. In recent years, actual speed-based measurements have been more widely available and provide an independent check of these industry-reported speeds. The broadest comparative perspective on Internet speed comes from crowdsourced data—data contributed by users around the world through a number of intermediaries. For example, the popular speedtest.net (offered by the analytics company Ookla) and the M-Lab consortium offer tools for users to measure their connectivity speeds along with other metrics, and in doing so compile a tremendous amount of data from users around the world. A limitation of this data is that those who choose to take these tests are a self-selected sample of individuals. Not only do we not know how well they represent the general population of Internet users for a given area, we don’t know how these selection biases vary across regions. It is a reasonable guess that those who take the test are more interested in the performance of their connection and that where they have options will tend to choose higher speed offerings. This could be balanced out by those that are

FIGURE 1: AVERAGE CONNECTION SPEED TO THE AKAMAI INTELLIGENT PLATFORM, OCTOBER 2013
taking the time to test their broadband speed out of frustration with their lousy connection.

Another prominent source of speed data is from the content distribution management company Akamai, which has servers around the world and is reported to handle a large percentage of global Internet traffic. Akamai regularly tests network latency between cities and monitors network traffic worldwide. The company’s interactive State of the Internet graph lets users compare average connection speeds and average peak connection speeds across countries (see Figure 1, above); additional visualizations display maps of viruses, attacks, connection speeds between cities, and overall traffic volume. Akamai and Ookla’s datasets are based on different measurement methodologies and draw on different samples, resulting in somewhat different results. Both sets of metrics have their advantages. The sampling structure of Akamai’s data—based on a large proportion of Internet connections—suggests that this is a more reliable measure of connectivity speeds. Ookla offers more frequent access to its data and provides public data on more countries. Additionally, Ookla offers a metric developed for assessing VoIP call quality.

In the United States, the Federal Communications Commission (FCC) has implemented a speed test system that uses a structured sample of households that have agreed to the installation of hardware designed to measure connectivity speeds. Under the “Test My ISP” program, the FCC provides 10,000 households with fixed line broadband Internet connections with a Whitebox, a modified router that connects to users’ existing routers and measures jitter, latency, packet loss, and a number of other performance metrics. The first report, issued in February 2013, indicated that many ISPs were meeting or exceeding the speeds they advertise. The FCC is partnering with broadband measurement service SamKnows, which is operating similar programs in the United Kingdom, Brazil, Singapore, and across Europe.

Systematically measuring the price that consumers pay for digital access is similarly plagued by data problems and is complicated by the existence of multiple service options offered at different prices across many different markets. Consumers typically have a choice of different connectivity options that vary by price and speed such that comparisons must be made of very different products. A number of market firms collect data based on market prices offered publicly by different service providers. Open sources of comparative pricing information include the OECD, the FCC, and Google. The ITU also reports per-country average monthly broadband subscription prices as reported by the ISP with the largest market share in each country.

ISPs use a variety of pricing strategies to attract and retain customers (for example, limited time deals or long term contracts at lower prices), which complicates attempts to compile and aggregate standard price options for residential or commercial broadband. The pricing of broadband products can be hard to isolate when Internet service is sold as a part of a bundle that includes telephone and television. Installation, equipment rental, and usage charges may or may not be explicit, and the presence of usage caps, which vary substantially across different providers and locations, further muddies the water. Moreover, the tremendous variation in connectivity speeds makes producing direct comparisons difficult.

Creating price indexes is a standard method for overcoming the presence of heterogeneous products, but this method hinges upon access to weights that determine the contribution of each product to the price index. The actual choices of different broadband products made by consumers
would provide the ideal set of weights for constructing a broadband price index constructed on the market share of different products. This data, however, is proprietary. An alternative approach is to construct a price index based on the range of advertised broadband products. An alternative approach is to construct a price index based on the range of advertised broadband products.

Some have sought to remedy the problem of comparing broadband products offered at very different speeds by converting prices to a per-unit basis, for example dollars per Mbps. This can be a misleading metric. On a per unit basis, a 1 gigabit connection at $200 per month, which translates to $0.02 per Mbps, would register as much less expensive than a $30 per month DSL connection that offers 4Mbps, or $7.50 per Mbps. Even if these comparisons were based on actual data throughput as opposed to advertised speed, this metric is flawed. The value of improving broadband connectivity is not linear; the benefit of moving from a connection of 5 Mbps to 10 Mbps is much higher than upgrading from 95 Mbps to 100 Mbps.

Ookla’s Net Index provides broadband subscription prices contributed by its users and reports the mean subscription price per country divided by the country’s per capita gross domestic product. According to the company’s estimates, the mean subscription price is less than 1 percent of per capita income in 11 countries, led by Luxembourg at 0.6 percent. In a quarter of the 64 countries for which they provide this data, the mean subscription price is greater than 5 percent of per capita income. According to the ITU, which reports the monthly broadband subscription price for the ISP with the largest market share, in 25 of the 165 countries for which data exists in 2010, the subscription price was less than 1 percent of monthly per capita income. In 78 countries, the monthly price in 2010 was greater than 5 percent of

FIGURE 2: FIXED BROADBAND MONTHLY SUBSCRIPTION PRICE AS A PERCENTAGE OF AVERAGE MONTHLY GDP VS. INTERNET USERS PER 100 PEOPLE, CENTRAL & SOUTH AMERICA
monthly per capita income. In 19 countries, the monthly cost was greater than average monthly per capital income. As should not come as a surprise, Internet penetration rates and subscription prices are negatively correlated (see Figure 2, above).

The OECD conducts what appears to be the most comprehensive effort to collect well-structured cross-country data related to telecommunications and uses this data to inform its biennial publication, Communications Outlook. The data and metrics compiled by the OECD cover many Internet-related topics, including broadband subscriber statistics, speeds, and prices; infrastructure measures such as access lines; and a range of market-related metrics, such as operator revenues and investment levels. Statistics on coverage, penetration, prices, services, and speeds are updated more frequently and are freely available on their website. This data is comprised primarily of structured surveys of businesses and consumers, some of which are administered directly by the OECD and some of which are collected by the representatives of member governments. The coverage of this data is restricted to the OECD member states.

Data on the routing and hosting infrastructure of the Internet—including the scale of IP networks, the hardware attached to the network, and the quantity of data that passes through the network—offers not only a picture of the state of the architecture and operation of the Internet in a country, but also rough insights into online activity. Much of this information is compiled from reports by the agencies and companies around the world that provide the core physical and virtual infrastructure for the Internet.

The allocation of the approximately 4.3 billion IPv4 IP addresses offers a highly skewed version of Internet activity, reflecting not only activity but the institutional history of the Internet. One IP address can host any number of content hosts, making this metric a poor proxy for online activity.

