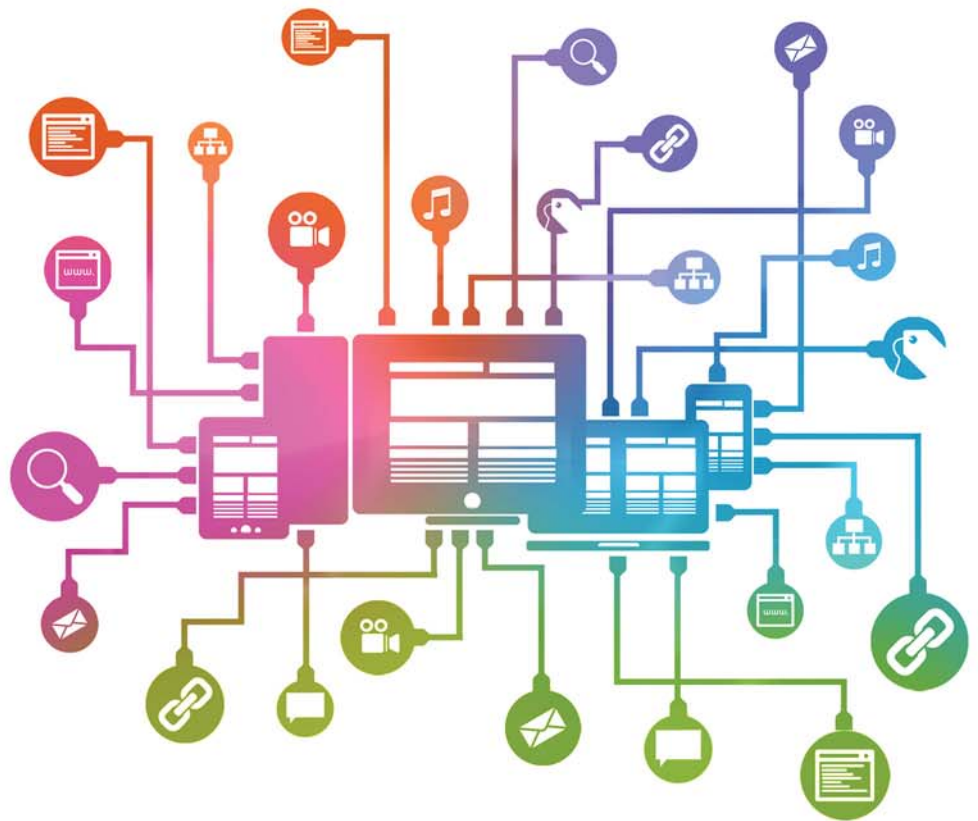


ONLINE PRIVACY AND ISPS:

ISP Access to Consumer Data is Limited and Often Less than Access by Others



Peter Swire, Associate Director, The Institute for Information Security & Privacy at Georgia Tech; Huang Professor of Law, Georgia Tech Scheller College of Business; and Senior Counsel, Alston & Bird LLP

Justin Hemmings, Research Associate, Georgia Tech Scheller College of Business and Policy Analyst, Alston & Bird LLP

Alana Kirkland, Associate Attorney, Alston & Bird LLP

Preface

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others

This Working Paper provides a detailed, factual description of today's online ecosystem for the United States, with attention to user privacy and the data collected about individual users. The Working Paper addresses a widely-held, but mistaken view about Internet Service Providers ("ISPs") and privacy. That view asserts that ISPs have comprehensive and unique access to, and knowledge about, users' online activity because ISPs operate the last mile of the network connecting end users to the Internet. Some have cited this view to suggest that ISPs' collection and use of their customers' online data may justify heightened privacy restrictions on ISPs.

This Working Paper takes no position on what rules should apply to ISPs and other players in the Internet ecosystem going forward. But public policy should be consistent and based on an up-to-date and accurate understanding of the facts of this ecosystem. The Working Paper addresses two fundamental points. First, ISP access to user data is not *comprehensive* – technological developments place substantial limits on ISPs' visibility. Second, ISP access to user data is not *unique* – other companies often have access to more information and a wider range of user information than ISPs. Policy decisions about possible privacy regulation of ISPs should be made based on an accurate understanding of these facts.

Technological Developments Place Substantial Limits on ISPs' Visibility into Users' Online Activity:

1. **From a single stationary device to multiple mobile devices and connections.** In the 1990s, a typical user accessed the Internet from a single, stationary home desktop connected by a single ISP. Today, in contrast, the average Internet user has 6.1 connected devices, many of which are mobile and connect from diverse and changing locations that are served by multiple ISPs. By 2014, 46 percent of mobile data traffic was offloaded to WiFi networks, and that figure will grow to 60 percent by 2020. Any one ISP today is therefore the conduit for only a fraction of a typical user's online activity.
2. **Pervasive encryption.** We present new evidence about the rapid shift to encryption, such as the HTTPS version of the basic web protocol. Today, all of the top 10 web sites either encrypt by default or upon user log-in, as do 42 of the top 50 sites. Based on analysis of one source of Internet backbone data, the HTTPS portion of total traffic has risen from 13 percent to 49 percent just since April 2014. An estimated 70 percent of traffic will be encrypted by the end of 2016. Encryption such as HTTPS blocks ISPs from having the ability to see users' content and detailed URLs. There clearly can be no "comprehensive" ISP visibility into user activity when ISPs are blocked from a growing majority of user activity.
3. **Shift in domain name lookup.** One integral function of ISPs has been to match the user's web address request to the correct domain and specific Internet Protocol ("IP") address. Today there is a still small, but growing, trend of Internet users utilizing proxy services that displace this traditional ISP function. Examples include Virtual Private Networks ("VPNs") and new proxy services offered by leading Internet companies. When a user accesses the Internet through an encrypted tunnel to one of these gateways, ISPs cannot even see the domain name that a user is visiting, much less the content of the packets they are sending and receiving.

Non-ISPs Often Have Access to More and a Wider Range of User Information than ISPs:

1. **Non-ISP services have unique insights into user activity.** At the same time that the above technological and marketplace developments are reducing the online visibility of ISPs, non-ISPs are increasingly gathering commercially valuable information about online user activity from multiple contexts, such as: (1) social networks; (2) search engines; (3) webmail and messaging; (4) operating systems; (5) mobile apps; (6) interest-based advertising; (7) browsers; (8) Internet video; and (9) e-commerce. This Working Paper explains the data flows and mechanisms for advertising for each of these contexts, many of which gather insights about users that are not available to ISPs. Traditional ISPs are not market leaders in any of these major areas; rather, they are just starting to compete in some of them.
2. **Non-ISPs dominate in cross-context tracking.** Each of the above-listed services and platforms gathers volumes of data about users, frequently with insights into content (social networks, webmail, etc.) and other information often characterized as sensitive in privacy debates. While it is analytically instructive to understand each service/platform, the real insights come from combining information from multiple services/platforms – what we call “cross-context tracking” linked to a particular user device or across devices. The 10 leading ad-selling companies earn over 70 percent of online advertising dollars, and none of them has gained this position based on its role as an ISP.
3. **Non-ISPs dominate in cross-device tracking.** Yesterday’s desktop has evolved into today’s tablets and smartphones, and tomorrow’s innumerable devices in the Internet of Things. A growing share of advertising tracking targets the user across multiple devices. Market leaders are companies for whom users log in across multiple devices, such as smartphones, tablets, and laptops. Today, cross-device data collection from logged-in and not logged-in users is led by non-ISPs.

In summary, based on a factual analysis of today’s Internet ecosystem in the United States, ISPs have neither comprehensive nor unique access to information about users’ online activity. Rather, the most commercially valuable information about online users, which can be used for targeted advertising and other purposes, is coming from other contexts. Market leaders are combining these contexts for insight into a wide range of activity on each device and across devices.

Executive Summary

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Executive Summary

Online Privacy and ISPs: ISP Access to Consumer Data Is Limited and Often Less than Access by Others¹

This Working Paper provides a detailed, factual description of today's Internet ecosystem for the United States, with attention to user privacy and the data collected about individual users. For two decades, there have been complex policy discussions about how to protect users' privacy online while also enabling the provision of advertising-supported content and robust commercial activity on the Internet.²

This Working Paper is intended to provide information useful to Congress, federal agencies, and the general public in consideration of online privacy issues. Among other relevant fora, in 2015 the Federal Communications Commission ("FCC") issued its Open Internet Order, which brings Internet Service Providers ("ISPs") under the common carrier requirements of Title II of the Telecommunications Act.³ Title II contains Section 222, which governs how telecommunications service providers use and disclose Customer Proprietary Network Information.⁴ In April 2015, the FCC held a hearing on broadband Internet privacy, for which one of the authors of this Working Paper was invited to testify.⁵

This Working Paper grew out of the April hearing, where there were large factual disagreements about important aspects of online privacy for broadband services newly covered by Title II. At the hearing, FCC officials expressed interest in better understanding these facts. *This Working Paper, in response, is intended to provide a factual and descriptive foundation for making public policy decisions about the privacy framework that should apply to ISPs and other companies that collect and use consumers' online data.*⁶

¹ The authors thank Marie Le Pichon for creating the Diagrams, which are under a Creative Commons Attribution 4.0 license and should be attributed to her. We also thank Brooks Dobbs and Addison Amiri for assistance on technological aspects of this Working Paper.

Research support for this Working Paper comes from Broadband for America, the Institute for Information Security and Privacy at Georgia Tech, and the Georgia Tech Scheller College of Business. The views expressed here are those of the authors.

² Peter Swire, "Markets, Self-Regulation, and Government Enforcement in the Protection of Personal Information," U.S. Department of Commerce, Aug. 15, 1997, (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=11472). This Working Paper addresses issues relevant to law and policy in the United States. Other nations have different privacy regimes, but this Working Paper does not specifically address practices outside of the U.S.

³ This Working Paper uses the familiar term Internet Service Provider ("ISP") in the way it is generally understood – an organization that connects users to the Internet. Discussions of data collected by an ISP refer to information received by a company specifically by virtue of its providing end users a connection to the Internet. In its Open Internet Order, the FCC used a somewhat different term: "Broadband Internet Access Services." The FCC defined these as a "mass-market" retail service by wire or radio that provides the capability to transmit data to and receive data from all or substantially all Internet endpoints, including any capabilities that are incidental to and enable the operation of the communications service, but excluding dial-up Internet access service. "In the Matter of Protecting and Promoting the Open Internet," *Report and Order*, FCC 15-24 app. A (2015) (hereinafter "The Open Internet Order").

⁴ The statutory cite is 47 U.S.C. §222. The FCC's regulations implementing Section 222 are at 47 C.F.R. § 64.2001 *et seq.*

⁵ "Federal Communications Commission Workshop on Broadband Consumer Privacy," *Federal Communications Commission*, April 2015, (<https://www.fcc.gov/news-events/events/2015/04/public-workshop-on-broadband-consumer-privacy>); Peter Swire, "Comments to the FCC on Broadband Consumer Privacy," presented before the Federal Communications Commission Workshop on Broadband Consumer Privacy, April 2015, (<https://transition.fcc.gov/cgb/outreach/FCC-testimony-CPNI-broadband.pdf>).

⁶ Knowing that the facts can be complex and difficult to understand, we are creating a mechanism to receive factual comments, with the intention of correcting mistakes or lack of clarity where such exist. Comments can be submitted to comments@iisp.gatech.edu, and any updates will appear on the website of the Institute for Information Security and Privacy at Georgia Tech.

The Working Paper addresses a widely-held, but mistaken view about ISPs and privacy. The view asserts that ISPs have comprehensive and unique access to, and knowledge about users' online activity because they operate the last mile of the network connecting end users to the Internet. Certain consumer advocates and others have cited this view to suggest that ISPs' collection and use of their customers' online data may justify heightened privacy restrictions on ISPs.

This Working Paper takes no position on what rules should apply to ISPs and other players in the Internet ecosystem going forward. But public policy should be consistent and based on an up-to-date and accurate understanding of the facts of this ecosystem. The Working Paper addresses two fundamental points. First, ISP access to user data is not *comprehensive* – technological developments place substantial limits on ISPs' visibility. Second, ISP access to user data is not *unique* – other companies often have access to more information and a wider range of user information than ISPs. Policy decisions about possible privacy regulation of ISPs should be made based on an accurate understanding of these facts.

Technological Developments Place Substantial Limits on ISPs' Visibility into Users' Online Activity:

1. **From a single stationary device to multiple mobile devices and connections.** In the 1990s, a typical user accessed the Internet from a single, stationary home desktop connected by a single ISP. Today, in contrast, the average Internet user has 6.1 connected devices, many of which are mobile and connect from diverse and changing locations that are served by multiple ISPs.⁷ By 2014, 46 percent of mobile data traffic was offloaded to WiFi networks, and that figure will grow to 60 percent by 2020.⁸ Any one ISP today is therefore the conduit for only a fraction of a typical user's online activity.
2. **Pervasive encryption.** We present new evidence about the rapid shift to encryption, such as the HTTPS version of the basic web protocol. Today, all of the top 10 websites either encrypt by default or upon user log-in, as do 42 of the top 50 sites.⁹ Based on analysis of one source of Internet backbone data, the HTTPS portion of total traffic has risen from 13 percent to 49 percent just since April 2014.¹⁰ An estimated 70 percent of traffic will be encrypted by the end of 2016.¹¹ Encryption such as HTTPS blocks ISPs from having the ability to see users' content and detailed URLs. There clearly can be no "comprehensive" ISP visibility into user activity when ISPs are blocked from a growing majority of user activity.
3. **Shift in domain name lookup.** One integral function of ISPs has been to match the user's web address request to the correct domain and specific Internet Protocol ("IP") address. Today there is still a small, but growing, trend of Internet users utilizing proxy services that displace this traditional ISP function. Examples include Virtual Private Networks ("VPNs") and new proxy services offered by leading Internet companies. When a user accesses the Internet through an encrypted tunnel to one of these gateways, ISPs cannot even see the domain name that a user is visiting, much less the content of the packets they are sending and receiving.

⁷ "The Zettabyte Era – Trends and Analysis," Cisco, May 2015, (www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html).

⁸ "Cisco Visual Networking Index (VNI) Mobile Forecast Projects Nearly 10-fold Global Mobile Data Traffic Growth over Next Five Years," Cisco, Feb. 3, 2015, (<http://newsroom.cisco.com/press-release-content?articleId=1578507>).

⁹ See Appendix 1 to Chapter 1.

¹⁰ See Appendix 2 to Chapter 1.

¹¹ "Sandvine: 70% of Global Traffic Will Be Encrypted In 2016," Sandvine, Feb. 11, 2016, (<https://www.sandvine.com/pr/2016/2/11/sandvine-70-of-global-internet-traffic-will-be-encrypted-in-2016.html>).

Non-ISPs Often Have Access to More and a Wider Range of User Information than ISPs:

1. **Non-ISP services have unique insights into user activity.** At the same time that the above technological and marketplace developments are reducing the online visibility of ISPs, non-ISPs are increasingly gathering commercially valuable information about online user activity from multiple contexts, such as: (1) social networks; (2) search engines; (3) webmail and messaging; (4) operating systems; (5) mobile apps; (6) interest-based advertising; (7) browsers; (8) Internet video; and (9) e-commerce. This Working Paper explains the data flows and mechanisms for advertising for each of these contexts, many of which gather insights about users that are not available to ISPs. ISPs are not market leaders in any of these major areas; rather, they are just starting to compete in some of them.
2. **Non-ISPs dominate in cross-context tracking.** Each of the above-listed services and platforms gathers volumes of data about users, often with insights into content (social networks, webmail, etc.) and other information often characterized as sensitive in privacy debates. While it is analytically instructive to understand each service/platform, the real insights come from combining information from multiple services/platforms – what we call “cross-context tracking” linked to a particular user device or across devices. The 10 leading ad-selling companies earn over 70 percent of online advertising dollars, and none of them has gained this position based on its role as an ISP.¹²
3. **Non-ISPs dominate in cross-device tracking.** Yesterday’s desktop has evolved into today’s tablets and smartphones, and tomorrow’s innumerable devices in the Internet of Things. A growing share of advertising tracking targets the user across multiple devices. Market leaders are companies for whom users log-in across multiple devices, such as smartphones, tablets, and laptops. Today, cross-device log-in is led by non-ISPs.

In summary, based on a factual analysis of today’s Internet ecosystem in the United States, ISPs have neither comprehensive nor unique access to information about users’ online activity. Rather, the most commercially valuable information about online users, which can be used for targeted advertising and other purposes, is coming from other contexts such as social networks and search. Market leaders are combining these contexts for insight into a wide range of activity on each device and across devices.

Meeting Privacy and Other Goals for the Internet

The White House and leading regulatory agencies have expressed strong support both for privacy protection when individuals are online, and for effective uses of data about users’ online activity. We briefly give examples of support both for uses of personal information and limits on such uses to frame the later description of modern online data collection and use.

The United States protects privacy with many detailed laws, regulations, enforcement regimes, self-regulatory codes, and in other ways.¹³ The Obama Administration has emphasized the importance of privacy online in numerous ways, including in its announcement of a Consumer Privacy Bill of Rights, stating: “Privacy protections are critical to maintaining consumer trust in networked technologies.”¹⁴ The Federal Trade Commission (“FTC”)

¹² “IAB Internet Advertising Revenue Report: 2015 First Six Month Results,” IAB & PwC, Oct. 2015, (http://www.iab.com/wp-content/uploads/2015/10/IAB_Internet_Advertising_Revenue_Report_HY_2015.pdf).