Statistics are also available on the number of domain names for each country. This shows the US dominating the space, with over 80 million of the global total of more than 130 million domains, followed by Germany with 6.5 million. Egypt has 32,000, Mozambique 163, and the Cayman Islands 1.4 million. The actual location of domains and servers may bear little relationship to the digital footprint of a country’s population, however: a reader in Moscow may read a post by a blogger who lives in London and posts content to a site hosted by a server in Sweden with a site name provided through a domain name registry located in the United States.

The number of Internet hosts—computers connected to the Internet—offers a somewhat better measure of a country’s digital presence. By one estimate, more than 900 million hosts exist, with the US accounting for over 500 million. The World Bank reports estimates of the number of secure Internet servers, those servers that use encryption for transactions. However, the fact the users can choose the country in which to host their content suggests that this measure is influenced by the competitive edge of each country’s hosting industry and may reflect the overall level of technological advancement of different countries, the level of digital economic activity, or the friendliness of the regulatory regime for those interested in hosting their digital presence outside of their own jurisdiction. This has important legal implications for operating Internet businesses, as well as practical ramifications for the regulation of speech online, but does not offer a reliable measure of Internet activity.

Researchers at CAIDA leverage the information provided for routing Internet traffic to compile data that supports the mapping of the major nodes of the Internet and the connections between these nodes.
Routing information identifies the major nodes in the Internet—typically ISPs or similar institutions—that manage networks. These are identified by autonomous system numbers (ASN). The large variation in the number of ASNs per country suggests that this is also a poor proxy for Internet activity, although this data does point to interesting differences in the complexity of the Internet networks in a given country (see Figure 3, above).\textsuperscript{32}

International bandwidth per capita, derived from the total capacity of Internet backbone transit operators at international connection points, gives us a proxy for the quantity of bits that flow into and out of a given country—assuming that capacity is correlated with demand—and therefore a sense of the level of connectedness with the rest of the world. However, international bandwidth should reveal a mix of inbound traffic to popular content hosts and outbound traffic to content hosted elsewhere, and may bear only a loose correlation to what citizens and businesses are doing online, as it doesn’t necessarily offer a reliable measure of domestic traffic. Each of these indirect measures provides meaningful data on important aspects of Internet access and activity, but requires much more context to fully understand what it means.

CONTROL

Once users are able to connect to the Internet, they face a number of further obstacles to access to information and freedom of speech. Table 2, below, outlines metrics for Internet controls. Both democratic and authoritarian governments regulate the content their citizens can view online; these regulations are becoming increasingly sophisticated, focused not only on the outright blocking of information but also on using more indirect mechanisms to control what users are able to see and do. In some cases, these mechanisms—which range from registration and licensing requirements to libel laws to cyberattacks—create a “chilling effect,” contributing to an environment in which users preemptively self-censor.

FIGURE 3: AUTONOMOUS SYSTEM DIAGRAMS OF CHINA (LEFT) AND NIGERIA (RIGHT)
TABLE 2: METRICS FOR INTERNET CONTROL

<table>
<thead>
<tr>
<th>Metric</th>
<th>Data Source(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take-downs</td>
<td>• Business self-reporting (e.g., Chilling Effects, company transparency reports)</td>
</tr>
<tr>
<td></td>
<td>• Content tracking (e.g., studies on the removal of Weibo posts)</td>
</tr>
<tr>
<td>Filtering</td>
<td>• Business self-reporting (company transparency reports)</td>
</tr>
<tr>
<td></td>
<td>• Distributed data collection and analysis (e.g., OpenNet Initiative, OONI)</td>
</tr>
<tr>
<td></td>
<td>• Crowdsourced reports (e.g., Alkasir, Herdict)</td>
</tr>
<tr>
<td></td>
<td>• Automated tools: websites, keywords (e.g., GreatFire, Is It Down Right Now)</td>
</tr>
<tr>
<td></td>
<td>• Social / media reports</td>
</tr>
<tr>
<td></td>
<td>• Leaked block lists</td>
</tr>
<tr>
<td>DDOS</td>
<td>• Distributed network data gathering (e.g., Arbor Networks, Akamai, Google)</td>
</tr>
<tr>
<td></td>
<td>• Surveys of websites and services</td>
</tr>
<tr>
<td></td>
<td>• Social / media reports</td>
</tr>
<tr>
<td>Malware and other attacks</td>
<td>• Malware analysis and signatures (anti-virus)</td>
</tr>
<tr>
<td></td>
<td>• Malware hosting (e.g., StopBadware, Google)</td>
</tr>
<tr>
<td></td>
<td>• Response coordination (e.g., CERT)</td>
</tr>
<tr>
<td>Legal restrictions</td>
<td>• Legal analysis</td>
</tr>
<tr>
<td></td>
<td>• Social / media reports</td>
</tr>
<tr>
<td>Non-technical controls</td>
<td>• Watchdog group reports</td>
</tr>
<tr>
<td></td>
<td>• Social / media reports</td>
</tr>
<tr>
<td>Self-censorship</td>
<td>• Surveys of Internet users and online organizations</td>
</tr>
</tbody>
</table>

Private companies are also increasingly involved in the Internet control landscape, whether due to government pressure or of their own volition. Governments seeking to censor content or monitor users frequently turn to content hosts and search engines. For instance, in the first six months of 2012, Google received 789 court orders and 1,002 requests for content removal from government entities in 33 countries. Email providers may also be compelled to identify and turn over email messages of a user of interest to the government.

Private companies also act on their own initiative to block access to content, as YouTube did when it temporarily blocked access in Egypt and Libya to an anti-Islam video that helped spark protests in the Middle East in the summer of 2012. Companies may even proactively adapt their services to comply with government regulation. Some Chinese blogging platforms, for instance, refuse to publish posts they deem—individually and in advance of specific government instructions—to be controversial. Other sites block access for entire nations of users in order to avoid these issues, as LinkedIn and Facebook both do in Syria. Social networking sites and web hosts also independently police content based on their own terms of service agreements and business interests.
Targeted campaigns against specific websites have also become a regular part of the Internet control landscape. Distributed denial of service (DDoS) attacks—which have nearly doubled in the past year—have taken down the websites of government agencies, banks, blogging platforms, online retailers, independent media, and other organizations around the world. While many of these attacks are propagated by individual Internet users or loosely connected networks of users, some involve government actors. Malware is also a concern: in June 2013, malware disguised as anti-circumvention software targeted members of the Syrian opposition; malware was used the same month to launch a DDoS attack against Korean government servers.

Researchers and analysts have developed a number of metrics, some well-tested and others more experimental, to document this maelstrom of occasionally conflicting efforts to wield control over online content and activities.

The OpenNet Initiative (ONI) is perhaps the most widely cited organization documenting government Internet filtering. Since 2006, the ONI has methodically tested the accessibility of hundreds of websites in dozens of countries around the world. Researchers maintain a global list of over 1000 (mostly English-language) potentially sensitive websites in nearly 40 thematic categories, including sites related to minority rights, religion, history, environmental issues, political reform, and public health. These thematic categories are grouped into four main areas: Conflict and Security, Internet Tools, Political, and Social. An independent list is also developed for each country tested, focusing on websites in local languages and specific to local issues, such as the sites of political parties or local blog hosts. Researchers within each country then used specialized software to test both lists of websites to see whether they are available, analyze the results of these tests, and assign the countries a filtering score in each of the four content areas (see Figure 4, below).

While the ONI employs a systematic methodology to assess filtering practices, it provides a fixed view of Internet censorship in a country at a particular time—it does not capture fluctuations in filtering that occur around elections, political anniversaries, or
Measuring Internet Activity: A (Selective) Review of Methods and Metrics

During times of social upheaval, Herdict, which collects user reports of website inaccessibility and accessibility in real time, was developed partially in response to this challenge. As of January 2013, Herdict had collected over 320,000 reports from 234 countries and territories. These reports can be sorted and visualized by location, URL, and date, creating a picture of changes as they are happening (see Figure 5, above).

The Open Observatory for Network Interference (OONI) has developed a free and open source software tool called OONI-probe to test traffic manipulation and content blocking. The results of these tests, run by volunteers who download and run the software, are collected and analyzed by OONI researchers; raw data is also made openly available under a Creative Commons attribution license. Developers interested in a specific aspect of network tampering can also write and run their own tests using the software.45

Alkasir, a circumvention software developed specifically to enable access to blocked news and information websites and communication tools, also tracks which sites its users access and provides this information via its Cyber-Censorship Map.46 Users can also view lists of the websites Alkasir has determined are blocked by individual ISPs.47

In China, GreatFire collects and shares both real-time and historical information about blocked sites and search terms (see Figure 6, below). GreatFire monitors a list of URLs that have been submitted by users, as well as URLs that other projects—including Herdict and China Digital Times—have indicated are blocked in China. GreatFire also monitors

---

**FIGURE 5: HERDICT MAP OF WEBSITE INACCESSIBILITY REPORTS (IN RED) AND ACCESSIBILITY REPORTS (IN GREEN)**

---

**FIGURE 6: GREATFIRE REPORT ON FACEBOOK**

**FACEBOOK.COM IS 100% BLOCKED IN CHINA**

<table>
<thead>
<tr>
<th>OCT. 2013</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
</tr>
</tbody>
</table>

**BLOCKED:**

100% (IN THE LAST 30 DAYS)

**OTHERWISE RESTRICTED:**

0% (IN THE LAST 30 DAYS)

**TEST NOW**

**TESTED SINCE:** Fri, Feb 18, 2011

**LAST TESTED:** Thu, Oct 31, 2013

**TESTS:** 12 TIMES (IN THE LAST 30 DAYS)
blocked keywords on Baidu, Google, Weibo, and Wikipedia.

Sites such as Is It Down Right Now and Down for Everyone Or Just Me allow users to check the status of individual websites, testing whether the site is generally globally accessible or not. Down for Everyone Or Just Me returns a simple “It’s just you” or “It’s not just you,” while Is It Down Right Now provides more detailed information about a site’s response times.

While tracking the accessibility of individual websites can provide a picture of the level of Internet filtering in a country, a number of countries have experienced large-scale Internet outages in the past few years. The Cooperative Association for Internet Data Analysis (CAIDA) is working to detect these outages and determine their root causes, whether political (as in Egypt and Libya in early 2011 and Syria in 2012) or natural disaster-related (as during the earthquakes in Japan and New Zealand in 2011).

A number of commercial organizations, including Renesys, Akamai, Arbor Networks, Kaspersky Labs, and a number of other companies also track Internet disruptions due to DDoS attacks and malware.

Controls imposed by private corporations are more difficult to measure. Some companies—notably Google and Twitter—self-report the number of take-down requests they receive from governments and the actions they take in response. Despite a recent increase in the number of companies issuing such reports, transparency like Google and Twitter’s is still not the norm. ISPs and private companies rarely report the requests they receive from governments or their actions in response, if any. Without the cooperation of the companies, gathering meaningful data is difficult. Often, different ISPs have different responses to government requests for censorship, and the ISPs may work in tandem with other mechanisms of content control.

While most studies of Internet censorship focus on blocked URLs and keywords, a recent study conducted by political scientists Gary King, Jennifer Pan, and Molly Roberts compares the substantive content of social media posts individually blocked by Chinese censors to those posts that are allowed to remain online in order to analyze the specific kinds of speech and expression Chinese authorities find objectionable. The study, published in 2012, finds that, rather than targeting posts critical of the government and its policies, Chinese government censors tend to focus on posts that are associated with collective action events.

Monitoring and measuring the impact of non-technical Internet controls—content take-downs related to copyright or defamation laws, or the use of legal or extralegal threats to prevent people from publishing information online—is even more difficult. Chilling Effects maintains a database of over 1,000,000 content take-down demands, aggregating reports from individuals as well as from Google and Twitter. A number of non-profit organizations, including Reporters Without Borders, Threatened Voices, and the Committee to Project Journalists, monitor arrests of, attacks on, threats against, and murders of bloggers and online journalists around the world. Self-censorship is even harder to measure, as obtaining accurate statistics requires the cooperation of individuals who are hesitant to talk in depth about their online activities. As such, few numbers exist, though an August 2011 survey conducted by the Berkman Center for Internet & Society studied nearly 100 Middle Eastern bloggers’ perceptions of online risk and their propensity to self-censor in the face of these risks.
ONLINE ACTIVITY: CONTENT AND COMMUNITIES

Measures of physical and functional access to digital networks and of the physical infrastructure of the networks themselves (the first set of metrics discussed earlier) offer a useful starting point for assessing the potential impact of the Internet on society. The second set of metrics, the quantitative and qualitative assessments that capture different content control techniques practiced around the world, continue to advance and offer a clearer sense of both the measures put in place to protect users against malicious activity online and the limitations placed by governments on freedom of expression and online organizing, but stop short of capturing the full impact of these policies. Ultimately, the best understanding of the efficacy of networked technologies to affect social change and promote human progress—and the costs of restrictions—comes through the observation and measure of what people do and accomplish with digital tools. This is telling both in what is taking place online, for example political campaigns or forums devoted to women’s health issues, and what is not.