¹³ See, e.g., Peter Swire & Kenesa Ahmad, *U.S. Private Sector Privacy: Law and Practice for Information Privacy Professionals*, International Association of Privacy Professionals (2012).

¹⁴ “Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy,” *The White House*, Feb. 2012, (<http://www.whitehouse.gov/sites/default/files/privacy-final.pdf>).

has made privacy protection online a major priority.¹⁵ As mentioned above, the FCC is now carefully studying privacy issues related to broadband Internet access services.

Along with privacy limits on data collection and use, there are benefits in our information age from gathering and using personal information. In its 2014 Big Data report, the Obama Administration discussed multiple benefits, such as improved fraud detection¹⁶ and cybersecurity,¹⁷ and “enormous benefits” associated with “targeted advertising.”¹⁸ That report stated: “Consumers are reaping the benefits of a robust digital ecosystem that offers a broad array of free content, products, and services.”¹⁹ Regulatory agencies have similarly recognized such benefits.²⁰

With these introductory comments in mind, we next outline the 10 Chapters that accompany this Executive Summary, addressing specific parts of the online ecosystem. Appendix 1 to this Executive Summary explains key terms we use in this Working Paper, including: availability vs. use; content vs. meta-data; cross-context tracking vs. cross-device tracking; ISP vs. non-ISP; and visibility and seeing.

Chapter 1: Limited Visibility of Internet Service Providers into Users’ Internet Activity

In providing the last-mile connection to the Internet for their customers, ISPs carry users’ data traffic on their network. In most cases, ISPs have relatively accurate information about a user’s name and billing address, and they may have users’ credit card information and phone number. This Chapter explains the technological and market changes that have made ISP visibility into users’ Internet activity far from comprehensive. We highlighted this Chapter’s major findings above: (1) the shift from a single stationary device and ISP to multiple mobile devices and ISPs, (2) pervasive encryption, and (3) the shift in domain name lookup.

Of these, the recent and rapid shift to HTTPS and other forms of encryption is perhaps the clearest and simplest way to explain why ISPs today and in the future do not have “comprehensive” access to users’ Internet activities. HTTPS blocks the possibility of ISP access to the content of users’ activities – the technology called “deep packet inspection” does not work on encrypted communications. HTTPS also blocks the possibility of ISP access to detailed URLs, which can reveal granular details of a user’s search or other online activities.

Taken together, the three technological developments described in this Chapter show fundamental changes in what information is even theoretically available in providing the last-mile connection – the job of an Internet Service Provider. In addition, the strong trends toward multiple and mobile devices and connections, encryption, and changes in Domain Name System (“DNS”) lookup are likely to continue.

¹⁵ At the time of writing, the most recent major FTC report is “Big Data: A Tool for Inclusion or Exclusion?” *Federal Trade Commission*, Jan. 2016, (<https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf>).

¹⁶ Executive Office of the President, “Big Data: Seizing Opportunities, Preserving Value,” *The White House*, May 2014, p. 39, (https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf).

¹⁷ *Id.* at 40.

¹⁸ *Id.* at 50.

¹⁹ *Id.* at 41.

²⁰ As a recent example, FTC Chairwoman Ramirez recently discussed benefits of cross-device tracking, including continuity in services across multiple devices and providing consumers with in-store discounts derived from searches on home computers. FTC Chairwoman Edith Ramirez, “Opening Remarks of FTC Chairwoman Edith Ramirez, Cross-Device Tracking: An FTC Workshop,” Nov. 16, 2015, (https://www.ftc.gov/system/files/documents/public_statements/881513/151116cross-devicetracking.pdf).

Diagram E-1 shows a funnel for what information is available about user activity going forward for ISPs. At the top are the multiple contexts discussed in the Working Paper, where different players in the online ecosystem see detailed URLs and content about user activity. Due to pervasive encryption, VPNs, and the other developments discussed here, technology often blocks ISP access to user traffic. Next, users are shifting to multiple devices and ISPs, so an ISP's connection to any one device is far less than complete, especially in the Internet of Things world we are rapidly entering. Finally, especially as WiFi hotspots become the majority of traffic, any one ISP only sees a fraction of the activity on any one device. In short, ISPs have far less than a comprehensive view of any user's Internet activity, and the rich information available to non-ISPs means that ISPs do not have unique visibility into users' online activity.

Appendix 1 to Chapter 1 shows the widespread use of encryption today by the top 50 Internet sites. Appendix 2 shows data about the recent and substantial shift to HTTPS for Internet backbone traffic.

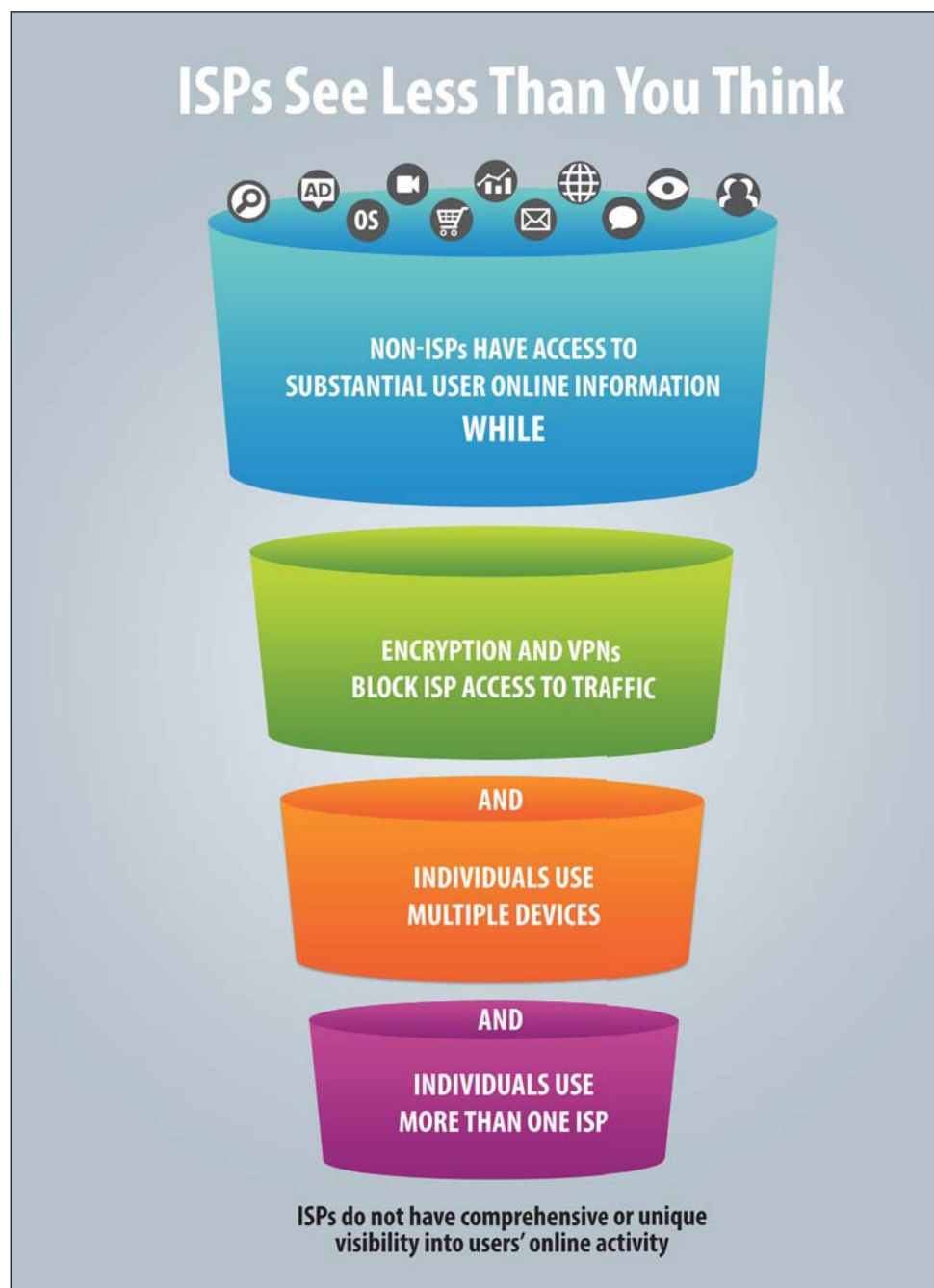


Diagram E-1

Chapter 2: Social Networks

Chapter 2 is the first of several Chapters that discuss certain categories, or “contexts,” that are important ways that various players in the Internet ecosystem gather information about users’ Internet activity. Each of these Chapters explains the prominent content and metadata that become available about users’ activity, especially for advertising purposes. Each Chapter then analyzes how the access of non-ISPs compares with the access of ISPs for this context of data gathering.

Chapter 2 examines the flow of data through social networks, including data to which social networks have privileged or unique access, and the value of that data for advertising services. For social networks, there are three main data streams:

- a. **User-Generated Data.** By design, social networks generally include a large amount of data supplied by the users themselves. Users create profiles including personal data such as name, city of birth, relationship status, and place of employment. Depending on the platform, users may post pictures, videos, URLs, and comments on posts of their personal and professional contacts.
- b. **Metadata.** Along with data supplied by the user, the social network gains granular information about the user’s interaction with the network, such as location data and activities of the user’s contacts, which can then be combined with user-supplied data about interests and preferences in various products and services.
- c. **Logged-In Users.** Social networks generally require an authenticated login from users, allowing for better tracking of that user. In particular, when a user signs in to the same social network on multiple devices, the social network can link each of those devices to the user for cross-device tracking. Social networks also use “plug-ins” on third-party websites to collect data about a logged-in user’s activity on those sites. All of that user’s activity can be accurately tied to the social network account and, depending on the amount and accuracy of information shared, to a specific, identified person.

ISPs do not have that level of insight or visibility of user behavior.

Chapter 3: Search Engines

For the past decade, search has generated almost half of all online advertising revenue, based especially on collection of two types of data: 1) user search queries, and 2) search results, including which results users ultimately click through to visit. The specificity of users’ search queries can provide key insights into their intent, including the users’ likelihood of purchase. When the search is performed over an HTTPS connection, as has become the norm, the ISP can only see which search engine was used and the host domain of the clicked link, but not the search query or the full URL that was clicked.

Chapter 4: Webmail and Messaging

Providers of webmail and other messaging services have the ability to scan the content and metadata of their users’ messages for purposes such as security and advertising.²¹ Scanning for security can reduce the transmission of spam, malware, and illegal content such as child pornography. Scanning can also identify keywords present in messages that are sent or received, which are then used to target advertising to the user. When webmail is accessed over an HTTP connection, ISPs could have the technical ability to perform deep packet inspection to access user content. Most webmail providers have recently moved to HTTPS by default, however, so ISPs are technologically blocked from this information.

²¹ For what is called end-to-end encryption, even the service provider cannot scan the content of the message.

Chapter 5: How Mobile Is Transforming Operating Systems

When it comes to the technical capability of tracking user activity, no software or service is as comprehensive as the operating system (“OS”). Especially with the dramatic rise in mobile computing, the OS today is becoming far more tightly linked with advertising-relevant data, in at least three major ways. First, the leading operate app stores that generate usage data and attract app developers by being advertising-friendly. Second, the OS facilitates collection of location data, available often both to the OS and app developers. ISPs have some capability to access “coarse” location information through triangulation of cell towers, but more precise location information is generally gathered by non-ISPs based on a Global Positioning System (“GPS”), WiFi hotspot, and other sources of location data. Third, personal assistants such as Apple’s Siri, Google’s Google Now, and Microsoft’s Cortana mean that OS systems gather detailed data from across the device in order to answer user queries. Previous separation between the OS and advertising is shifting greatly in the mobile setting.

Chapter 6: Interest-Based Advertising and Tracking

Many players in the online advertising ecosystem gather data about the online activity of users and devices. Going beyond the earlier scope of online behavioral advertising (“OBA”), this Chapter provides new Diagrams and explanation for the system of interest-based advertising (“IBA”), a broader term that includes the increasingly common practice of adding offline information to cookie-based, mobile advertising ID-based, and other online information. Notably, new Diagrams show the roles of publishers, supply-side platforms, advertising exchanges, demand-side platforms, and marketers, for both the mobile and non-mobile advertising ecosystem. ISPs historically have not been leading players in the IBA system, and the leading roles have been played by non-ISPs, who often are leaders as well at cross-context and cross-device tracking.

Chapter 7: Browsers, Internet Video, and E-commerce

This Chapter more briefly examines three additional contexts that are relevant to non-ISP collection of data. Major browsers vary in how extensively they collect user information, but the amount collected can be significant. For instance, most browsers carefully analyze user behavior to suggest search terms while the user is typing and then later use that information to autofill online forms by default. When users are logged-in, their browsing information can be integrated with information from the other contexts engaged in by that browser company. By contrast, ISPs are not developers of any of the major browsers and do not have access to this information.

For Internet video accessed through a browser or a mobile app, the party hosting the video content has the same ability to gain information about the user as any other site hosting content. Third-party ads are served in connection with video content the same as for other content. When Internet video is delivered over a HTTPS connection, the ISP can only see the host domain.

E-commerce sites (first-party retailers) often create long-standing relationships with their consumers. Due to purchases on the site, e-commerce sites usually have relatively accurate and detailed information about a user’s name, credit cards, billing and shipping addresses, and phone numbers. E-commerce sites can also develop profiles of what their users purchase, which are more valuable the more often the user comes back to the same site. ISPs, by contrast, are not market leaders in their own e-commerce efforts, and they do not have first-party access to the variety or volume of information other e-commerce sites have.

Chapter 8: Cross-Context Tracking

This Chapter defines cross-context tracking, and discusses two ways that companies can build a context map for users. Cross-context tracking is the combination of different types of data, such as those discussed in the preceding Chapters – ISPs, social networks, search, webmail and other messaging, operating systems, mobile apps, interest-based advertising, browsers, Internet video, and e-commerce. The same company within the advertising ecosystem often plays a role in multiple contexts, such as an operating system company that also provides a search engine, or a social network company that also has an advertising network. These companies often perform cross-context tracking in two ways:

- a. **Logged-in (deterministic) cross-context tracking.** When a user logs-in to the same service in multiple contexts, that company can accurately map activity in each context to the logged-in account. For example, if a user searches for a location in a search engine, and then links to a driving navigation service provided by the same company, the company can attribute all of that activity to the individual account.
- b. **Not logged-in (probabilistic) cross-context tracking.** Not logged-in context maps are built around an individual user or device, but without the definitive log-in event as a catalyst. Instead, companies can compare data collected in each of their service contexts and use a proprietary algorithm to estimate when different activities are performed by the same user or device. These not logged-in maps can be used independently or to augment an existing logged-in cross-context map with additional data from outside that company's contexts, or as a commodity to be sold to other advertising entities.

The rise of cross-context tracking, often by companies with leading market roles in multiple contexts, heightens the value to advertisers of the insights into users' Internet activities that come from each context. We provide a cross-context chart for major ISPs and other companies, listed by over five percent of the market, market presence, or not in the market. The chart illustrates that the "unique" insights into user online activity most thoroughly is available to companies that have not historically been ISPs.

Chapter 9: Cross-Device Tracking

This Chapter explores the ways in which different entities can create cross-device maps for users. As with cross-context tracking, companies can create cross-device maps based on logged-in (deterministic) tracking or not logged-in (probabilistic) tracking.

Building on the earlier Chapters' discussions of the various technologies, this Chapter provides a summary of how different parts of the ecosystem work together. An accurate device map, especially when combined with an accurate cross-context map, provides distinct advantages for advertisers:

- a. **Frequency Capping.** By being able to track each context and device a user engages with, advertisers can make sure that no individual user sees a single advertisement more often than desired.
- b. **Attribution.** Cross-device tracking can allow advertisers to accurately attribute sales conversion to previous-in-time advertising impressions, including reduction in fraud. For example, if a user sees an ad on her smartphone and then performs a search on her desktop for that product, an accurate cross-device map demonstrates that the smartphone ad was effective in driving the purchase.
- c. **Improved Advertising Targeting.** By collecting data across multiple devices, advertisers have a fuller picture of the user to whom they are targeting advertisements, allowing for a higher likely return-on-investment for each advertisement.

- d. **Sequenced Advertising.** Cross-device tracking can enable companies to conduct sequenced advertising campaigns. Regardless of the device used, the advertiser can make sure that each ad in sequence is served to the user in the intended order.
- e. **Tracking Simultaneity.** Cross-device tracking can also allow for multi-screen tracking of users. If a user is watching content on their smart TV while also using a tablet, an accurate device map can allow a company to know what ads are being served to the smart TV and sync those ads with the ones served to the user's tablet.

In this emerging ecosystem, ISPs are merely one source of data, and their subscriber relationships provide a diminishing portion of any user's history of Internet activity, as users shift to an expected average of 11.6 devices by 2019.²² A single cross-device tracking company works with numerous sources of information, few of them related to the ISP function, to gather and analyze data in creating the device map.

Chapter 10: Conclusion

In summary, based on detailed analysis of today's Internet ecosystem in the United States, this Working Paper concludes in Chapter 10 that the evidence does not support a claim that ISPs have "comprehensive" knowledge about their subscribers' Internet activity, for encryption and other technological reasons. Similarly, ISPs lack "unique" insight into users' activity, given the many contexts where other players in the ecosystem gain insight but ISPs do not, and the leading role in cross-context and cross-device tracking played by non-ISPs.

This Working Paper takes no position on what rules should apply to ISPs, or to providers of services in the other contexts (often called "edge providers"). However, public policy should be consistent and based on an accurate understanding of the facts. The following Chapters provide details and citations to further explain today's online ecosystem.

²² "VNI Forecast Highlights," Cisco, (http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights/index.html).

Appendix 1

Some Key Terms

Availability and Use: In this Working Paper, we distinguish between the overall “availability” of certain data and the actual “use” of that data. “Availability” means that no technological barriers prevent an entity from accessing data. Just because data is available in this sense, however, does not mean that the entity actually views and uses such data. In addition, just because data is “available” in this sense does not mean that it is practical to use it.

Example of deep packet inspection (“DPI”). Deep packet inspection “inspects the content portion of the [web] traffic flowing through the network in real-time.”¹ In the 1990s, the content of user traffic was rarely encrypted, so that traffic may have been theoretically “available” to an ISP in the sense used here – no technological barriers to ISP access. However, our research shows that the actual “availability” was (and remains) constrained by the cost, capacity, and speed of DPI equipment at that time, which typically could not inspect more than a small fraction of the traffic flowing through an ISP network. Due to these limits on availability, ISPs’ actual “use” of DPI for marketing purposes has been limited compared to uses of data in other contexts.

Content vs. Metadata: This Working Paper discusses various types of “content” and “metadata” that may be visible to ISPs and non-ISPs. In an email, for example, the to/from information is metadata, while the subject line and text are content.

Example of host names and detailed URLs. Recent cases are treating URLs that are simply host names (such as www.example.com) as metadata or routing information. By contrast, some cases suggest that if the full URL contains search terms input by the user, the full URL should be treated as content.² The content/metadata distinction is useful even though reasonable people may differ about where the line should be.

Cross-Context Tracking: There has been considerable public attention to cross-device tracking, where a company can link multiple devices to a user, such as a laptop, tablet, and mobile phone. This Working Paper introduces the term “cross-context tracking,” which we explore in depth in Chapter 8. In earlier Chapters, we examine various “contexts” one at a time, to show the data flows for each, including: ISPs; social networks; search; webmail and other messaging services; interest-based advertising; operating systems; browsers; Internet video; and e-commerce.

Example of cross-context tracking. Suppose a user is logged-in to a webmail account, and then uses other services, such as a navigation service or social network, offered by the same company. Regardless of whether the user accesses these services on one or multiple devices, it is “cross-context tracking.”

ISP vs. Non-ISP: In this Working Paper, we use the term “Internet Service Provider” (“ISP”) in the way it is generally understood – an organization that connects users to the Internet. Discussions of data collected by an “ISP” refer to information received by a company specifically by virtue of it providing end users with a connection to the Internet. Non-ISPs refer to companies that receive information through any other mechanism. One task of the Working Paper is to assess the modern technology and market realities of ISP vs. non-ISP data collection and use.

¹ George Ou, “Understanding Deep Packet Inspection (DPI) Technology,” *Digital Society*, Oct. 23, 2009, (<http://www.digitalsociety.org/files/gou/DPI-Final-10-23-09.pdf>).

² “In Re: Google Inc. Cookie Placement Consumer Privacy Litigation,” (Nov. 10, 2015), (<http://www2.ca3.uscourts.gov/opinarch/134300p.pdf>).

Example of ISP vs. non-ISP. A company provides broadband connectivity to the Internet. The same company (and its affiliates) provides other services, such as search, webmail, and online advertising. We discuss the company as an ISP for its broadband connectivity and as a non-ISP for the other services.

Visibility and Seeing: We use the terms “visibility” and “seeing” as synonyms for “availability,” defined above. One theme of this Working Paper is that ISPs today have much less visibility into users’ Internet activity than they would have had in the early days of the Internet if they had deployed DPI at scale. We emphasize three technological barriers to visibility: (1) the shift to multiple and mobile user devices; (2) the prevalence of encryption; and (3) the growth of VPNs and proxy servers.

Example of operating systems. Operating systems, by definition, have access to all the data and programs on the device. Modern devices such as smartphones are usually connected to the Internet, so it is a choice by the company that develops the operating system what data will be reported back to the company. As explained in Chapter 5, trends such as the use of personal assistants like Cortana, Google Now, and Siri mean that operating systems in practice can and do access data from a wide range of applications running on the device. In our terminology, the OS companies have “visibility” or can “see” that data, although they may decide not to actually access or “use” that data in practice.

TABLE OF CONTENTS

Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others

Summary of Contents:

Preface	2
Executive Summary	5
Appendix 1: Some Key Terms	15
Chapter 1: Limited Visibility of Internet Service Providers Into Users' Internet Activity	22
Appendix 1: Encryption for Top 50 Web Sites	36
Appendix 2: The Growing Prevalence of HTTPS as Fraction of Internet Traffic	38
Chapter 2: Social Networks	42
Chapter 3: Search Engines	50
Chapter 4: Webmail and Messaging	58
Chapter 5: How Mobile Is Transforming Operating Systems	65
Chapter 6: Interest-Based Advertising ("IBA") and Tracking	81
Chapter 7: Browsers, Internet Video, and E-commerce	89
Chapter 8: Cross-Context Tracking	100
Appendix 1: Cross-Context Chart Citations	108
Chapter 9: Cross-Device Tracking	115
Chapter 10: Conclusion	122

TABLE OF CONTENTS

Chapter 1: Limited Visibility of Internet Service Providers Into Users' Internet Activity22

- A. Users Have Multiple Devices and Are Mobile
- B. ISPs See Less Because Encryption is Becoming Pervasive
 - 1. How encryption blocks ISP visibility
 - 2. The increasing prevalence of encryption
- C. VPNs and Proxy Services Further Reduce ISP Visibility
 - 1. DNS process
 - 2. Virtual Private Networks ("VPN")
 - 3. Third-party proxy services
- D. Conclusions on the Visibility of ISPs Into Internet Usage
 - 1. Mobile and multiple devices
 - 2. HTTPS and other encryption
 - 3. VPNs and other proxy services

Appendix 1: Encryption for Top 50 Web Sites

Appendix 2: The Growing Prevalence of HTTPS as Fraction of Internet Traffic

Chapter 2: Social Networks.....42

- A. How Social Networks Gather Commercially Valuable Information
 - 1. User-generated content
 - 2. Metadata
 - 3. Logged-in users
- B. Network Effects
- C. How This Data Helps Advertisers
 - 1. Targeted advertising based on detailed profiles
 - 2. From social network to advertising network
- D. Conclusion

Chapter 3: Search Engines50

- A. Search Engine Data Flows
 - 1. Search queries and search results
 - 2. Google Search

- B. An Analogy of Search to Deep Packet Inspection (“DPI”)
- C. The Utility of Search for Targeted Advertising
 - 1. Insight into users’ intent
 - 2. Targeted advertising
 - 3. Efficient auction ecosystem
 - 4. Links with other applications, such as digital assistant programs
- D. Conclusion

Chapter 4: Webmail and Messaging58

- A. Data and Metadata in Webmail
- B. Purposes of Collecting Data
 - 1. Security
 - 2. Advertising
 - i. Google and Yahoo terms of use
 - ii. Google’s Customer Match
- C. The New Prevalence of HTTPS
 - 1. The overall shift toward encrypted email
 - 2. The Google Gmail example
- D. Conclusion

Chapter 5: How Mobile is Transforming Operating Systems65

- A. History of OS Collection of User Information
- B. How the Growth of Mobile Impacts Operating Systems
 - 1. Mobile device identifiers
 - 2. Mobile applications
 - 3. Mobile location tracking
- C. Recent OS Changes
 - 1. Apple
 - 2. Google Android
 - 3. Microsoft Windows 10
 - 4. How changes in mobile affect advertisers
- D. How ISPs Compare to Operating Systems

Chapter 6: Interest-Based Advertising (“IBA”) and Tracking81

- A. Introduction to the Online Advertising Ecosystem
 - 1. IBA example (Non-Mobile)
 - 2. IBA example (Mobile)
- B. Features of the IBA Ecosystem
- C. Limited Role of ISPs in the IBA Ecosystem

Chapter 7: Browsers, Internet Video, and E-commerce85

- A. Browsers
 - 1. Telemetry
 - 2. Private browsing
 - 3. Integration of search and other functionality
 - 4. Form autofill
 - 5. Data for advertising
 - 6. ISP lack of visibility of browser activity
- B. Internet Video
 - 1. How Internet video services collect user data
 - 2. How ISPs compare to Internet video providers
- C. E-commerce
 - 1. How e-commerce advertising data is collected
 - i. Data provided during purchase
 - ii. Appended data
 - 2. How e-commerce data affects advertising
 - 3. Why ISPs are not major e-commerce sites
- D. Conclusion

Chapter 8: Cross-Context Tracking.....100

- A. Cross-Context Tracking Data Flows
 - 1. Logged-in cross-context tracking (deterministic tracking)
 - 2. Not logged-in cross-context tracking (probabilistic tracking)
- B. Examples of Cross-Context Tracking
 - 1. Unified search and web browsers
 - 2. Combination of a social network with an advertising network
 - 3. Unified privacy policy across services

- C. Impact of Cross-Context Tracking on Advertising
- D. The Diminishing Visibility of ISPs in Cross-Context Tracking

Appendix 1: Cross Context Chart Citations

Chapter 9: Cross-Device Tracking115

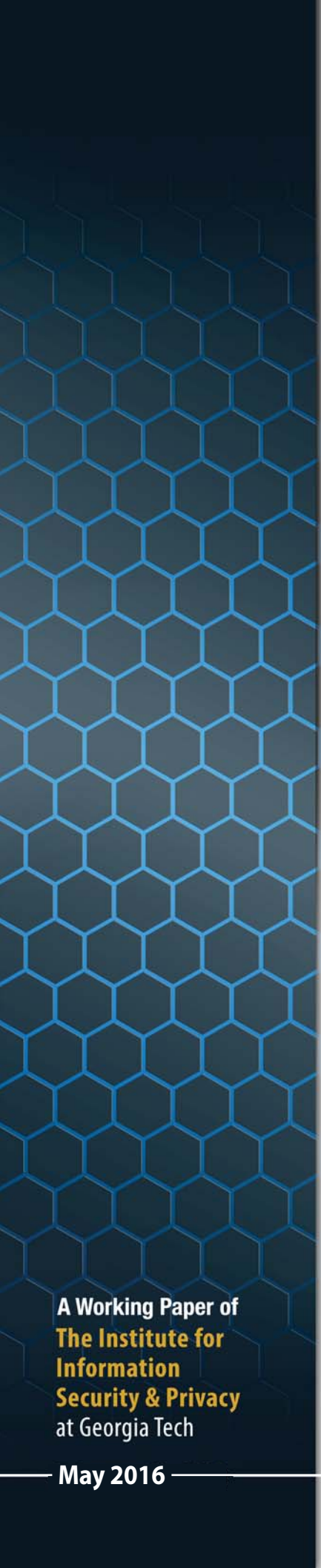
- A. Cross-Device Tracking Data Flows
 - 1. Logged-in cross-device tracking (deterministic tracking)
 - 2. Not logged-in cross-device tracking (probabilistic tracking)
- B. Impact of Cross-Device Tracking on Advertising
 - 1. Frequency capping
 - 2. Attribution
 - 3. Improved advertisement targeting
 - 4. Sequenced advertising
 - 5. Tracking simultaneity
 - 6. Summary on advertising uses
- C. The Limited Visibility of ISPs for Cross-Device Tracking

Chapter 10: Conclusion122

List of Acronyms

Addendum as of May 2016

Letter to Feamster



Limited Visibility of Internet Service Providers Into Users' Internet Activity

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 1:

Limited Visibility of Internet Service Providers Into Users' Internet Activity

This Chapter explains the increasingly limited ability of Internet Service Providers ("ISPs") to see a user's Internet activity, especially compared to the early days of the Internet. In the 1990s, ISPs at the technical level had a theoretical ability to view both metadata and content that a user accessed. Such high visibility was due to the combination of three phenomena: (1) for typical users, Internet activity was conducted through a home computer; (2) most Internet traffic was unencrypted, enabling ISPs to view both metadata (URLs) and content; and (3) ISPs typically provided both the connection to the Internet as well as the lookup service for the Domain Name System ("DNS").¹ This Chapter explains why even this theoretical ability for ISPs to see a large fraction of a user's Internet activity does not exist today. The following Chapters explain that other ecosystem participants now have the ability to see a large portion of a user's Internet activity.

In providing the last-mile connection to the Internet for their customers, ISPs carry users' data traffic on their network. In most cases, ISPs have relatively accurate information about a subscriber's name and billing address, and may have their credit card information and phone number. For URLs, an ISP previously had a theoretical capability to see a user's URLs because: (1) all or a large fraction of a user's Internet activity occurred through the home ISP; (2) the URLs were unencrypted; and (3) the DNS lookup meant that the ISP necessarily saw the URL as it connected the Internet Protocol ("IP") address of the user to the IP address of the destination server.² Technical limitations and business decisions limited the actual collection of URL data.

For unencrypted traffic, the ISPs at a technical level had at least a theoretical capability to do what is often called deep packet inspection ("DPI"). That is, an ISP, should it have chosen to do so, could have scanned the full contents of a packet coming through its system. Our understanding is that limited processing power and storage capabilities placed technological and cost limits on this capability.

Each of the above ways consumers used the Internet has greatly changed and is continuing to change. First, users today often connect to the Internet with multiple devices and from multiple locations, and at far higher speeds. This means that any single ISP views a diminishing portion of a user's Internet activity, and that the portion they do not carry represents an enormous and growing volume of data and transactions. Second, encryption is becoming pervasive, as exemplified in a 2015 *Fortune* story titled "Most Internet Traffic Will Be Encrypted by Year's End. Here's Why."³ The biggest shift is from HTTP to HTTPS (the secure version of the Hypertext Transfer Protocol, which uses Transport Layer Security, or TLS), with a resulting shift to encrypted content for webmail, social networks, search, and many other important kinds of Internet activity. With encrypted content, ISPs cannot see detailed URLs and content even if they try. Third, multiple changes, including widespread use of Virtual Private Networks ("VPNs") and third-party proxy services, are further limiting ISP visibility.

These three major and continuing technological trends have not been recognized widely enough to date. These developments contrast with the widely-held, but mistaken, view that operation of the last mile provides comprehensive and unique advantages to track consumer behavior. These claims are factually inaccurate today

¹ In this period, due to relatively slow access network speeds, the volume of data and number of transactions were manageably small, and encryption was rare.

² "Managing Domain Name Servers," *Network Solutions*, (<http://www.networksolutions.com/support/what-is-a-domain-name-server-dns-and-how-does-it-work/>).

³ Robert Hackett, "Most Internet Traffic Will Be Encrypted by Year End. Here's Why," *Fortune*, p. 4, April 30, 2015, (<http://fortune.com/2015/04/30/netflix-internet-traffic-encrypted/>).

and in the future.⁴ Changing technology and market practices create effective barriers to ISPs' visibility into their users' Internet activities.

Taken together, the technological developments described in this Chapter mean that providing the last-mile connection – the primary role of an ISP – enables the ISP to see a far smaller portion of a user's URLs and content than was previously possible. And the strong trends toward multiple devices and connections, encryption, and changes in VPNs and proxy servers are likely to continue.

Recently, some ISPs have introduced optional programs that include the use of a customer's URL history for customized advertising, to the extent that the ISP has access to that data.⁵ In addition, public WiFi services provided by retailers and other third parties on a free, ad-supported basis often use data collected from network usage for customized advertising.⁶ This may include the user's URL history while using the public WiFi service. These types of ISP programs are similar to the interest-based advertising programs discussed in Chapter 6, but are subject to the technical limitations discussed in this Chapter – encryption, multiple devices and ISPs, and changes in VPNs and proxy servers.

A. Users Have Multiple Devices and Are Mobile

The first technological development since the 1990s that reduces ISP visibility into a user's history of Internet activity is intuitive – any single ISP has a far less comprehensive view of a user's Internet activity than it theoretically used to have (or could have had if it had made the large investments to do so). In the early days of the Internet, most individuals worked on desktop computers, which were fixed in place and typically connected via a single ISP. Some Internet use was done via laptops, which often would log-in from one or two locations, such as home and work. Even for laptops, a large fraction of individuals' computing occurred through dial-up, so that the user would often rely on a single ISP such as AOL or AT&T Worldnet.

ISPs today face a more fractured world. The personal computing world is becoming more mobile, with tablet sales as large as laptops and desktops combined,⁷ and the share of desktop computers continuing to fall.⁸ More generally, the use of smartphones has exploded as a portion of users' overall Internet activity.

⁴ For instance, the FTC's 2012 Privacy Report stated that ISPs were able to comprehensively track consumers' online activity and "develop highly detailed and comprehensive profiles of their customers," which can then be used for marketing and advertising purposes. "Protecting Consumer Privacy in an Era of Rapid Change," *Federal Trade Commission*, p. 56, March 2012, (<https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf>).

⁵ See, e.g., "U-verse with AT&T GigaPower Internet Preferences Program," *AT&T*, (<https://www.att.com/esupport/article.html#!dsl-high-speed/KM1011211>) (opt-in program that uses a web browsing information from a customer's Gigabit Internet access service in exchange for a monthly discount on the top speed tier of service to customize ads and marketing offers); "Verizon Selects Program FAQs," *Verizon*, (<http://www.verizonwireless.com/support/verizon-selects-faqs/>) (opt-in program that uses web browsing information and other data from a customer's wireless service to personalize marketing offers); Sprint Mobile Advertising Program, *Sprint*, (http://newsroom.sprint.com/article_display.cfm?article_id=1623#ad) (opt-in program that uses web browsing information and other data from a customer's wireless service to serve more relevant ads).