This third category—tracking Internet activity, including the content that is produced and accessed and the communities that coalesce around common interests—is also the most challenging of the three to conceptualize and to measure. Ideally, we want to know how well citizens, companies, and public agencies are using the Internet for social, economic, and political ends, and to what extent this contributes to human development and wellbeing. This category covers a wide range of actions, such as the number of people engaged in different types of activities online; the type and quantity of information being produced, accessed, and shared; the formation of civic organizations, professional groups, and new media entities online; the growth and scale of e-commerce and online business; the use of online resources for health, education, and science; and the delivery of government services online and use of digital information to contribute to better governance.

Important qualitative dimensions exist around these activities that cannot be ignored. To what extent does access to the Internet contribute to a better informed populace that can effectively participate in public life and contribute to economic growth and human development? Do citizens have access to information that is accurate? Do they know how to identify misinformation? Does increasing engagement with digital media lead to better policy decisions?

As digital tools permeate all aspects of life in connected communities, the prospect of measuring Internet activity runs the risk of becoming an unbounded exercise. This is both aided and complicated by the influx of data. A useful foundation and anchor would be consistent measures of online activity. Yet we must recognize that a great proportion of digitally mediated activity cannot be attributed to the affordances of digital technology but instead simply reflects the migration and mirroring of political, economic, and social transactions to the digital realm. A markedly different approach is aimed at measuring the incremental shift in activity that results from the introduction and adoption of digital technologies. This angle is better suited to inform many of the policy questions mentioned earlier.

An example of the activity-based measurement approach would be to track the adoption of digital technologies and resources in the classroom. An impact-based approach would seek to estimate the impact of using digital tools on educational achievement. Outcomes and marginal impacts are typically more challenging to observe and estimate; producing evidence-based responses to Internet-related policy questions is a
formidable undertaking for which we are currently under-resourced.

There is however a strong linkage between activity measures and impact measures, and compiling standard metrics of Internet activity is a useful and necessary step towards better understanding the impact of digital activity. Such data contributes to a broad range of the studies that inform what we presume to know about these complex processes, such as longitudinal studies or cross-national comparative studies and time series, which are a core part of unpacking these complex interactions and motivating policy makers and publics.

A diverse set of data sources and measures, outlined in Table 3, below, support efforts to track activity.

### TABLE 3: DATA SOURCES FOR MEASURING ONLINE ACTIVITY

<table>
<thead>
<tr>
<th>General Data Type(s)/Origin(s)</th>
<th>Specific Source(s)</th>
</tr>
</thead>
</table>
| Reporting on individual behavior | • Client-side behavioral monitoring software (e.g., ComScore, Alexa)  
• Cookies and browsing history  
• Consumer surveys |
| Network monitoring: location, type, and quantity of traffic | • Monitoring by ISPs  
• Monitoring by network services (e.g., content distribution networks, Internet security companies) |
| Data collection by websites and services: visitors, contributors, content, links, comments, languages, locations | • Websites, including social media platforms (e.g., user-generated content sites, social network sites, blogging and micro-blogging sites)  
• Search data |
| Social media mapping: link-and/or content-based | • Landscape mapping: platform/service-based mapping (e.g., Twitter, Facebook, blogosphere)  
• Topical or issue-based mapping |
| Qualitative assessments | • Expert opinion surveys |

The raw number of people online is a useful proxy for Internet activity, but does not adequately capture what people actually do online; it doesn’t reflect the constraints and risks associated with government restrictions, the social practices, the depth or dearth of content that is locally relevant in a local language, or the influence of bandwidth or usage-based caps. It is roughly analogous to counting students as a measure of education.

The current state of data collection related to digital activity at the global level is well reflected by the projects that compile and aggregate different metrics to produce indices designed to summarize global online usage and progress. The Networked Readiness Index, produced by the World Economic Forum, seeks to measure the relative success of countries in leveraging the economic potential of information and communication technologies (ICTs) with a particular focus on competitiveness.\(^{66}\) The index is built upon four dimensions: the policy environment; the readiness of the individuals, businesses, and government to apply information and
communication technologies; the actual use of ICTs; and the impact of ICTs. Each of these dimensions is constructed by combining country-level metrics compiled by organizations such as the ITU and World Bank with surveys of business leaders to address the more complex, qualitative, and less well-documented aspects.

The Web Index, produced by the World Wide Web Foundation, incorporates 85 metrics to create an index of the use and impact of the World Wide Web for 61 countries, divided into three sub-indices: communications and institutional infrastructure; web content and web use; and political, economic, and social impact. Although this project is focused primarily on the World Wide Web, the methods and challenges are applicable to other digital realms as well. The datasets that are folded into the index are comprised almost entirely of two types of data: standardized telecommunications and socioeconomic data compiled by governments and international organizations, and transnational surveys of experts and business leaders. As in the case of OECD data, the structured metrics produced by governments and international organizations are also based not on direct measures of Internet activity but on surveys of users and businesses.

To address the more complex issues, such as those concerning political rights, the use of the web for political mobilization, and the availability of accurate health, education, economic, and other types of information, the Web Index conducts its own survey of experts as well as drawing on the results of the World Economic Forum survey. It also includes performance scores compiled by Freedom House and Reporters Without Borders, which themselves are constructed upon a mix of quantitative and qualitative measures that rely heavily on expert opinion. The notable outlier in the data that comprises the Web Index is the number of web pages in each language found on Wikipedia, which is used as a proxy for web content by language. Structured data on Internet activity with global coverage is hard to come by, which explains in part why these indices rely on the opinion of experts.

Surveys, not only of experts but also of businesses and users, continue to occupy a central role in understanding individual behavior. At the national level, user surveys based on probability sampling methods are arguably the best source of information on Internet adoption and usage patterns at a nation-wide level. A prime example is the work carried out by the Pew Research Center’s Internet and American Life project, which conducts regular surveys about online practices of the American public. It is from these surveys that we know that 27 percent of US Facebook users plan to take a break from the site in 2013, 58 percent of US mobile phone users used their phones to compare online prices while in a physical store during the 2012 holiday shopping season, and 20 percent of US adults who monitor their own health indicators do so online. In Europe, Eurostat conducts surveys that include online activity, among other topics, and cover EU member states, candidates for EU accession, and European Free Trade Association countries. The surveys conducted as part of the organization’s information society program cover topics such as access to information about health online. Icelanders are the leaders in this area, with 61 percent of individuals between the ages of 16 and 74 accessing health-related information online, compared to a 27-country average of 38 percent. For user-generated content, Estonia leads the pack with 21 percent uploading self-created content, compared to an average of 11 percent across 27 countries.