⁶ See, e.g., "Enjoy Fast WiFi," *Starbucks*, (free Starbucks Wi-Fi provided by Google) (<http://www.starbucks.com/coffeehouse/wireless-internet>) and LinkNYC free Wi-Fi service in New York City, which is provided by CityBridge (<http://www.link.nyc/>).

⁷ Angela Moscaritolo, "Tablets to Make Up Half the PC Market in 2014," *PCMag*, Nov. 26, 2013, (<http://www.pcmag.com/article2/0,2817,2427623,00.asp>).

⁸ Jordan Weissman, "The End of the Home Computer: Why PC Sales Are Collapsing," *The Atlantic*, April 11, 2013, (<http://www.theatlantic.com/business/archive/2013/04/the-end-of-the-home-computer-why-pc-sales-are-collapsing/274899/>).

As discussed throughout this Working Paper, users constantly shift from one ISP to another, not just for home and work, but among many WiFi hotspots and other locations from which they connect to the Internet.⁹ An estimated 46 percent of all mobile data traffic was offloaded to WiFi networks in 2014,¹⁰ which is projected to increase to 60 percent by 2020.¹¹ The old image was a pipe from the Internet to the user's home computer, with the pipe controlled by the ISP. The image today, for an ISP, is only episodic glimpses of any particular device.¹² In addition, as discussed in Chapter 9 on cross-device tracking, the number and variety of devices is exploding as we enter the era of the Internet of Things,¹³ so that the ability of any one ISP to see a user's total activity diminishes even further.

Just as consumers switch among ISPs during the day, they also switch ISP subscriptions more often than many would think. Mobile already constitutes a majority of broadband Internet activity¹⁴ and multiple mobile carriers are available to most American consumers.¹⁵ According to the FCC, between a fifth and a third of wireless subscribers switch their carriers annually.¹⁶ Moreover, wireline switching is substantial – one out of six customers switches wireline providers every year, and 37 percent of customers will switch every three years.¹⁷

B. ISPs See Less Because Encryption Is Becoming Pervasive

Encryption is pervasively limiting the ability of ISPs to see user Internet activity. A major reason is the shift from the traditional Internet protocol (HTTP or Hyper Text Transfer Protocol) to an encrypted protocol (HTTPS or Hyper Text Transfer Protocol Secure). We first explain how HTTPS and other encryption on both browsers and apps blocks

⁹ At the beginning of 2015, one study showed that 91 percent of users owned a desktop or laptop. Smartphone use has climbed sharply, to 80 percent. In addition to desktops, laptops, and smartphones, nearly 50 percent of users reported owning a tablet, 37 percent owned gaming consoles, and another 34 percent reported owning smart televisions. See Jason Mander, "80% of internet users own a smartphone," *GlobalWebIndex*, Jan. 5, 2015, (<http://www.globalwebindex.net/blog/80-of-internet-users-own-a-smartphone>).

¹⁰ "Cisco Visual Networking Index, Forecast and Methodology, 2014-2019 Working Paper," *Cisco*, May 27, 2015, (http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html).

¹¹ "Juniper Mobile Data Onload & Offload Report (June 2015)," *Juniper*, (<http://www.juniperresearch.com/researchstore/enabling-technologies/mobile-data-onload-offload/wifi-small-cell-network-strategies>).

¹² The episodic glimpses are more continuous when the user switches from one connection (such as home) with a particular ISP, and goes to another connection (such as a WiFi hotspot) with the same ISP.

¹³ The Internet of Things era has taken off in recent years. In 2011, the number of "things" (rather, devices connected to the Internet) surpassed the number of people. Due to the rapid emergence of these new technologies, the FTC hosted its first workshop specifically addressing the Internet of Things on November 19, 2013. The title of this workshop was "The Internet of Things: Privacy and Security in a Connected World." In the FTC Staff Report that accompanied the workshop, the FTC cited statistics stating that by the end of 2015, there would be 25 billion connected devices, and by 2020, 50 billion. Additionally, there are estimates that by 2020, 90 percent of consumer cars will have an Internet connection, an increase from the 10 percent in 2013. All of these statistics demonstrate that there has been and there will continue to be an increasing variety of connected devices. See "Internet of Things: Privacy and Security in a Connected World," *Federal Trade Commission*, p. 1, Nov. 19, 2013, (<https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf>).

¹⁴ Katie Benner & Conor Dougherty, "Publishers Straddle the Apple-Google, App-Web Divide," *The New York Times*, Oct. 18, 2015, (http://www.nytimes.com/2015/10/19/technology/publishers-straddle-the-apple-google-app-web-divide.html?_r=0).

¹⁵ 82 percent of mobile broadband Internet users have a choice of at least four providers, and 98.8 percent have at least two. See "Seventeenth Annual Mobile Wireless Competition Report," *Federal Communications Commission*, DA 14-1862 ¶ 51, rel. Dec. 18, 2014, (https://apps.fcc.gov/edocs_public/attachmatch/DA-14-1862A1.pdf); "2015 Broadband Progress Report and Notice of Inquiry on Immediate Action to Accelerate Deployment," *Federal Communications Commission*, FCC 15-10 109, rel. Feb. 4, 2015, (https://apps.fcc.gov/edocs_public/attachmatch/FCC-15-10A1.pdf).

¹⁶ "Annual Report and Analysis of Competitive Market Conditions with Respect to Mobile Wireless, Including Commercial Mobile Services: Fifteenth Report," *Federal Communications Commission*, June 27, 2011, (https://apps.fcc.gov/edocs_public/attachmatch/FCC-11-103A1.pdf).

¹⁷ "Broadband Decisions: What Drives Consumers to Switch-or Stick with-Their Broadband Internet Provider," *Federal Communications Commission*, Dec. 2010, (https://apps.fcc.gov/edocs_public/attachmatch/DOC-303264A1.pdf).

ISP visibility of user content and full URLs.¹⁸ We next document the recent and growing dominance of encryption, technologically blocking ISPs from viewing most user activity, including content often characterized as sensitive.

1. How encryption blocks ISP visibility

Diagram 1-A helps us explain the role of the ISP and how it changes with the shift to HTTPS.

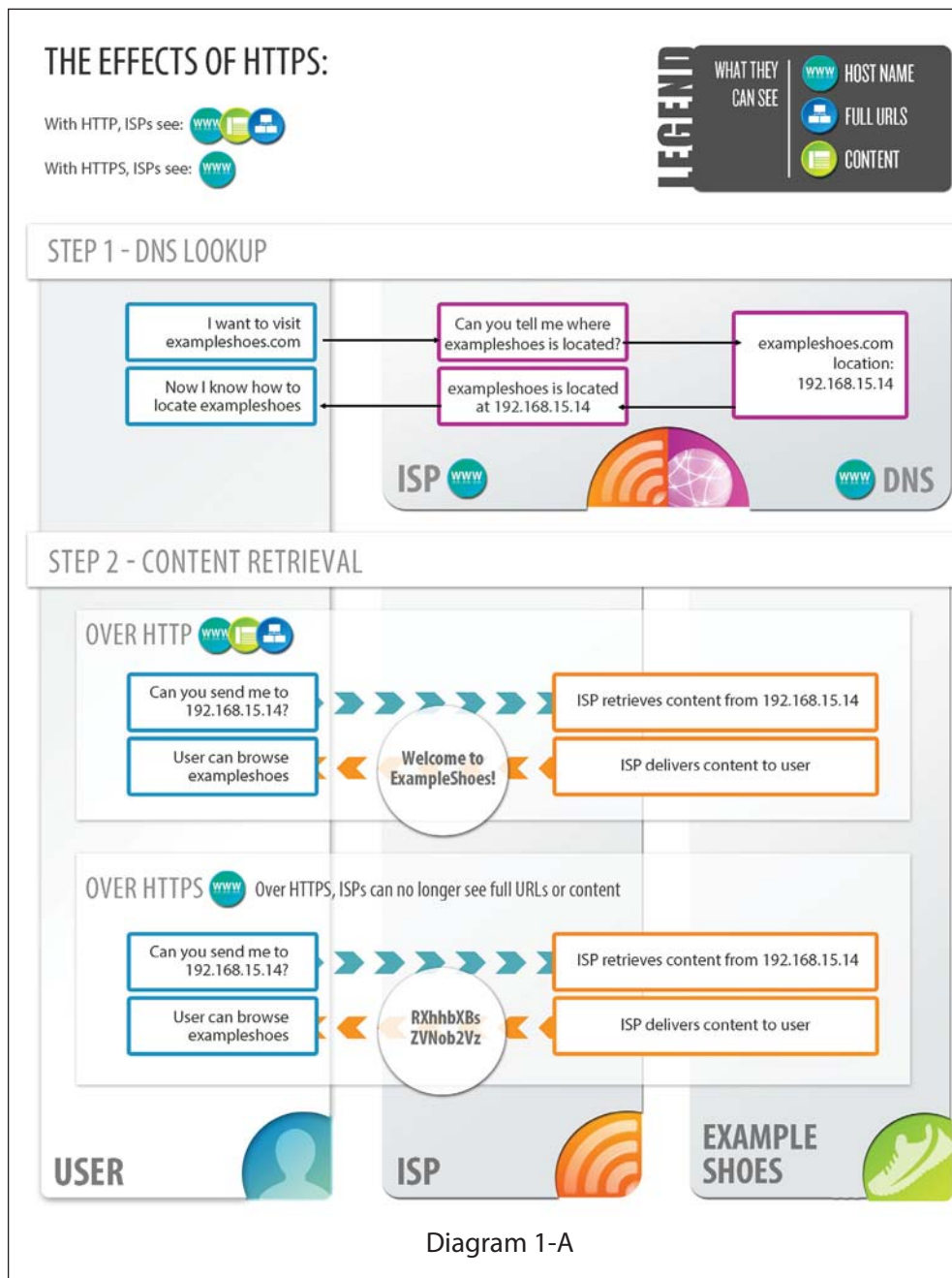


Diagram 1-A begins with the user wishing to visit ExampleShoes.com. The first stage is what is called DNS lookup, in the top right corner of the Diagram. The user types a human-friendly version of the domain name into their web browser: ExampleShoes.com. The ISP receives that request and looks up the IP address for that domain, which in the Diagram is 192.168.15.14. Below, we discuss changes in DNS lookup.

¹⁸ The encryption also blocks access in transit to the content and full URLs to others, including cyber-criminals, government agencies, and anyone else who lacks keys to the encrypted traffic.

The user's computer next sends its request for 192.168.15.14 to the ISP,¹⁹ which retrieves content from that IP address and routes it back to the user. That role of the ISP is shown in the middle of the Diagram, near the bottom and above ExampleShoes.com.²⁰

The bottom right corner contrasts the visibility of the ISP under HTTP vs. HTTPS. Under the long-used HTTP protocol, the ISP had the theoretical ability to see all of the bits flowing between ExampleShoes.com and the user:

1. The host name (or "domain name") such as <http://ExampleShoes.com>.²¹
2. The full URL (or "detailed" URL), such as <http://ExampleShoes.com/sneakers/women/white.htm>.
3. The content, such as photographs and text sent to the user concerning the white sneakers.

In the Diagram, these three categories of visibility for HTTP browsing are shown with three icons, for host name, full URL, and content.

By contrast, the newer and now widely-adopted HTTPS blocks the ISP from seeing the full URL and the content. As indicated in the bottom right corner of the Diagram, the ISP can see only the host name²² when data flows between ExampleShoes.com and the user.²³

We provide a basic explanation of HTTPS here,²⁴ and those interested can find more detailed explanations easily. A user starts by typing a URL into her laptop's web browser.²⁵ This HTTPS page will then be encrypted by a secure protocol using Transport Layer Security,²⁶ which establishes a "handshake" between the user and the site. Creating that handshake involves the exchange of encryption keys between the user and the site.²⁷ Once the handshake is established, the user's browser and the website can exchange communications securely. Any entity between the two, such as the ISP, can see only encrypted bits as they flow through the Internet. The ISP thus sees, at most, the IP address to which the user originally connected, such as the web server for ExampleShoes.com, information which is necessary in order to correctly route the packets.

¹⁹ This IP address is for illustrative purposes only and is part of private IPv4 addresses per RFC 1918 (<https://tools.ietf.org/html/rfc1918>).

²⁰ The Diagram omits some complexity. In modern practice, much content reaches the ISP from a Content Delivery Network ("CDN"), such as Akamai, Amazon CloudFront, and AT&T's Digital Media Solutions. The CDNs may manage delivery of content from a website such as ExampleShoes.com. In that event, the response to the ISP domain name lookup is dynamically generated based on the source network, geographic location, time of day, and other factors. These CDNs themselves may gain information about individual users that can be used for advertising purposes.

²¹ One technical term for what this paper often calls the "host name" is a fully qualified domain name ("FQDN").

²² The ISPs can in many cases also see other information such as the size, timing, and destination of packets. The commercial usefulness of this data will depend on how useful such data is compared to going to the website visited by the user (or others in the ecosystem who may have access to it), which will have more complete information about the content and detailed URLs that the user has accessed.

²³ Even many requests to the host are often invisible to the ISP due to local caching – often, an ISP can see the first request to a web site in the day, but caching technology means that subsequent requests by the user are routed by an entity other than the ISP. Yu Ng, "In the World of DNS, Cache is King," *catchpoint*, July 15, 2014, (<http://blog.catchpoint.com/2014/07/15/world-dns-cache-king/>).

²⁴ For a basic text introduction, see "What is HTTPS?" *Instant SSL by Comodo*, (<https://www.instantssl.com/ssl-certificate-products/https.html>). For a short video explaining HTTPS, see https://www.youtube.com/watch?v=_p-LNLv49Ug.

²⁵ The user may type in an HTTPS URL herself, or the URL may convert to HTTPS when the user clicks "enter" because the website provides encryption by default (e.g., financial websites). Appendix 1 shows that roughly half of the top 50 web sites currently provide encryption by default.

²⁶ TLS is a successor to the previous Secure Sockets Layer ("SSL") protocol, which is no longer considered secure.

²⁷ Both TLS and SSL rely on an asymmetric Public Key Infrastructure ("PKI") system, where each side shares its "public" key but keeps its "private" key secret. "Anything encrypted with the public key can only be decrypted by the private key, and vice-versa." "What is HTTPS?" *Instant SSL by Comodo*, (<https://www.instantssl.com/ssl-certificate-products/https.html>). When a user requests an HTTPS connection, the website initially sends its TLS certificate to the user's browser. The certificate contains the public key needed to begin the session.

2. The increasing prevalence of encryption

The prevalence of HTTPS and other encrypted web traffic has grown enormously and is expected to continue to grow. In the 1990s, encrypted URLs and content were relatively rare. By 2015, the extent of the change was captured in the title of a Fortune magazine article: “Most Internet Traffic Will Be Encrypted by Year End. Here’s Why.”²⁸

The evidence of widespread encryption is clear, including much of the individual user traffic of greatest interest to online advertisers:

1. Encryption today is pervasive for banking, e-commerce, and other websites that gather or transmit financial information. Encrypted payment information paved the way in many settings for more widespread use of encryption.
2. Appendix 1 shows the dominance of encrypted traffic for the major consumer sites today.²⁹ Based on Alexa’s ranking of the top 50 Internet sites:
 - a. All of the top 10 sites either use HTTPS by default or shift to HTTPS when the user logs-in;
 - b. 42 of the top 50 sites either use HTTPS by default or shift to HTTPS when the user logs-in; and
 - c. 24 of the top 50 sites use HTTP by default.
3. Appendix 2 shows the rapid and recent growth of HTTPS according to Internet backbone data from the Center for Applied Internet Data Analysis (“CAIDA”). The first chart shows the large increase in share of HTTPS in the past two years:
 - a. In early 2014, HTTPS was a small fraction of total traffic. For April 2014, HTTPS was 13.3 percent.³⁰
 - b. HTTPS traffic increased considerably by the beginning of 2015.
 - c. By early 2016, HTTPS traffic was greater than HTTP traffic.
4. Appendix 2 also shows detailed data for a large sample of Internet backbone traffic in one week in February 2016:
 - a. HTTPS traffic was 38 percent larger than HTTP traffic.³¹
 - b. HTTPS traffic accounted for nearly half of total traffic (48.6 percent).³²
3. Sandvine studies confirm the prevalence of encrypted traffic:
 - a. By early 2015, a majority of non-video web traffic was already encrypted.³³ “Non-video” is an important category because video constitutes such a large fraction of bits flowing to consumers over the Internet.
 - b. Encryption is spreading to more video. Sandvine has reported that, by the end of 2016, more than two-thirds of North America’s Internet traffic will be encrypted due to Netflix’s decision to use HTTPS,³⁴ although only about 10 percent of Netflix’s traffic was encrypted as of February 2016.³⁵

²⁸ *Id.*

²⁹ Alexa lists the 50 top websites. For each site, we had a researcher visit to determine the status of encryption as of February 2016.

³⁰ Diagram 3 in Appendix 2 is a screen shot showing the share of HTTPS traffic in April 2014.

³¹ An average of 48.6 percent of HTTPS and 35.1 percent of HTTP. $48.6/35.1 = 1.38$, or 38 percent higher for HTTPS.

³² Diagram 4 in Appendix 2 is a screen shot showing the share of HTTPS traffic in the week in February 2016.

³³ “Global Internet Phenomena Spotlight: Encrypted Internet Traffic,” *Sandvine*, April 2015.

³⁴ *Id.*

³⁵ “2016 Global Internet Phenomena, Spotlight: Encrypted Internet Traffic,” *Sandvine*, Feb. 2016.

- c. As of February 2016, Sandvine forecasts that 70 percent of global Internet traffic will be encrypted in 2016, with many networks exceeding 80 percent.³⁶

Along with this evidence of the decisive shift toward encryption, multiple reasons are pushing toward even more pervasive encryption:

1. In the wake of the revelations by Edward Snowden that began in 2013, global technology companies have embraced encryption as a way to highlight to consumers that their services can be trusted by users globally.³⁷ The “Let’s Encrypt” project from the Internet Security Research Group, for instance, is a program designed to make HTTPS easy to enable for all websites, and is sponsored by major technology companies.³⁸
2. High-profile technology organizations such as the Internet Society and the Internet Engineering Task Force (“IETF”) have strongly supported encryption: “The Internet Society believes that encryption should be the norm for Internet traffic.”³⁹ The IETF has released two protocols supporting encryption by default, RFC 6973, “Privacy Considerations for Internet Protocols” and RFC 7258, “Pervasive Monitoring Is an Attack.”⁴⁰
3. According to the Media Trust Company: “In 2015, several of the largest advertising platforms encouraged the adoption of HTTPS-encrypted advertisements across mobile, desktop, and video display.”⁴¹ According to the Interactive Advertising Bureau, almost 80 percent of its members’ ad delivery platforms support encrypted ads.⁴²
4. As part of its “HTTPS Everywhere” campaign, in 2014 Google announced that it would begin using HTTPS as a ranking signal, to encourage website owners to switch from HTTP to HTTPS and promote security.⁴³ This announcement means that if two websites are relevant to a search query, an HTTPS website will rank higher than a comparable HTTP website.⁴⁴ Since Google has a large market share in search, this announcement creates a commercial motive for websites to switch to HTTPS to achieve higher search rankings.
5. A recent “Worldwide Survey of Encryption Products” documented 865 hardware or software products incorporating encryption from 55 different countries, including 546 encryption products from outside of the United States.⁴⁵ A third of the non-U.S. products surveyed are open source. The widespread availability of

³⁶ *Id.*

³⁷ For a collection of sources about increasing industry encryption since Snowden, see Peter Swire, Testimony before the Senate Judiciary Committee on “Going Dark: Encryption, Technology, and the Balance between Public Safety and Privacy,” at fn. 18, July 8, 2015 (<http://www.judiciary.senate.gov/imo/media/doc/07-08-15%20Swire%20Testimony.pdf>).

³⁸ “Let’s Encrypt,” *Let’s Encrypt*, (<https://letsencrypt.org>).

³⁹ Internet Society, “Encryption,” *Internet Society*, (<https://www.internetsociety.org/encryption>).

⁴⁰ IETF RFC 6973, “Privacy Considerations for Internet Protocols,” (<http://tools.ietf.org/html/rfc6973>), IETF RFC 7258, “Pervasive Monitoring Is an Attack,” (<http://tools.ietf.org/html/rfc7258>).

⁴¹ “Media Scanner™ Enforces and Maintains Encryption Compliance” The Media Trust, (<https://themediatruster.com/media-scanner-for-encryption-compliance-ad-tags.php>); “A Reminder for Advertisers: Better Encryption, Better Creative, Better Reach,” *Yahoo*, Feb. 25, 2015 (<https://advertising.yahoo.com/Blog/REMINDER-FOR-ADVERTISERS.html>).

⁴² Lucian Constantin, “Google’s Push to Encrypt Ads Will Improve Security, But Won’t Kill Malicious Advertising,” *PCWorld*, April 20, 2015 (<http://www.pcworld.com/article/2912092/googles-push-to-encrypt-ads-will-improve-security-but-wont-kill-malvertising.html>).

⁴³ “HTTPS as a Ranking Signal,” *Google Webmaster Central Blog*, Aug. 6, 2014, (<http://googlewebmastercentral.blogspot.com/2014/08/https-as-ranking-signal.html>).

⁴⁴ Google’s ranking search algorithm has numerous “signals,” such as the web page age, user friendly layout, placement of keywords, etc. Since 2014, HTTPS has become one of the signals.

⁴⁵ Bruce Schneier, Kathleen Seidel & Saranya Vijayakumar, “A Worldwide Survey of Encryption Products,” Feb. 11, 2016, (https://www.schneier.com/cryptography/archives/2016/02/a_worldwide_survey_o.html).

encryption, including from open source projects, means that any U.S. or other government efforts to reverse the trend toward encryption will be extremely difficult to achieve.

All of these recent developments contrast with earlier findings in separate articles by Christopher Soghoian⁴⁶ and Peter Swire⁴⁷ that pervasive encryption was developing but was not yet in place in 2012, the year when the FTC Privacy Report stated that ISPs could develop “comprehensive profiles” of their customers. Today, the norm has become that deep links and content are encrypted on the Internet. For all of the encrypted links and content, ISPs are technically blocked from seeing user data.

C. VPNs and Proxy Services Further Reduce ISP Visibility

This section discusses how the increasing availability of competing proxy services, combined with encryption, further reduces the ability of ISPs to see the URLs and content accessed by their subscribers. As shown in Diagram 1-A, ISPs have always integrated DNS lookup services with Internet access. DNS lookup converts human-readable website names (i.e., the domain name) to computer-readable numerical IP addresses. Where the ISP also provides DNS lookup, which is currently the typical situation, the ISP necessarily sees the level of detail that a user provides to route the user to the desired web location. DNS lookup thus provided ISPs with the ability to see a substantial portion of a user’s URLs.

Increasingly, however, the user has encrypted communications with a third-party proxy, and it is the other DNS server rather than the ISP that performs DNS lookup. The discussion of encryption just explained that protocols such as HTTPS block ISPs from the content and detailed URLs, but still allow the ISP to see the host name the user visits, such as www.ExampleShoes.com. Proxy services go a step further and block ISP visibility even for the host name. We now explain how this works for VPNs and third-party proxy services.

1. DNS process

An ISP previously could have visibility into a user’s history of URLs because a user’s ISP provided the DNS server that would connect the IP address of the user to the IP address of the server.⁴⁸ For example, as shown in Diagram 1-A, a user might type www.ExampleShoes.com into her computer. The user’s computer and router cannot read the domain name, but instead require an IP address in order to transmit the communication from the user’s computer to the Example Shoes site. Therefore, traditionally, the user would use the DNS lookup service provided by her ISP, and the ISP would see that request to look up the name www.ExampleShoes.com, as well as the date and time that this request was made. The ISP would then resolve the name www.ExampleShoes.com to an IP address and the user would be able to visit this particular web page. In the era of unencrypted traffic, the ISP would have the technical capability to see the detailed URL as well, such as www.ExampleShoes.com/sneakers.

⁴⁶ Christopher Soghoian, “Caught in the Cloud: Privacy, Encryption, and Government Back Doors in the Web 2.0 Era,” 8 *Journal on Telecommunications and High Technology Law* 8, 359 (2012), available at <http://ssrn.com/abstract=1421553>.

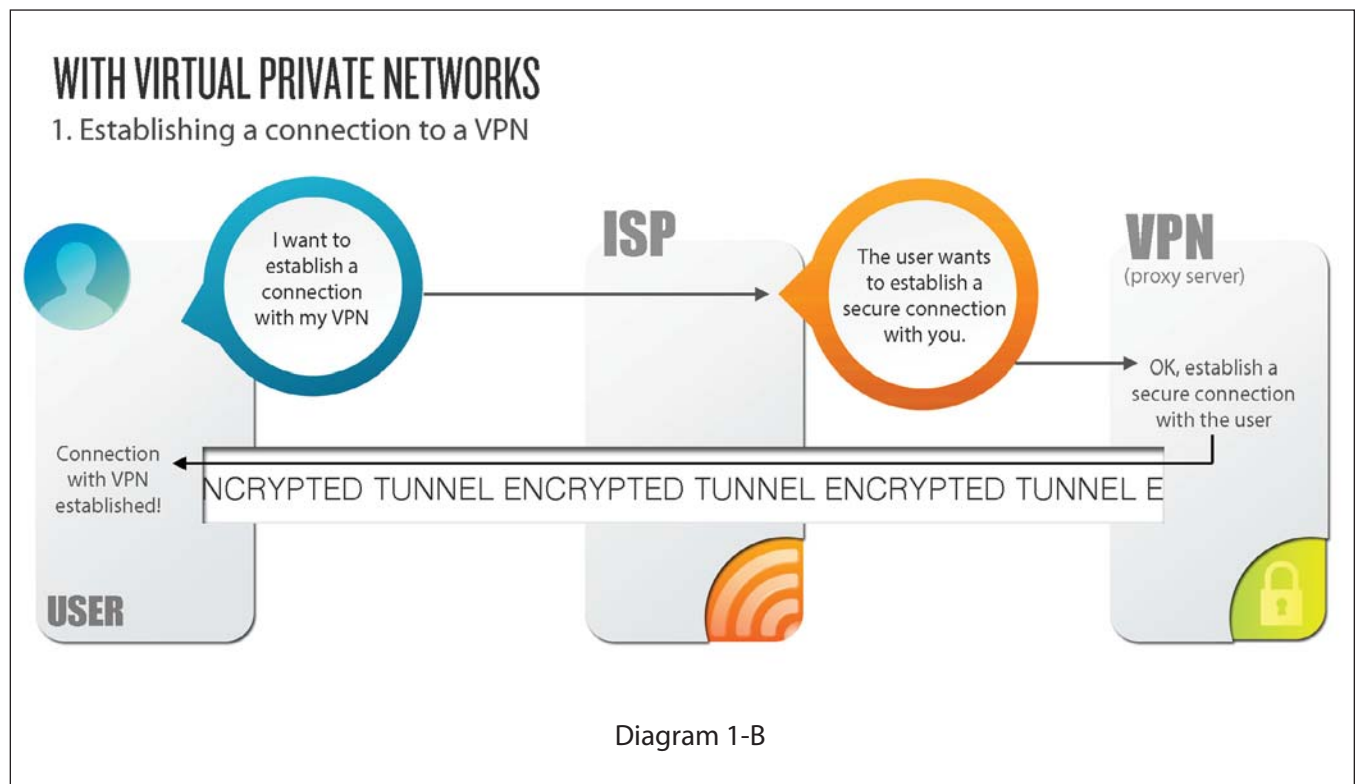
⁴⁷ Peter Swire, “From Real-Time Intercepts to Stored Records: Why Encryption Drives the Government to Seek Access to the Cloud,” *International Data Privacy Law* (2012), available at <http://ssrn.com/abstract=2038871>.

⁴⁸ ISPs cannot see every request as the DNS lookup is locally cached (i.e., remembered). “Time to Live” (“TTL”) prevents computing systems from having to go through the entire DNS process each time a website is requested. Once there has been a request for a web page, there is a question as to whether the computer will have to re-ask the DNS server for the same web page each time the user wants to visit that particular web page or if the computer will remember the web page internally. TTL dictates how long it will be until a computer refreshes its DNS related information by specifying the number of seconds the computer record can be cached (i.e., remembered) by the DNS server. The amount of time can range from zero seconds (no DNS memory) to hours, days, and weeks. Yu Ng, “In the World of DNS, Cache is King,” *catchpoint*, July 15, 2014, (<http://blog.catchpoint.com/2014/07/15/world-dns-cache-king/>).

2. Virtual Private Networks (“VPNs”)

VPNs are a prominent example of a service that uses encryption and third-party proxy servers.⁴⁹ Where VPNs are in place, the ISPs are blocked from seeing the deep links and content (as with other encryption), and also the domain name the user visits.⁵⁰

Diagram 1-B shows that the first stage is for the user to establish a secure connection with the VPN server. When the secure connection is established, the ISP can see the URL for the proxy server’s domain. Next, the handshake takes place between the user and the proxy server, which establishes what the Diagram calls an “encrypted tunnel” between the user and the proxy server, blocking all further visibility of the ISP into what the user accesses.⁵¹ In fact, this is a virtual tunnel, not a physical one, but the effect on the ISP is the same – the ISP sees only encrypted 1s and 0s and cannot see the domain name, deep links, or content the user visits.

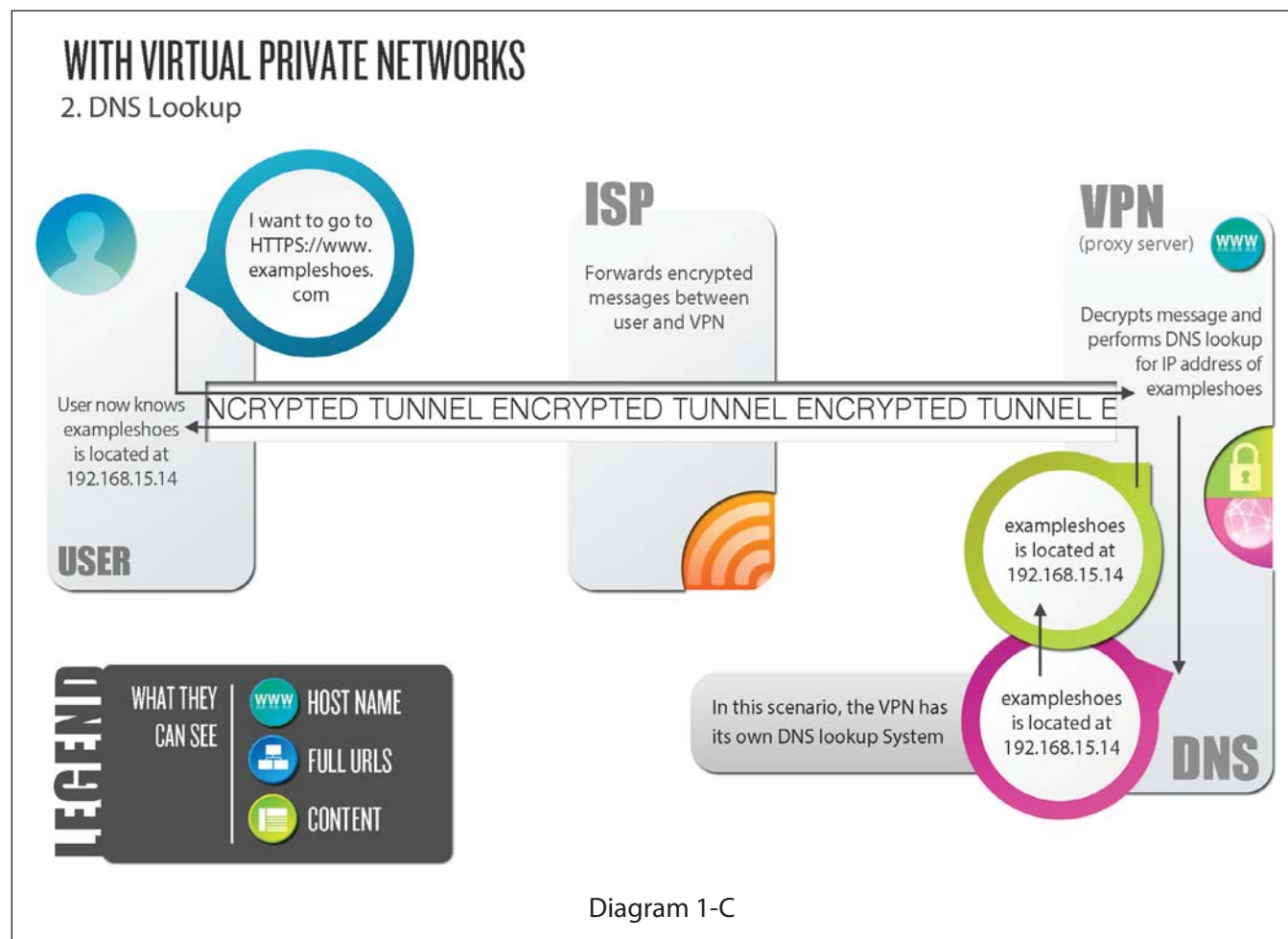


⁴⁹ A proxy server is a server that acts as an intermediary “between an endpoint device, such as a computer, and another server from which a user . . . is requesting a service.” “Proxy Server,” *WhatIs.com*, p. 1, (<http://whatis.techtarget.com/definition/proxy-server>).

⁵⁰ As noted in the Upturn report, a properly configured VPN will mask DNS queries, though not all VPNs are so configured. See, What ISPs Can See, *Upturn*, Mar. 2016, (<https://www.teamupturn.com/reports/2016/what-isps-can-see>). For example, Cisco’s support for its VPN client details the multiple ways DNS lookups can operate, depending on multiple factors. See, What ISPs Can See, *Upturn*, Mar. 2016, (<https://www.teamupturn.com/reports/2016/what-isps-can-see>). While this is true, the remainder of this section will describe how a VPN properly configured to mask DNS lookup will operate.

⁵¹ Stephanie Crawford and Jeff Tyson, “How VPNs Work,” *HowStuffWorks.com*, April 14, 2011, (<http://computer.howstuffworks.com/vpn5.htm>). For another succinct explanation of VPNs, see “Virtual Private Networks,” (<https://technet.microsoft.com/en-us/library/cc977889.aspx>).