Several transnational surveys also include coverage of online behavior. On the commercial side, the 2010 TNS Digital Life study surveyed nearly 50,000 online consumers in 46 countries and combined the survey results with Clickstream data to obtain
both self-reported and passively collected information about online behavior. On the non-profit side, the Internet Society’s 2012 Global Internet User Survey examined the attitudes and behaviors of 10,000 people in 20 countries regarding the Internet’s role in human rights, development, and education; the data from this survey is openly available online. Transnational surveys with a more specific focus also exist, such as Microsoft’s 2012 survey of nearly 8000 children in 25 countries, which asked respondents about their online relationships and behaviors to shed light on online bullying; however, the logistics and cost of carrying out large population representative surveys across dozens of countries limit their frequency and coverage.

Survey methods have evolved over time along with communications technologies, from a reliance on face-to-face interviews to random digit dialing telephone interviews in countries where a vast majority of citizens had landline telephones in their homes. Telephone surveys today must account for a growing proportion of the population that have dropped their landline altogether in favor of cell phones, and many surveys are now conducted entirely online. Many online surveys seek to implement similarly rigorous probability sampling techniques; others attempt to mitigate errors rooted in non-probability sampling with post-stratification weighing, which applies weights to the results of surveys using socioeconomic information on the respondents to correct for samples that are biased towards certain demographic segments. While the costs of conducting opt-in surveys online have dropped to almost zero (click here to take our survey!), the strong selection bias inherent in opt-in methods mean that these surveys may not offer a reliable basis for making inferences about the preferences, beliefs, and actions of online users, although there are signs that rigorous online surveys are gaining ground on telephone-based surveys.

Surveys based on probability sampling methodologies (where the probability of each respondent being chosen from the general population is known) continue to offer the most reliable source of information about online activity. One of the ironies of the study of digital phenomena is that the traditional survey-based methods for understanding human behavior have not diminished in their importance and the advent of digital communication has not eliminated the considerable cost of implementing scientifically robust surveys.

In addition to survey data, a number of direct measures of Internet activity exist: search data, website visits, clicks, purchases, accounts and subscriptions, downloads, uploads, traffic volume, links between websites and between and among social media sites, posts on social media and social networking sites (e.g., blogs, micro-blogs, social networking sites, and video sharing sites), social media site metrics (e.g., hashtags, followers, or likes), comments on posts and articles, social linking sites, edits to websites and wikis, and more. Each of these transactions offers insights into the motives and preferences of Internet users and content producers. This data is also typically limited in coverage to a modest slice of digital life, which makes it difficult to draw inferences about broader populations of Internet users.

Internet service providers, who control network connectivity for a vast proportion of household and business users across the globe, are in a unique position to carry out traffic monitoring. This offers a front row seat to the online behavior of their users. Some ISP data is collected and aggregated by government regulators, though this data usually constitutes a small portion of what is accessible. Household and business broadband subscription data mentioned earlier is based on ISP reports that are typically collected and aggregated by national authorities. The companies that provide hardware and software to ISPs are another
source of traffic information; Sandvine is one example. Data on individual Internet usage collected by ISPs is rarely shared with researchers, and for good reason, given the privacy concerns. Even in aggregate form, there is no standard publicly available source of Internet traffic data.

While generally indicative of levels of activity, overall quantity of data flow is not equivalent to the rate of citizen participation or value of that participation, if for no other reason than that a substantial proportion of throughput is taken up by video streaming: Sandvine reports that in North America, Netflix comprises one-third of peak period traffic, while in other regions, YouTube is the largest source of real-time entertainment traffic. Although watching a stream of one’s favorite TV show over the Internet instead of over the air is a sign of the times, this behavior doesn’t say much about the social and political impact of ICTs.

Several web analytics companies gather data by recruiting Internet users to install software on their computers that tracks their browsing behavior. This data is then packaged and sold to companies that want to understand online behavior and browsing habits. Alexa and ComScore have made a business of this, while also offering free public access to portions of their data. This data covers potentially a wide set of Internet users and might be a closer approximation to representative surveys, but it is still subject to the same limitation: we don’t know how the sample of people willing to install software that will monitor their behavior differs from the population of Internet users. Do they represent the gullible, naive, or adventuresome? Does this alter their online behavior? On the plus side, these companies do track actual behavior, a marked improvement over respondent reported behavior. While they do offer some public access to their data, access to the most detailed data is reserved for their paying clientele.

Websites as well as Internet application providers collect information on their users and visitors as a part of business operations and to varying degrees share this information with researchers and the general public. Search data is a powerful gauge of the interests and preferences of Internet users and offers an unparalleled view into the questions on people’s minds, at least for those online and the queries that they seek to answer on the Internet. These queries are logged and aggregated by search engines and can be analyzed over time and geographic distribution. This data underlies the tools available to the public at Google Insights and Google Trends. A prominent example of the power of this data is Google Flu Trends, which draws on the frequency of searches for flu-related keywords to assess the spread of illness. Other potentially powerful applications include gaining new insights into drug interactions or predicting voter turnout based on regional differences in search activity. It is unclear how far we might go in mining this type of data to better understand global Internet activity; it is possible that we’ve barely scratched the surface of this unique view into the collective digital mind.

The analytics behind Internet search engines are another potent source of information. PageRank, the algorithm used by Google to estimate the relevance of web pages in respect to search queries, offers a link-based measure of the prominence of different websites. The pricing of Internet advertising, which is an important source of revenue for online enterprises large and small, requires monitoring of traffic and activity on sites to determine how much to charge for serving advertisements. This is turn becomes another perspective on Internet user behavior and traffic, at least the commercial aspects of online life.

For some applications, the data sharing and acquisition methods are easily integrated or even baked into the service. Most blogs offer
an RSS feed that allows interested researchers ready access to their posts. The introduction of APIs to core data has also been a boon to data collection and reporting, for example as implemented by Twitter and Facebook. This has helped to spawn third-party social media tracking and analysis companies, such as SocialBakers, which provides some freely available data on Facebook, Twitter, YouTube, Google+, and LinkedIn use along with more advanced, specialized analysis for paying clients; Crimson Hexagon, which incorporates sentiment analysis into its social media research; and Morningside Analytics, which focuses on the networks behind online conversations and information flows.

Contributions to user-generated content platforms in particular offer an informative measure of the location of Internet activity. We know where contributions to Flickr originate; that, after the US, the most Facebook users are found in Brazil, India, and Indonesia; and the language and origin of edits to Wikipedia edits.