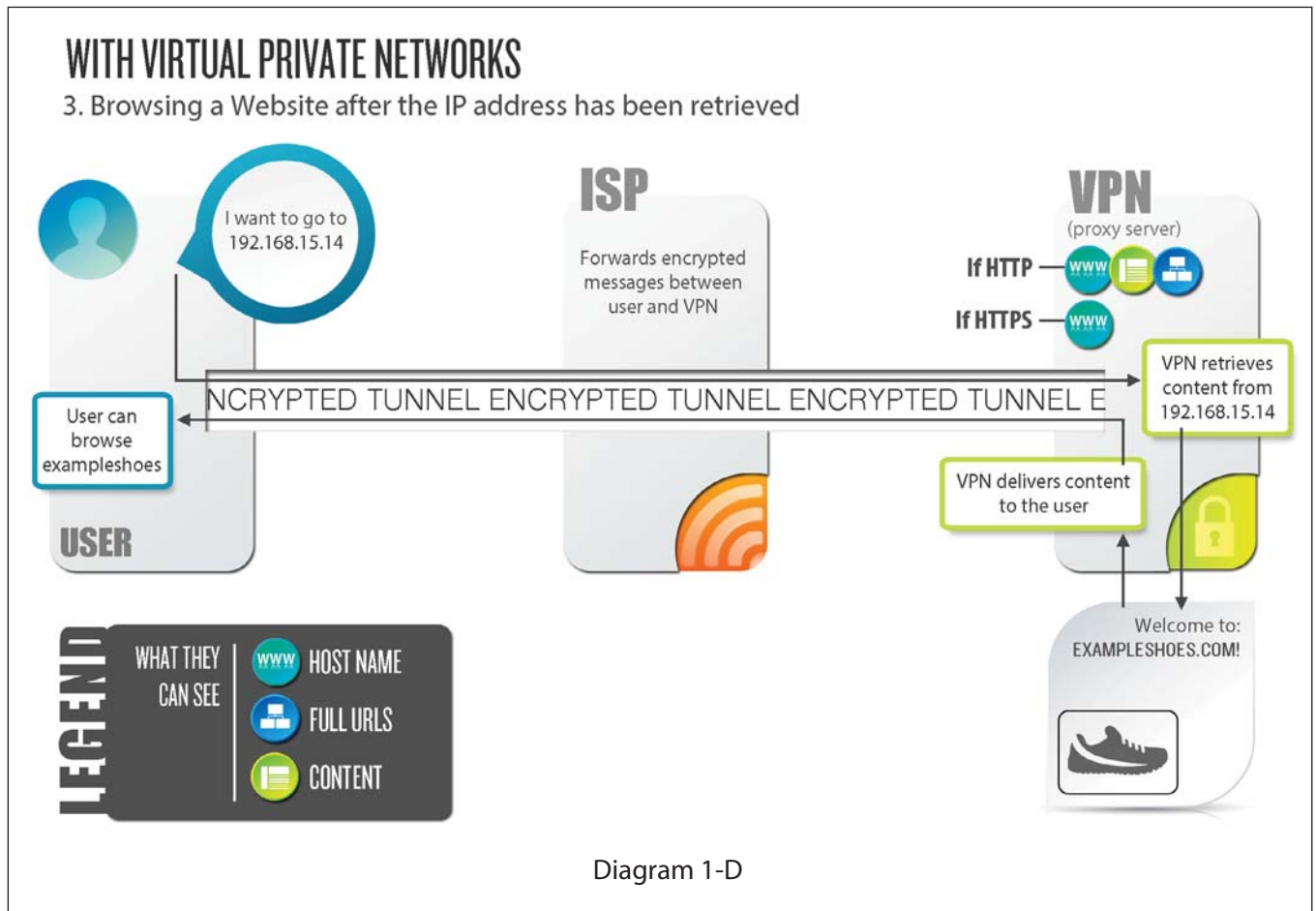
Diagram 1-C shows what happens after the secure tunnel is established between the user and the VPN (proxy server). Diagram 1-C shows that the DNS lookup happens at the proxy server. The user types in www.exampleshoes.com and the proxy server looks up the IP address and learns that it is 192.168.15.14. As is true generally once a VPN is established, the ISP does not see the domain name, deep links, or content.⁵² Diagram 1-C illustrates, however, how a traditional role of an ISP – looking up the IP address to route the user to a website – gets shifted to the entity operating the proxy server.⁵³



⁵³ The visibility of the ISP, or the VPN in this example, also exists for what is called a “5-tuple.” An ISP can see the source IP address/port number, destination IP address/port number and the protocol in use; these five pieces of information are summarized as a 5-tuple, when a user connects to a given web site. These pieces of data are required for the ISP to perform basic routing functions – that is, for the Internet as we know it to work (i.e., to process packets to/from the user and to offer connectivity). The source IP address is issued by the ISP to the user. The destination IP address indicates the address of the site the user is going to. The port numbers and protocol give an indication of what type of application is being run (e.g. mail, or WWW, or video) but generally not the specific application itself or what the content is. The 5-tuple is the key data exported from networking equipment via the commonly used IPFIX protocol (RFC 5101). As explained in RFC 5472, “IP Flow Information Export (IPFIX) Applicability,” IPFIX (also known as Netflow) data is most useful for network management but lacks the precision for reliably determining specific applications and usage. Further, when an encrypted VPN/proxy is used, the information about the website is encrypted and is no longer visible to the ISP -- the only 5-tuple the ISP sees is that of the proxy/VPN, the encrypted connection (i.e. 5-tuple showing a link from the user to the proxy/VPN) not that of the traffic flowing through the VPN.

In comparing the visibility of an ISP to other players in the ecosystem, the 5-tuple lacks the precision for reliably determining specific application and usage. The destination IP address/port number is known to the destination, such as the destination web site, as well as any entity downstream that learns the information from the destination website. The source IP address/port number is also known to the destination – that information is logged by the destination as part of the basic operation of the Internet. In the terms used in this Working Paper, the 5-tuple is far from unique to ISPs, and provides little information beyond the destination IP address.

Diagram 1-D shows how the spread of encryption affects the VPN proxy server. Where the connection is via HTTP, the proxy server stands in the same position that the ISPs used to have – the proxy server can see the domain name, deep links, and content. Where the connection is via HTTPS, by contrast, the proxy server can still see the domain name, but it is blocked from seeing the deep links and content. This last point underscores an effect of using HTTPS – only the user and the website, and not intermediate entities such as the ISP and the proxy server, can see the deep links and content.



VPNs can also secure content through Simple Mail Transfer Protocol Secure (“SMTPS”) and IP Security (“IPSEC”). SMTPS is a secure standard for Simple Mail Transfer Protocol (“SMTP”) a protocol for sending email messages between servers.⁵⁴ Many webmail systems use SMTPS to send messages from one server to another server. The connection is secured by SSL or TLS, meaning that the webmail communications are encrypted. IPSEC is a secure protocol that supports two encryption modes: transport and tunnel.⁵⁵ Transport mode encrypts the data portion of the packet, but the header is untouched.⁵⁶ Tunnel mode is more secure and encrypts both the data portion and the header.⁵⁷ Both SMTPS and IPSEC contribute to the fact that ISPs are not able to view the history of a user’s content because the content and sometimes even header are encrypted.

⁵⁴ Vangie Beal, "SMTP-Simple Mail Transfer Protocol," *Webopedia*, (<http://www.webopedia.com/TERM/S/SMTP.html>).

⁵⁵ Vangie Beal, "IPsec," *Webopedia*, (<http://www.webopedia.com/TERM/I/IPsec.html>).

56 *Id.*

57 *Id.*

VPN use is substantial and growing. Many corporations deploy VPNs as a security measure to prevent information about their employees' web communications from being visible to outside parties. A 2014 survey by GlobalWebIndex indicated that VPN use was increasing globally in response to privacy concerns stemming from government surveillance revelations and increasing worries about the way commercial entities were tracking users, as well as to circumvent geographic content restrictions.⁵⁸ This was further supported by a 2015 report that indicated several trends, including the use of VPN and third-party proxy servers (discussed below), were causing hundreds of millions of internet users to become invisible in traditional internet studies.⁵⁹

3. Third-party proxy services

VPNs are one type of third-party proxy service, whose use is familiar to many as a security tool for communicating between corporations and employees. Third-party proxy services are becoming more prevalent for other reasons as well, including faster browsing. At least 30 million people in the U.S. use a VPN or other proxy service,⁶⁰ amounting to 79 million connections and 2,779 terabytes.⁶¹ This number likely will climb sharply in coming years.

Google, for instance, has introduced its "Data Saver" proxy service, to provide "faster, safer, and cheaper web browsing."⁶² The service acts as an intermediary between the user's browser and website destination, performing DNS requests, retrieving Internet content on behalf of the user, and encapsulating all Internet content inside of an encrypted tunnel.⁶³ Traditional web protocols have been efficient at transferring an individual file, but they can have difficulty transferring a large number of small files.⁶⁴ Internet content is becoming more complex, as web pages often have numerous images, CSS files, and external JavaScript. Loading all of these individual files takes time due to the overhead of separately requesting them.⁶⁵ Encryption further slows page load times. Data Saver addresses this issue by aggregating and compressing these requests so that users' devices do not have to individually request each feature. As with other proxy servers, ISPs can only see the request to the Data Saver server, but not the domain name, deep links, or user content.

Although we are not aware of statistics about Data Saver adoption, the scale is likely to become substantial because it is integrated with Google's operating system and web browser. Google encourages users to activate Data Saver as part of the Android operating system registration process, and offers Data Saver as an option for its Chrome web browser, including both Android and iOS devices.

⁵⁸ Jason Mander, "GWI Infographic: VPN Users," *GlobalWebIndex*, Oct. 24, 2014, (<http://www.globalwebindex.net/blog/vpn-infographic>).

⁵⁹ See generally, "The Missing Billion," *GlobalWebIndex*, 2015, (<https://app.globalwebindex.net/products/report/the-missing-billion>). The study analyzes how "passive web analytics has skewed our understanding of the global internet population - with trends including VPN usage, device sharing and mobile-only access causing hundreds of millions of Internet users, especially in fast-growth markets, to become invisible in traditional internet studies." See also Chris Smith, "Seriously Dark Traffic: 500 Mil. People Globally Hide Their IP Addresses" *Digiday*, Nov. 18, 2014, (<http://digiday.com/publishers/vpn-hide-ip-address-distort-analytics/>).

⁶⁰ See Chris Smith, "Seriously Dark Traffic: 500 Mil. People Globally Hide Their IP Addresses," *Digiday*, Nov. 18, 2014, (<http://digiday.com/publishers/vpn-hide-ip-address-distort-analytics/>).

⁶¹ "VPN Gate User Countries Realtime Ranking," *VPN Gate*, (<http://www.vpngate.net/en/region.aspx>) (last visited Dec. 28, 2015).

⁶² "Data Saver," *Chrome*, (<https://developer.chrome.com/multidevice/data-compression>).

⁶³ *Id.* The Data Saver proxy service now receives support from Google in place of the earlier SPDY protocol. "Hello HTTP/2, Goodbye SPDY," *Chrome*, Feb. 9, 2015, (http://blog.chromium.org/2015/02/hello-http2-goodbye-spdy-http-is_9.html).

⁶⁴ Lljitsch van Beijnum, "SPDY: Google Wants to Speed Up the Web by Ditching HTTP," *Ars Technica*, Nov. 12, 2009, (<http://arstechnica.com/business/2009/11/spdy-google-wants-to-speed-up-the-web-by-ditching-http/>).

⁶⁵ *Id.*

Other proxy servers may become prevalent as well. In 2013, Facebook purchased Onavo, which offers a similar compression/proxy service for mobile devices.⁶⁶ Onavo's products are positioned as ways to help consumers reduce their bandwidth usage, while also giving Facebook potential visibility into mobile web browsing and app usage. Facebook also recently announced that it is adding support for the Tor third-party proxy service on Android devices.⁶⁷ As such proxy servers are adopted, they block ISP visibility about the domain names, deep links, and content seen by users.

D. Conclusions on the Visibility of ISPs into Internet Usage

Far more than most people have realized, ISPs today and in the future face blockades to their technical ability to view users' Internet activity. This Chapter has documented three layers of technical blockades:

1. **Mobile and multiple devices.** As users become more mobile and use multiple devices, a single ISP has far less of a comprehensive view of a user's Internet activity. Roughly half of mobile traffic is offloaded to WiFi hotspots today, and that fraction will grow rapidly. The image of an ISP having "comprehensive" visibility due to its provision of home broadband service is outdated.
2. **HTTPS and other encryption.** Encryption pervasively blocks an ISP's visibility of users' deep links and content. The Chapter documents the spread of HTTPS, which is the new normal. For understanding the online advertising ecosystem, the lack of ISP visibility for encrypted communications matters for both quantitative reasons (most Internet traffic) and qualitative reasons (encryption for key content such as search, social networks, webmail, and increasingly online advertising itself).
3. **VPNs and other proxy services.** The spread of VPNs and other third-party proxy services adds another layer of encryption, beyond the use of HTTPS. Such services block ISP visibility of domain names as well as deep links and content. These proxy services can provide both greater security and faster speed for Internet activity, and adoption will likely rise sharply with support from players such as Facebook and Google.

Taken together, these technological blockades mean that ISPs do not and will not have "comprehensive" visibility into user Internet activity. In the new and future ecosystem, the ISPs that provide the last mile connection can see only narrowing subsets of the URLs and content that flow to users.

⁶⁶ Mike Isaac, "Facebook Acquires Israeli Mobile Analytics Startup Onavo," *All Things D*, Oct. 13, 2013, (<http://allthingsd.com/20131013/facebook-acquires-israeli-mobile-data-management-startup-onavo/>).

⁶⁷ "Adding Tor Support on Android," *Facebook*, Jan. 19, 2016, (https://www.facebook.com/notes/facebook-over-tor/adding-tor-support-on-android/814612545312134?_rdr=p).

Appendix 1: Encryption for Top 50 Websites

This Appendix reports on encryption for the top 50 websites listed by Alexa.com in February 2016.

1. All of the top 10 sites either use HTTPS by default or shift to HTTPS when the user logs-in.
2. 42 of the top 50 sites either use HTTPS by default or shift to HTTPS when the user logs-in.
3. 24 of the top 50 sites use HTTPS by default, even without user log-in.

In this Appendix, green means HTTPS by default. Yellow means encryption when the user logs-in. Red means does not support encryption by default.

As shown in Diagram 1-A, when a session is encrypted with HTTPS, the ISP can see the host name, such as <https://www.exampleshoes.com>, but the encryption blocks the ISP from seeing the full URL or the content.

1. Google. HTTPS by default: <https://www.google.com/>
2. Facebook. HTTPS by default: <https://www.facebook.com/>
3. YouTube. HTTPS by default: <https://www.youtube.com/>
4. Amazon.com. HTTPS upon log-in: <http://www.amazon.com/>
5. Yahoo.com. HTTPS by default: <https://www.yahoo.com/>
6. Wikipedia.org. HTTPS by default: <https://www.wikipedia.org/>
7. Ebay.com. HTTPS upon log-in: <http://www.ebay.com/>
8. Twitter.com. HTTPS by default: <https://twitter.com/>
9. Reddit.com. HTTPS by default: <https://www.reddit.com/>
10. Netflix.com. HTTPS by default: <https://www.netflix.com/>
11. Live.com. HTTPS by default: <https://login.live.com/>
12. LinkedIn.com. HTTPS by default: <https://www.linkedin.com/>
13. Pinterest.com. HTTPS by default: <https://www.pinterest.com/>
14. Craigslist.org. HTTPS upon log-in: <http://washingtondc.craigslist.org/>
15. Go.com. Does not support encryption by default: <http://go.com/>
16. Imgur.com. HTTPS upon log-in: <http://imgur.com/>
17. Chase.com. HTTPS by default: <https://www.chase.com/>
18. Paypal.com. HTTPS by default: <https://www.paypal.com/home>
19. Tumblr.com. HTTPS by default: <https://www.tumblr.com/>
20. Cnn.com. Does not support encryption by default: <http://www.cnn.com/>
21. Instagram.com. HTTPS by default: <https://www.instagram.com/>

22. Bing.com. HTTPS upon log-in: <http://www.bing.com/>
23. Imdb.com. Does not support encryption by default: <http://www.imdb.com/>
24. Espn.go.com. Does not support encryption by default: <http://espn.go.com/>
25. Nytimes.com. Does not support encryption by default: <http://www.nytimes.com/>
26. Bankofamerica.com. HTTPS by default: <https://www.bankofamerica.com/>
27. Weather.com. HTTPS by default: <https://weather.com/>
28. Msn.com. HTTPS upon log-in: <http://www.msn.com/>
29. T.co by. Site becomes encrypted as soon as you follow through to Twitter: <http://t.co/>
30. Blogspot.com. HTTPS upon log-in: <http://blogspot.com/>
31. Wellsfargo.com. HTTPS by default: <https://www.wellsfargo.com/>
32. Yelp.com. HTTPS upon log-in: <http://www.yelp.com/>
33. Office.com. HTTPS by default: <https://www.office.com/>
34. Zillow.com. HTTPS upon log-in: <http://www.zillow.com/>
35. Walmart.com. HTTPS upon log-in: <http://www.walmart.com/>
36. Apple.com. HTTPS upon log-in: <http://www.apple.com/>
37. Huffingtonpost.com. Does not support encryption by default: <http://www.huffingtonpost.com/>
38. Intuit.com. HTTPS upon log-in: <http://www.intuit.com/>
39. Wordpress.com. HTTPS by default: <https://wordpress.com/>
40. Etsy.com. HTTPS by default: <https://www.etsy.com/>
41. Buzzfeed.com. Does not support encryption by default: <http://www.buzzfeed.com/>
42. Microsoftonline.com. HTTPS by default: <https://login.microsoftonline.com/>
43. Aol.com. HTTPS upon log-in: <http://www.aol.com/>
44. Microsoft.com. HTTPS by default: <https://www.microsoft.com/en-us/>
45. Target.com. HTTPS upon log-in: <http://www.target.com/>
46. Comcast.net. HTTPS upon log-in: <http://my.xfinity.com/?cid=cust>
47. Washingtonpost.com. HTTPS by default: <https://www.washingtonpost.com/regional/>
48. Foxnews.com. Does not support encryption by default: <http://www.foxnews.com/>
49. Bestbuy.com. HTTPS upon log-in: <http://www.bestbuy.com/>
50. Wikia.com. HTTPS upon log-in: <http://www.wikia.com/fandom>

Appendix 2: The Growing Prevalence of HTTPS as a Fraction of Internet Traffic

This Appendix supplies data from the Center for Applied Internet Data Analysis (CAIDA) about the growing prevalence of HTTPS. CAIDA “is a collaborative undertaking among organizations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure.” (<http://www.caida.org/home/about/>).