Twitter has proven to be particularly attractive to researchers, in part due to the willingness of the platform to allow researches access to data through its API and in part because of the intriguing view Twitter gives into current events, media attention, the emergence of memes and the structure of social relationships online. One of the common approaches to studying Twitter is to bound the analysis by tracking the use and spread of a particular hashtag or hashtags. For example, Devin Gaffney explores word frequency, account creation, and retweeting during and after the 2009 Iranian presidential election. A 2012 paper by Panos Panagiotopoulos, Alinaghi Ziaee Bigdeli, and Steven Sams focused on the use of Twitter by local government authorities in London during the summer 2011 riots. While these examples illuminate the mechanics of online public discourse around specific topics in specific regions and countries at specific times, they can’t be easily aggregated to measure overall Internet activity. Larger scale studies mitigate some of these concerns. For example, Leetaru et al. extract geolocation information from the Twitter accounts of a sample of a month of Twitter activity (over 1.5 billion tweets from over 70 million accounts) to produce a map global Twitter activity.

Twitter is also being used to a means to estimate the attention of Internet users to different news stories and memes online, similar in scope and scale to the use of search term data mentioned earlier. Twitter data also provides interesting crowdsourced monitoring of global events. For example, social media marketing company SocialFlow has posted analyses of prominent hashtags used to share information during Hurricane Sandy and of hashtag use by different groups of users during the October 3, 2012 presidential debate. Tufekci points out several problems with studies that are structured around certain hashtags. The use of hashtags are not consistently good proxies for capturing an online debate, hashtags may mean different things in different contexts, and debates may persist despite users dropping the use of related hashtags, among other several factors that confound reliable interpretation.

Facebook has emerged as an important platform for digital expression and networking. The potential for learning about the mechanisms and influence of networks through the study of Facebook is intriguing and has been the focus of much research attention over the past several years. However, unlike blogs and micro-blogging platforms, which lean heavily towards the public sphere, Facebook straddles the public and private realms. This complicates efforts to study Facebook, as it is more difficult to acquire data, and researchers must grapple with privacy concerns.
A recurring issue with information gleaned from individual websites is representativeness; extrapolating the behavior of the users of a particular website or platform to a broader population is problematic. We are, in effect, observing a segment of the digital world without understanding how this piece fits into the whole. This limitation is less acute where large social networking and social media sites, such as Facebook and Twitter, encompass a great majority of a given digital domain. To study microblogging in the US, access to Twitter is enough to offer nearly universal coverage, as is now true of Facebook and social networking in the US. In Russia, LiveJournal users comprise a great majority of the political discussion core of the Russian language blogosphere. This is not the case, however, across the globe. In China, Sina Weibo dominates microblogging, but an important segment of microblogging occurs within a select population of Twitter users. Even where a single provider captures nearly all of a given segment of Internet activity, we are reminded that those users likely do not represent the general population. While we know that Twitter users in the US tend to be younger and more likely to identify with the Democratic Party than the general public, their reactions to news events at times appear to be more liberal and at times more conservative.

The study of online platforms, including social network sites, blogs, microblogs, and other digital media, represents an approach to studying online activity that is fundamentally different from the survey-based methods described earlier. Instead of asking questions of individual representatives, many of the platform-level studies strive to capture a big picture view of digital activity at a landscape or ecosystem level. This landscape approach may also focus more attention on the structure of networks and communities as well as the interactions among participants.

Compared to approaches that measure and aggregate individual activity, digital landscape approaches may also more directly address collective behavior. Measures of individual
activity online are a strong indication of the reach and influence of digital technologies: number of people online, online media consumption, contributions to user-generated content, hours spent on social networking sites, and so on. These metrics do not capture, however, the prevalence and impact of collective action online. Studying group activity requires more elaborate data collection efforts and more advanced analytical techniques. A number of studies focus on describing the structure, content, and participation in online communities, looking the composition of networks. Indeed, the digital age has proven to be a boon for the application and development of network science. Observable online social networking and linking behavior in particular have opened immense opportunities for network scientists.96

Collaborative and interactive online activity is often discussed in the language of the networked public sphere, which offers a useful conceptual foundation for studying online activity. The networked public sphere represents not only the sum of the various digital media outlets resting on different platforms but also the interactions between these actors and sites that give rise to digital networks. Benkler describes the networked public sphere as an alternative arena for public discourse and political debate, an arena that is less dominated by large media entities, less subject to government control, and more open to wider participation. The networked public sphere provides an alternative structure for highlighting stories and sources based on relevance and credibility.

Studying the online public sphere offers a promising avenue for basing research on the direct observation of digital activity while also addressing the problem of representativeness. A complicating factor is that the networked public sphere spans several different platforms, ranging from blogs and micro-blogs to chat rooms and social networking sites, which may be hosted by a large number of providers. Data collection efforts that span multiple sources can address this challenge. The linking structure of these networks can be used to infer which sites are the most influential, allowing a more targeted study that does not cover the entire universe of digital sources. Adamic and Glance's landmark study employed this approach to study the English language blogosphere, focusing on the linking behavior of 40 A-list political bloggers along with a sample of more than 1,000 liberal and conservative blogs.97

A series of studies published by the Berkman Center researchers in collaboration with John Kelly of Morningside Analytics have used a network mapping approach to study the Persian, Arabic, and Russian-language blogospheres (see Figure 8). These studies identify clusters based on linking behavior and outward facing attention to digital media and other blogs, and are able to track the prominence and spread of ideas in the blogosphere. Broadly speaking, these maps and associated analytics describe the
Measuring Internet Activity: A (Selective) Review of Methods and Metrics

The political and social structure of a backbone of the networked public sphere and potentially an online reflection of offline institutions. Each of these studies map several thousand blogs selected by the number of in-links. While these studies do not claim to offer a broad summary of digital activity, they do offer insights into the most politically salient portions of a key digital platform. The study of the Persian-language blogs, for example, depicts a blogosphere that is characterized by bi-polar political structure with supporters of the current government aligned more closely with conservative blogs on one side and proponents of political reform aligned with more secular-minded bloggers on the other.98

In an analogous study of Russian Twitter, Kelly et al. began with a large corpus of Russian language tweets—over 50 million—that covered the vast majority of the Russian language activity of Twitter in 2010. The team then mapped over 10,000 users and clustered these users into topical and regional networks.99 Employing a similar link-based clustering approach, Taneja and Wu cluster activity associated with the 1000 most-visited websites globally to demonstrate that Internet activity is segregated by language and geography.100

Benkler et al. take a similar approach to the study of the networked public sphere by mapping various digital media sources, including blogs, various online media sources, non-profits and other interest groups, associated with a particular public debate, in this case the SOPA-PIPA debate.101 After compiling over 9,000 stories that discussed this topic, the authors are able to describe the evolution of the debate over time to describe the more influential voices and themes that emerge over an eighteen-month period. Rather than attempting to draw a content-agnostic landscape map, this approach maps the actors and organizations that engage with a given topic with a focus on the temporal dimensions of the debate and mobilization.