The information for these charts comes from a CAIDA probe in Chicago on an Internet backbone link that goes between Chicago and Seattle. (<http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA>). In these diagrams, teal represents the HTTPS traffic, and lime green the traffic that uses HTTP.

The first chart shows the large increase in share of HTTPS in the last two years:

- In early 2014, the teal for HTTPS was a small fraction of total traffic. For April 2014, HTTPS was 13.3 percent.¹
- HTTPS traffic increased considerably by the beginning of 2015.
- By early 2016, HTTPS traffic was greater than HTTP traffic.

The second chart focuses on one week in February, 2016, when we ran the most current figures:

- By this time, HTTPS traffic was 38 percent larger than HTTP traffic.²
- By this time, HTTPS traffic accounted for essentially half of total traffic (48.6 percent).³

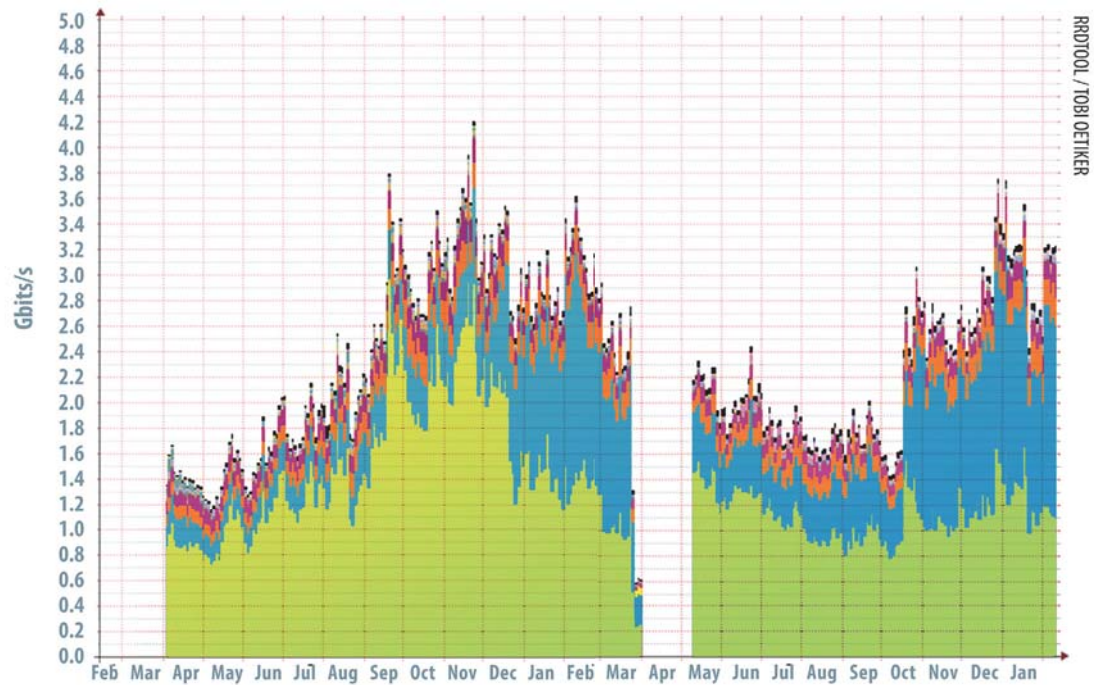
As shown in Diagram 1-A, when a session is encrypted with HTTPS, the ISP can see the host name, such as <https://www.exampleshoes.com>, but the encryption blocks the ISP from seeing the full URL or the content.

¹ Diagram 3 is a screen shot showing the share of HTTPS traffic in April, 2014.

² HTTPS was 48.6 percent and HTTP was 35.1 percent. $48.6/35.1=1.38$, or 38% higher for HTTPS.

³ Diagram 4 is a screen shot showing the share of HTTPS traffic from February 3 to February 10, 2016.

APPLICATION BITS/S - 2 YEARS



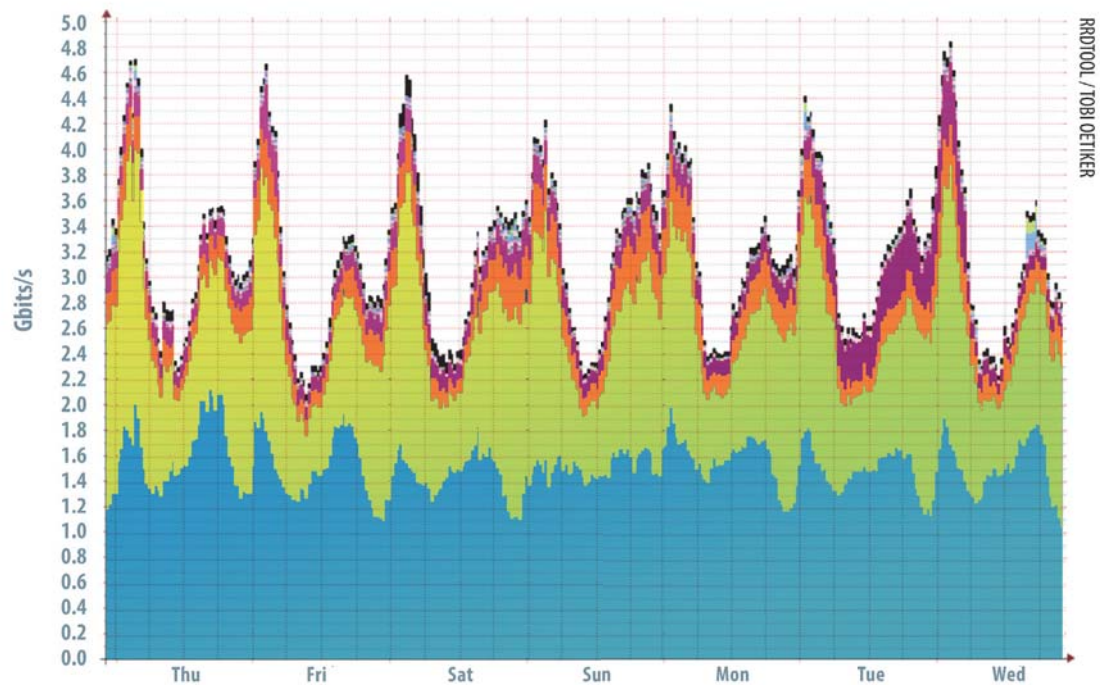
FEBRUARY 10, 2014 - FEBRUARY 10, 2016 UTC

Application	Min	Avg	Max
HTTP	231.80M	1.34G	3.11G
HTTPS	135.66M	609.09M	1.74G
UNKNOWN_UDP	39.29M	156.90M	300.38M
UNKNOWN_TCP	22.27M	120.44M	346.39M
RTMP	5.83M	16.57M	38.26M
ABACAST	82.15k	12.71M	97.93M
ICHAT	4.50M	11.95M	38.90M
NTP	85.83k	9.50M	220.68M
NOPTS_UDP	705.36k	6.12M	67.61M
SQUID	54.51k	6.03M	95.29M
QUAKE	1.16M	5.72M	11.69M
DNS	884.53k	5.16M	43.11M
SSH	178.65k	3.66M	24.43M
MS_LIVE	163.82k	3.15M	28.68M
other	0.00	33.75M	115.25M

generated 2016-01-11 21:03UTC
created with CAIDA's CoralReef (c) 2012 UC Regents

Diagram 1 – the 2 year series

APPLICATION BITS/S - 1 WEEK



FEBRUARY 03, 2016 - FEBRUARY 10, 2016 UTC

Application	Min	Avg	Max
HTTPS	1.04G	1.53G	2.12G
HTTP	480.93M	1.14G	2.31G
UNKNOWN_UDP	89.05M	232.12M	454.18M
UNKNOWN_TCP	80.40M	180.76M	516.36M
ICHTAT	628.40k	17.17M	47.39M
RTMP	2.93M	13.60M	32.11M
SSH	2.76M	13.19M	55.05M
QUAKE	1.33M	10.17M	54.71M
MS_LIVE	855.50k	9.09M	28.01M
NOPTS_UDP	272.94k	8.14M	193.55M
DNS	1.80M	6.65M	117.27M
ABACAST	739.76k	1.75M	10.07M
SQUID	139.32k	1.74M	12.02M
NTP	45.40k	1.66M	36.47M
other	23.32M	48.84M	167.74M

generated 2016-01-11 21:03UTC
created with CAIDA's CoralReef (c) 2012 UC Regents

Diagram 2 – Feb 2016

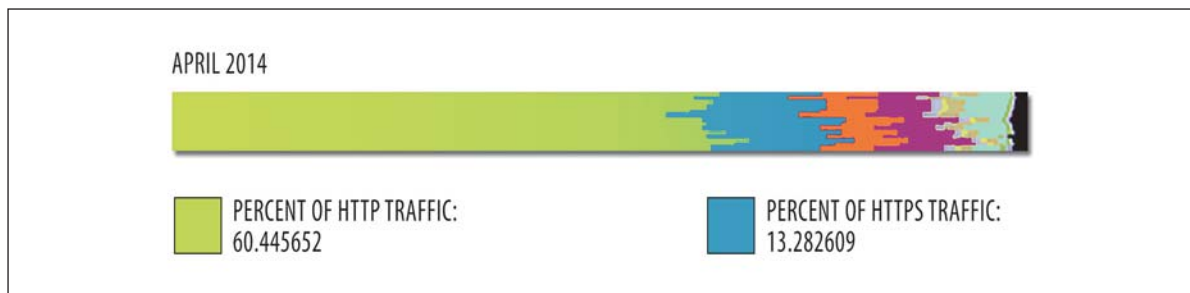


Diagram 3 – screen shot for April 2014 statistics.

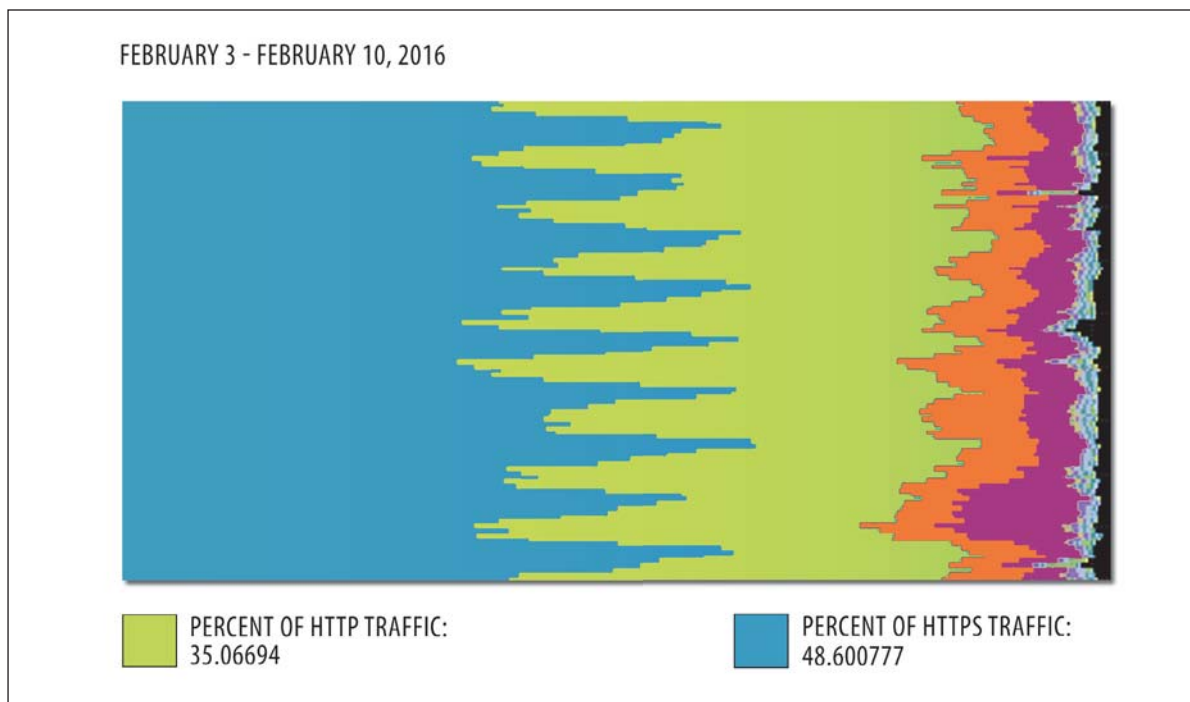


Diagram 4 - screen shot for week in Feb 2016 statistics.



Social Networks

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 2: Social Networks

Social networks¹ are able to gain unique insights into the online activity of their users, and sometimes even non-users, because they have the means to collect commercially valuable information through the use of various strategies, services, and technologies. The amount of commercially valuable data that social networks are able to access and collect often exceeds that of Internet Service Providers (“ISPs”)² due to the important role these websites play in consumers’ lives. The average online American spends two hours daily on social media.³ Further, on average, Americans check their Facebook, Twitter, and other social media accounts seventeen (17) times per day.⁴ Because users frequently interact with social networks, social networks have unique access to users’ data and are therefore able to gain deep insights into users’ online activity.⁵ Social networks may use that collected data for numerous purposes, such as targeted advertising. According to the Interactive Advertising Bureau (“IAB”), social media advertising revenue increased at a 55 percent annual rate from 2012 to 2014.⁶ Facebook alone had mobile advertising revenues of over \$7.39 billion in 2014⁷ and total advertising revenues of over \$11.5 billion.⁸

This Chapter first discusses various ways social networks are able to collect commercially valuable user data. Because users are often logged-in for extended periods, social networks can connect disparate actions to a particular user account including taking advantage of the rise of social plug-ins on other websites, thus permitting higher visibility into a user’s Internet activity history. Second, the Chapter explains the importance of what economists call “network effects” for the value of information that accrues to successful social networks. Third, the Chapter discusses the ways social networks can use this data for advertising purposes, specifically with respect to targeted advertising programs such as Facebook’s Atlas ad-serving platform. Lastly, the Chapter concludes by comparing the increasing and rich content available to social networks with the lack of similar content available to ISPs.

This comparison of social networks and ISPs undermines the widely-held, but mistaken, view that ISPs have comprehensive and unique knowledge about user online activity because they operate the last mile of the network. For the important realm of social media, companies that contain ISPs have not been prominent

¹ For purposes of this Working Paper, a “Social Network” will be defined as an online service, platform, or website that focuses on facilitating the building of social networks or social connections among people who, for example, share interests, activities, backgrounds, or real-life connections. Social networks allow users to share ideas, activities, events, and interests with their individual networks. Examples of social networks include Facebook, Google+, LinkedIn, Pinterest, and Twitter. “Social Networking,” *Mashable*, (<http://mashable.com/category/social-networking/>).

² As used throughout this Working Paper, “Internet Service Provider” (“ISP”) is defined as the company that connects an individual user to the Internet.

³ “Social Networking Eats Up 3+ Hours Per Day for the Average American User,” *Marketing Charts*, Jan. 9, 2013, (<http://www.marketingcharts.com/wp/interactive/social-networking-eats-up-3-hours-per-day-for-the-average-american-user-26049/>).

⁴ Lulu Chang, “Americans Spend An Alarming Amount of Time Checking Social Media on their Phones,” *Digital Trends*, June 13, 2015, (<http://www.digitaltrends.com/mobile/informate-report-social-media-smartphone-use/>).

⁵ Cooper Smith, “SOCIAL BIG DATA: Each Social Network Is Using a Very Different Data Lens to Understand and Target Users,” *Business Insider*, Mar. 12, 2014, (<http://www.businessinsider.com/social-big-data-the-type-of-data-collected-by-social-networks-3-2014-3>).

⁶ “IAB internet advertising revenue report: 2014 full year results,” *IAB*, April 2015, (http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_FY_2014.pdf).

⁷ Steven Max Patterson, “Google, Facebook Combined for 50% of Mobile Ad Revenues in 2014,” *Network World*, Feb. 6, 2015, (<http://www.networkworld.com/article/2881132/wireless/google-facebook-combined-for-50-of-mobile-ad-revenues-in-2014.html>).