In this section, we barely scratch the surface of a rapidly expanding area of activity by researchers and analysts. Our nascent understanding of digital activity is supported by a plethora of ways to capture different aspects of the complex dynamics of online content and communities. We are better poised than ever to study and understand human interactions using digital data. Yet we are still quite far from an empirically robust, comprehensive, and reliable measure of Internet activity. In the next section, we describe several of the ongoing challenges in this field.

QUESTIONS AND CHALLENGES MOVING FORWARD

As we describe in the prior sections, a growing array of methods and metrics exist for assessing Internet activity, fed by a diverse set data sources. Each of the many approaches and methods described above—from surveys of users, businesses, and experts to traffic monitoring to landscape mapping—offers its own strengths and limitations (see Table 4). Taken together, they begin to fill in some of the gaps between our current understanding of digital activity and current policy questions, but further improvements are needed.

The increasing availability of new sources of data—typically transactional data that is less susceptible to misreporting—that will contribute to our understanding of the world and help to guide decisions has caused much excitement among researchers and policy analysts. This is both a wonderful opportunity and an area rife with potential pitfalls. As
### TABLE 4: METHODS FOR MEASURING INTERNET ACTIVITY

<table>
<thead>
<tr>
<th>Method</th>
<th>Strengths</th>
<th>Limitations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual surveys</td>
<td>• Robust sampling can offer data that is representative of general populations.</td>
<td>• Expensive to employ. • Response biases/rely on individuals to accurately recall and report information. • Opt-in surveys lose representativeness.</td>
<td>• Pew Internet • Eurostat</td>
</tr>
<tr>
<td>Business surveys</td>
<td>• Can provide near-comprehensive coverage of sector. • Reflect actual transactional data.</td>
<td>• Reporting biases/industry incentives to over or under report. • Typically exclude proprietary information from public release.</td>
<td>• Wireline and wireless subscription data</td>
</tr>
<tr>
<td>Expert surveys</td>
<td>• Able to address complex impact assessments.</td>
<td>• Often highly subjective. • Difficult to ensure cross-respondent consistency.</td>
<td>• World Economic Forum • Web Index</td>
</tr>
<tr>
<td>Crowdsourcing</td>
<td>• Often able to achieve broad coverage. • Inexpensive. • Can be highly responsive to changing contexts and events.</td>
<td>• Chronic representativeness issues.</td>
<td>• Herdict • Ookla</td>
</tr>
<tr>
<td>Web analytics</td>
<td>• Can provide comprehensive view of platforms/websites.</td>
<td>• Typically limited to specific platforms and sites. • Limited to transactional behavior. • Often proprietary.</td>
<td>• YouTube • Wikipedia • Facebook • Twitter • Search</td>
</tr>
<tr>
<td>Client-side monitoring</td>
<td>• Detailed individual online behavior.</td>
<td>• Sampling/representativeness issues.</td>
<td>• Comscore, Alexa, browsing trackers</td>
</tr>
<tr>
<td>Network monitoring</td>
<td>• Can offer broad view of users’ Internet transactions and traffic.</td>
<td>• Proprietary.</td>
<td>• Sandvine • Akamai</td>
</tr>
<tr>
<td>Routing and hosting infrastructure</td>
<td>• View of infrastructure for conducting Internet business.</td>
<td>• Unreliable measure of user access and behavior.</td>
<td>• Internet hosts • IP addresses • International bandwidth</td>
</tr>
</tbody>
</table>
Distributed data gathering and diagnostic assessments
- With adequate coverage, able to collect empirical data not otherwise available.
- Can be difficult and expensive to implement.
- OpenNet Initiative

Landscape mapping
- Able to infer structure and influence within a network.
- Can be implemented with link-based or content-based methods.
- Link-based applications may not reflect readership, exposure, and influence.
- Drawing inferences beyond the study network is questionable.
- Focus is typically on high-influence nodes with less or no coverage of smaller nodes.
- Blog, microblog, social networking, and digital media mapping

Rufus Pollock points out, advances in computing have led to the “the mass democratization of the means of access, storage and processing of data,” providing the tools to conduct new and sophisticated analyses of digital activity (and other types of data) to researchers worldwide. Jon Kleinberg describes the convergence of social and technological networks that yields a wealth of data and a watershed for researchers: “As such, we are witnessing a revolution in the measurement of collective human behavior and the beginning of a new research area—one that analyzes and builds theories of large social systems by using their reflections in massive datasets.” He also points out difficulties that accompany these opportunities:

Massive datasets can allow us to see patterns that are genuine, yet literally invisible at smaller scales. But working at a large scale introduces its own difficulties. One doesn’t necessarily know what any one particular individual or social connection signifies; and the friendships, opinions, and personal information that are revealed online come in varying degrees of reliability. One is observing social activity in aggregate, but at a fine-grained level the data is more difficult to interpret.

Indeed, data about Internet activity rarely offers clean insights into human society. For example, a study of tweets about choosing not to vote only reveals information about those who have chosen both to participate in Twitter and to self-identify as unlikely voters—hardly a representative sample of a national population.

The advent of “big data” is similarly dual-edged, and introduces an array of promising avenues for understanding digital processes as well as new challenges. The term itself is nebulous, used both to mean data that cannot be manipulated or analyzed on a personal computer and data that contains higher quality observations, regardless of size. Big data per se is not the focus of this paper but offers an intriguing set of possible approaches to measuring Internet activity. While the existence of petabytes of information about online activity has generated significant enthusiasm among business analysts, data scientists caution that finding the signal among the noise can be challenging. As articulated by danah boyd and Kate Crawford, “big data does not mean better data”—the proportion of useful
information that can be gleaned from data decreases as the volume of data increases.\textsuperscript{110}

Zeynep Tufekci describes a set of the methodological questions associated with the use of big data: “Although big data is being variously touted as the key to rigor in social science and as an important basis for policy, this emergent field suffers from inadequate attention to methodological and conceptual issues.”\textsuperscript{111} Methodological questions emerge as techniques historically applied to smaller datasets are scaled up, sometimes poorly, and as new techniques and approaches are developed, tested, and validated.\textsuperscript{112} Other methodological concerns stem from the provenance of digital data. The availability of data derived from digital activity offers researchers more opportunities to use “found” or “trace” data rather than data gathered to answer specific research questions, using data collection methods tailored specifically to strengthen the reliability and explanatory power of the data.\textsuperscript{113} For some, the introduction of big data amplifies the need for strong theory and robust methods; others see opportunities in data mining to unearth patterns with less emphasis on establishing causal mechanisms.\textsuperscript{114}

The following summarizes some of the major methodological and theoretical challenges facing those who seek to measure and analyze digital activity, at both the small and large scales, in order to answer the policy questions described above. The issues span the full spectrum of the analytical cycle, starting with basic availability and access to data, to measurement errors, sampling problems, and problems that stem from the analysis of complex processes. Although none of these challenges are unique to the collection and analysis of digital data, the relative incidence and impact on research is particular to this field.

MISSING DATA, AGGREGATE DATA, UNEVEN DISTRIBUTION OF DATA

In some cases, data of great interest is not gathered, or is missing for large swathes of the globe. Other variables of interest, such as creativity and innovation, are inherently difficult to measure. Drawing inferences from aggregate data, whether it be at the country level, by language group, or by time period, obscures much important variation.

BIG DATA, CHEAP DATA, EXPENSIVE DATA

The Internet is awash in data, much of which is available through bulk download, APIs, and/or RSS feeds. A vast majority of blogs are offered up to the public sphere. Twitter, Facebook, and other platforms provide APIs to their data. However, despite the institutional and technical infrastructure put in place to facilitate capturing and analyzing this data, much of the analytical capacity resides in closed and semi-closed architectures. The technological hardware, software, and skills needed to harvest this data means that both acquiring data and analyzing it can still be prohibitively expensive. Many private sector firms specialize in monitoring social media for commercial purposes, but fewer data gathering efforts are designed specifically for general public access or public interest research.

PUBLIC SPHERE, PRIVATE SPHERE

Much of the speech that occurs on digital forums and platforms is not fully public and hence difficult to capture, record, and aggregate. Many of the most meaningful transactions take part outside of the view of the public Internet, and much more never enters the digital realm. These private and semi-private conversations—deemed “dark social” from the perspective of public social media observers—certainly inform and are partially reflected in the public sphere, but our view of digital activity is a highly selective and incomplete version of human communication and interaction.
SIGNAL AND NOISE
Online activity is often not what it purports to be. Commercial spam floods online discussions, and voices are at times sock puppets or the product of astroturfing. Government-sponsored information campaigns—e.g., the 50 Cent Army in China, pro-regime bloggers in Iran, and Kremlin-backed voices in Russia—seek to influence and color online discussions. Attribution problems are a persistent feature of online activity; it is often difficult if not impossible to know who is responsible for online speech. In the same vein, mapping online voices to sites, regions, or countries is subject to considerable uncertainty and error.

REPRESENTATIVENESS
Much of what we discern about digital activity is based on observations of different slices of online life, perhaps from a debate on Twitter, a Facebook group, a network of bloggers, or an online community forum. Participation online is conducted by a collection of self-selected individuals, and participation in different forums and platforms varies by the individual. This makes drawing inferences about wider circles of online activity based on any of these enclaves or drawing conclusions about wider social trends based on digital activity shaky at best. While we have access to more and more information, personal, professional, and commercial, our ability to translate this into knowledge about the preferences and activities of the general population is still severely hobbled.

COMPLEX PROCESSES AND A MYRIAD OF INTERDEPENDENT FACTORS
The questions and targets for observation related to Internet activity—whether related to physical infrastructure or social, political, and economic outcomes—are the result of the interaction of a complex set of factors. Individual decisions online are the product of preferences and motivations that are shaped by personal capability, physical and economic access to digital platforms, the ability of governments and private actors to restrict and influence online behavior, market forces, the architecture of online tools and platforms, and the informational and organizational context created by online communities and sources of information available to individual users. Collective action online reflects all of these individual motives and constraints along with the added complexities of network interactions and emergent group behavior.

Moreover, many of the variables of interest are highly covariant. Income growth, broadband penetration rates, and connectivity speeds tend to move together, along with educational achievement, health outcomes, good governance, and participation in online communities and discussions. This reduces our ability to draw lessons from different outcomes and establish cause and effect.

CONCLUSIONS
While the Internet is rising in importance as a multi-purpose platform for social, economic and political ends, we are seldom in the position to apply robust empirical evidence to inform pressing policy decisions.

Many of the standardized measures of Internet activity are improving. We now have time series data that offers a reasonably accurate view of the numbers of households online in different countries, the quality of service and speeds at which they connect to the Internet, and the prices they pay. Yet the economic and development costs being borne by countries that are slower to bring their citizens online are still poorly documented.

Similarly, over the past several years a number of studies have helped us to better understand the various strategies adopted by governments and private actors to limit freedom of expression and online organizing. However, the costs of these measures in terms of innovation, economic output, and social and cultural growth are not well understood.
Several promising approaches may emerge to fill the persistent gap between policy questions and the current state of knowledge. In-depth study of particular segments of the online world, network mapping, and ecosystem-level approaches are a promising source of better understanding. Yet while many new methodological approaches are emerging to address the problems of representativeness and to bridge the gap between observational and survey methods, several formidable obstacles remain. Much of the best empirical data on Internet behavior and interactions is derived from commercial products and are proprietary. Efforts to foster greater openness and to fund open data alternatives may help, though privacy concerns are likely to remain a challenge.

The power of big data and web analytics is yet to catch up with the data needs of policy making. We currently have a mosaic of different measures and metrics—data compiled by business and governments, data about specific sites, data collected by users, survey data, expert surveys, and social network and content analysis—to apply to our study of digital activity and its impacts. For policy makers, however, the traditional offline methods—surveys of users, businesses, and experts—still form the core of the base of knowledge about digital phenomena.
NOTES

3 Ibid.
7 International Telecommunication Union, World Telecommunication/ICT Indicators 2012.
17 Examples include PointTopic (http://point-topic.com/) and Telegeography (http://www.telegeography.com/).
20 “International Broadband Pricing Study: Dataset for public use,” Google Policy by the
Registries vary in the quality and quantity of IP allocation data they make publicly available; AFRINIC provides quite a lot of easily accessible information (http://www.afrinic.net/en/services/statistics), while APNIC’s statistics (http://www.apnic.net/publications/research-and-insights/stats/) are harder to navigate.


28 IANA provides a table of the allocation of IPv4 address space to various registries (http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml).


28 IANA provides a table of the allocation of IPv4 address space to various registries (http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml).


While ICTs historically included many analog devices and applications, the various definitions of related sets of technologies (e.g., digital technologies, the Internet) are rapidly converging. For the purposes of this paper, we use the terms loosely, interchangeably and without prejudice.


Measuring Internet Activity: A (Selective) Review of Methods and Metrics


106 A recent analysis by Microsoft Research found that the bulk of data analysis jobs conducted at even the largest of technology companies—often described as “big data”—could be conducted on an average laptop. See: Raja Appuswamy, et al., “Nobody ever got fired for buying a cluster,” Microsoft Research Technical Report, January 2, 2013, http://research.microsoft.com/apps/pubs/default.aspx?id=179615.