⁸ “Facebook Reports First Quarter 2014 Results,” *Facebook*, April 23, 2014, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=842071>); “Facebook Reports Second Quarter 2014 Results,” *Facebook*, July, 23, 2014, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=861599>); “Facebook Reports Third Quarter 2014 Results,” *Facebook*, Oct. 28, 2014 (<http://investor.fb.com/releasedetail.cfm?ReleaseID=878726>); “Facebook Reports Fourth Quarter and Full Year 2014 Results,” *Facebook*, Jan. 28, 2015, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=893395>).

players. Rather, non-ISPs, such as the social networks discussed in this Chapter, are themselves able to gain much insight into online user activity, often greater than that of an ISP. Social networks have effective means to gather commercially valuable information about their users, and sometimes non-users. In the terms used in this Working Paper, social networks are one “context” in which online information is gathered about users. As will be discussed in Chapter 8, information gathered in one context, such as social networks, is often combined effectively with information from other contexts, resulting in what we call “cross-context tracking.” Further, as will be discussed in Chapter 9, information can also be combined in cross-device tracking with information from other devices, especially when users are logged-in (as they are for many social networks), further enhancing the insights about users available to such non-ISPs.

A. How Social Networks Gather Commercially Valuable Information

1. User-generated content

A recent study indicates that 76 percent of all U.S. Internet users use at least one social networking site, while 90 percent of young adults (ages 18-29) use social media.⁹ Social networks have access to content that users expressly provide to them (i.e., user-generated content). For example, some social networks require that users create a profile to use their services. By creating a profile, users generally provide important personal data such as name, city of birth, birth date, city of residence, relationship status, email address, education, current employment, and previous employment. In order to use some of the premium functions of the particular social network, users may provide even more specific information such as their race, ethnicity, religious views, relationships, family members, favorite musicians, favorite movies, favorite foods, places the user has visited, and events the user has attended or will attend. Further, many social networks request a profile picture, and they allow users to upload additional photos and videos of themselves and friends.¹⁰

All of this data that users expressly provide to social networks permits social networks to construct a profile of who the user is and her likes and dislikes.¹¹ Further, the social network can in many instances connect all of the data to the account of a named user, rather than a cookie or other device identifier.

The data permits social networks to achieve one of their primary goals of facilitating social connections. Specifically, for members of the site with friends and followers, the social network knows about users’ most intricate relationships and can therefore connect users with other people. Individuals are experts about themselves; as individuals use a social network and connect with other people, they have many reasons to provide accurate and granular data that may not be available to advertisers in other contexts.

Another category of user-provided content is payment information, such as credit or debit cards. Some social networks collect this information when a user joins the service or buys applications or other services through the social network. This payment data contains authenticated information, such as address and phone number, supplying more information and more accurate information than may otherwise be true for user-generated content. Such payment information is also available to others in the online ecosystem, including e-commerce companies and others who receive payment directly from the individual.

⁹ Andrew Perrin, “Social Media Usage: 2005-2015 Survey,” *Pew Research Center*, Oct. 8, 2015, (<http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/>).

¹⁰ In 2013, Facebook users alone uploaded more than 350 million new photos each day and uploaded more than 250 billion photos in total. See Cooper Smith, “Facebook Users Are Uploading 350 Million New Photos Each Day,” *Business Insider*, Sep. 18, 2013, (<http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9>).

¹¹ This data can be associated with the company’s own identifier, such as a Facebook or Twitter ID, which is then used for advertising purposes but is not explicitly personally-identifiable information.

2. Metadata

In addition to user-generated content, social networks gain metadata about what the users provide. Consider a status post to a social network. The user provides user-generated content, such as: "I am at a concert seeing my favorite band." A dozen friends make comments and 50 friends like the post. The user-generated content is what the individual posts, the dozen comments, and the 50 likes. Metadata includes all of the other information a social network may generate from this post, such as: the date and time of the user-generated content, the location of users when they interact with the site, analysis of the social graphs of all the individuals (who connects with whom), and inferences about the popularity of the band or other advertising-relevant information.

Photos illustrate the information available to a social network that go beyond the raw data provided by the user. Depending on who or what is depicted in the photo, the social network may be able to gain insight about the identity of the user's friends and family, the user's favorite hobbies, and where the user is located. One way social networks can gain this insight is through facial verification technology, which recognizes when two images show the same face.¹² For example, Facebook created a facial recognition technology called DeepFace¹³ that allows Facebook to identify a person in the photograph (based on the other tagged photos on Facebook) and tag the user automatically.¹⁴ DeepFace takes the face in a photograph and corrects the angle of the face so the person in the picture faces forward, using a 3-D model of an "average" forward-looking face.¹⁵ Once that step is completed, deep learning occurs through a simulated neural network, working out a numerical description of the reoriented face.¹⁶ This deep network has more than 120 million parameters using several locally-connected layers.¹⁷ In a research paper, Facebook representatives stated that Facebook trained this technology "on the largest facial dataset to date, an identity labeled dataset of four million facial images belonging to more than 4,000 identities."¹⁸ Facebook's DeepFace has an accuracy of 97.35 percent, "reducing the error of the current state of the art by more than 27 percent, closely approaching human-level performance."¹⁹ This technology is just one example of how social networks go beyond user-generated content (what users explicitly provide). Analysis of the user-generated content and metadata about the services used by subscribers can lead to numerous inferences about the users, often with advertising-relevant targeting that is not available through other channels.

Apps accessed through the social network platform (whether free or purchased) can add to the data available to a social network. The social network may see metadata about app usage, such as which apps a user chooses, when the user is logged-into the app, as well as the user's location (even when the user is not logged-into the app).

¹² Tom Simonite, "Facebook Creates Software That Matches Faces Almost as Well as You Do," *MIT Technology Review*, March 17, 2014, (<http://www.technologyreview.com/news/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do/>).

¹³ Marc'Aurelio Ranzato, Yaniv Taigman, Ming Yang, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," Facebook AI Research and Tel Aviv University, (https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf); Darrell Etherington, "Facebook's DeepFace Project Nears Human Accuracy in Identifying Faces," *TechCrunch*, March 18, 2014, (<http://techcrunch.com/2014/03/18/faceook-deepface-facial-recognition/>).

¹⁴ Facebook began rolling out this technology to its users in early 2015, with the exception of EU users.

¹⁵ Tom Simonite, "Facebook Creates Software That Matches Faces Almost as Well as You Do," *MIT Technology Review*, March 17, 2014, (<http://www.technologyreview.com/news/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do/>).

¹⁶ Amit Chowdhry, "Facebook's DeepFace Software Can Match Faces With 97.25% Accuracy," *Forbes*, March 18, 2014, (<http://www.forbes.com/sites/amitchowdhry/2014/03/18/facebooks-deepface-software-can-match-faces-with-97-25-accuracy/>); citing Tom Simonite, "Facebook Creates Software That Matches Faces Almost as Well as You Do," *MIT Technology Review*, March 17, 2014, (<http://www.technologyreview.com/news/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do/>).

¹⁷ Marc'Aurelio Ranzato, Yaniv Taigman, Ming Yang, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," Facebook AI Research and Tel Aviv University, p. 1 (https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf).

¹⁸ *Id.*

¹⁹ *Id.*

Social plug-ins have become a highly visible aspect of web surfing. Users often see a half dozen social plug-ins, or more, when visiting web pages. As described in more detail in Chapter 8 on cross-context tracking, social plug-ins can provide a great deal of information about a user's activities on websites and apps outside the social network itself. The most widely-used example of social plug-ins is the Facebook "Like" button, which appears today in over 32 percent of the top 10,000 websites; these websites span many categories, including health and government.²⁰ The combination of user-generated content and metadata can provide more comprehensive visibility about Internet use than is available for ISPs, which are constrained by technology and market factors such as encryption, as discussed in Chapter 1.

3. Logged-in users

Enhanced knowledge about a user often occurs when a user is logged-in to a social network or other service. The activity of a logged-in user can be reliably attributed to the user's account, contributing to advertising-relevant insights about that user. This ability to associate account activity with a logged-in user is much different than the traditional third-party cookie tracking, which is further explained in Chapter 6 on interest-based advertising. With third-party cookie tracking, the activity the cookie receives is linked to a single identifier (i.e., account), but often not to a named individual or an offline identity.

Logged-in activity can range in usefulness from log in by an unnamed user to high-quality user authentication. At a minimum, logged-in status provides reliable linking of the range of activity in the single account. Many users of social networks use their real names; using real names even may be the social network's policy.²¹ Beyond that, there are significant incentives for users to provide accurate identification so their friends and business contacts can accurately communicate with them. In addition, as mentioned above, social networks often get their users' payment card information, providing highly authenticated name, address, and telephone information. When a real name is associated with an account, the social network can purchase additional information about the user from data marketplaces, which is a useful practice for targeted advertising.

B. Network Effects

What economists call "network effects" help many social networks to grow and collect users' data. A service or product "displays positive network effects when more usage of the product by any user increases the product's value for other users."²² Specifically, when more users use the social network and its services, the social network's value increases. Facebook is an example of one social network that has benefited from network effects. The fact that so many individuals already use Facebook is an important reason new users choose to join Facebook. This network effect is one reason Facebook has such a large consumer base and therefore can collect personal data.²³

²⁰ Gunes Acar, Brendan Van Alsenoy, Claudia Diaz, Frank Piessens, Bart Preneel, "Facebook Tracking through Social Plug-ins," Belgian Privacy Commission, Ver. 1.1, June 24, 2015, p. 2, (https://securehomes.esat.kuleuven.be/~gacar/fb_tracking/fb_plug-ins.pdf).

²¹ "What Names Are Allowed on Facebook," *Facebook*, (<https://www.facebook.com/help/112146705538576>). "Create or Change Your Google+ Profile Name," Google+ Help, (<https://support.google.com/plus/answer/1228271?hl=en>).

²² Eric Jorgenson, "The Power of Network Effects: Why They Make Such Valuable Companies, and How to Harness Them," *Evergreen*, June 22, 2015, (<https://medium.com/evergreen-business-weekly/the-power-of-network-effects-why-they-make-such-valuable-companies-and-how-to-harness-them-5d3fbc3659f8>).

²³ As of April 2015, Facebook had more than 1.44 billion monthly active users, with 1.25 billion also being mobile users. On a daily basis, more than 936 million people use Facebook, and 798 million people are mobile daily active users. Further, 65 percent of Facebook's members use the service daily, and 64 percent of its mobile members use it daily. Emil Protalinski, "Facebook Passes 1.44B Monthly Active Users and 1.25B Mobile Users; 65% Are Now Daily Users," *Venture Beat*, April 22, 2015, (<http://venturebeat.com/2015/04/22/facebook-passes-1-44b-monthly-active-users-1-25b-mobile-users-and-936-million-daily-users/>).

Network effects contribute to the incentives for websites to display social plug-ins, such as those of Facebook, Pinterest, and Twitter. When users click on the social plug-in button, the website often gets what amounts to “free advertising” within the social network because other social network users see the photo, article, or website links. The social network, in turn, benefits from the increased user engagement these links create. Websites have particular interest in getting that “free advertising” from the most popular social networks, reinforcing their incentive to highlight the most popular plug-ins. Beyond this, data associated with the plug-ins may provide additional reasons for websites to participate. While search engines often do not give much detail on their ranking algorithms, Microsoft’s Bing has publicly stated that it uses “Like” data for its search algorithms.²⁴ In short, websites and social networks have multiple reasons to facilitate the linking, with the strongest effects coming from the most popular networks.

C. How This Data Helps Advertisers

1. Targeted advertising based on detailed profiles

Once social networks have collected the data discussed above, they are able to use it for a number of different purposes. Notably, social networks are able to use the data for targeted advertisements, that is, offering ads specific to the individual or the individual’s interests and demographics. Such targeted advertising can be particularly effective because the social network has already gained much insight into the user and the user’s online activities through the collected data.

After social networks have collected and analyzed the data, it can be put to use either for advertisements visible on the network itself, or for advertisements that users see elsewhere. The discussion here highlights Facebook, which has the most detailed public reports about its practices. Facebook has the largest advertising operation for ads that appear within the network itself, with global mobile advertising revenues of over \$7.39 billion in 2014²⁵ and total advertising revenues of over \$11.5 billion.²⁶ Companies interested in advertising on Facebook can provide detailed parameters for the target audience; Facebook, relying on the detailed information about each user as described above, can deliver the ads to that audience. As one author has explained, advertisers can narrow down the parameters, for example, to “[s]omeone engaged to be married, who lives in New York, between the ages of 20-30, who likes swimming, and who drives a BMW.”²⁷ As discussed in more detail in Chapter 6 on interest-based advertising, social networks can also sell or share data that enables advertising on sites other than the social networks themselves. In general, the more detailed information a social network (or other actor) can gain, the more relevant the individual’s advertisements will be.

Thus far, this Working Paper has discussed “online” data, that is, data collected through the Internet. Advertisers have an incentive to connect online data with data collected offline, such as in-person sales, catalogues, and individuals’ public records. Online data holders can benefit, for instance, by gaining access to offline purchase

²⁴ Adam Ostrow, “Facebook and Bing’s Plan to Make Search Social,” *Mashable*, Oct. 30, 2010, (http://mashable.com/2010/10/13/facebook-bing-social-search/#9hL_zEWMFZqL).

²⁵ Steven Max Patterson, “Google, Facebook Combined for 50% of Mobile Ad Revenues in 2014,” *Network World*, Feb. 6, 2015, (<http://www.networkworld.com/article/2881132/wireless/google-facebook-combined-for-50-of-mobile-ad-revenues-in-2014.html>).

²⁶ “Facebook Reports First Quarter 2014 Results,” *Facebook*, April 23, 2014, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=842071>); “Facebook Reports Second Quarter 2014 Results,” *Facebook*, July 23, 2014, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=861599>); “Facebook Reports Third Quarter 2014 Results,” *Facebook*, Oct. 28, 2014, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=878726>); “Facebook Reports Fourth Quarter and Full Year 2014 Results,” *Facebook*, Jan. 28, 2015, (<http://investor.fb.com/releasedetail.cfm?ReleaseID=893395>).

²⁷ Thorin Klosowski, “How Facebook Uses Your Data to Target Ads, Even Offline,” *Lifehacker*, April 11, 2013, (<http://lifehacker.com/5994380/how-facebook-uses-your-data-to-target-ads-even-offline>).

records or detailed demographics through public records services. Offline data holders can benefit, for instance, from user-generated content provided to social networks and from metadata about a user's activities. Both sides can benefit from attribution – the ability to link online and offline behavior – which can provide evidence that a particular sale was attributable to a particular ad campaign. Knowledge about attribution reduces uncertainty and increases an advertiser's willingness to purchase effective ads.

An article by Thoren Klosowski highlighted the partnerships Facebook has forged with data collection companies, often from offline sources including Acxiom, BlueKai, Datalogix, and Epsilon.²⁸ These companies collect information through store loyalty cards, mailing lists, public records information (including home or car ownership), and cookies.²⁹ Drawing on these outside data sources, Facebook can create more detailed profiles of its users, which can lead to more specific targeted advertisements.

A more recent development is Facebook's Custom Audience service, which allows an advertiser to upload an email list and compare that data with who is on Facebook.³⁰ If the advertiser is able to pair the email address it has with one registered on Facebook, the advertiser can serve a targeted advertisement to that user.³¹

In short, social networks' abundance of data has become an important source for targeted advertising.

2. From social network to advertising network

Social networks' importance to online advertising has continued to expand, creating a convergence of social networking with online advertising networks.³² In 2014, Facebook announced Facebook Atlas as a separate product from its other advertising business that still relies on the same data set.³³ Atlas is an ad server platform for targeting, serving, and measuring ads in a digital space; in contrast, Facebook's internal ad products only serve and measure ads on the social network. Atlas is capable of measuring activity on ads served by a different platform as long as they contain an Atlas tag or pixel.

Atlas is positioned to be particularly effective as a cross-device measurement tool,³⁴ as it relies on Facebook's data-derived identifiers in conjunction with cookies or other traditional tracking technologies.³⁵ Atlas also is positioned to be particularly effective at serving and tracking mobile device advertising; Atlas claims it can access 50 percent of a mobile user's Internet activity history, and it can potentially measure three of every five minutes of a mobile user's Internet activity history.³⁶

²⁸ *Id.*

²⁹ Steve Kroft, "The Data Brokers: Selling Your Personal Information," *CBS News*, March 9, 2014, (<http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/>).

³⁰ "Target Facebook Ads to People on Your Contact List," *Facebook for Business*, (<https://www.facebook.com/business/a/custom-audiences>).

³¹ *Id.*

³² Gunes Acar, Brendan Van Alsenoy, Claudia Diaz, Frank Piessens, Bart Preneel, "Facebook Tracking through Social Plug-ins," Belgian Privacy Commission, Ver. 1.1, June 24, 2015, (https://securehomes.esat.kuleuven.be/~gacar/fb_tracking/fb_plug-ins.pdf).

³³ Justin Lafferty, "Facebook Announces the New Atlas: A Cross-Platform Ad Network," *AdWeek*, Sep. 29, 2014, (<http://www.adweek.com/socialtimes/facebook-announces-the-new-atlas-a-cross-platform-ad-network/301054>).

³⁴ See Chapter 9 Cross-Device Tracking.

³⁵ Will Oremus, "On Facebook's New Ad Platform, Your Data Will Follow Everywhere," *Slate*, Sep. 29, 2014, (http://www.slate.com/blogs/future_tense/2014/09/29/facebook_atlas_ad_platform_your_data_will_follow_you_across_web_apps_devices.html).

³⁶ *Id.* at 3.

Atlas connects offline purchases and conversions to digital ad impressions,³⁷ which improves the cross-context tracking we discuss in Chapter 8. Atlas can leverage relationships with offline retailers to link offline customer records with the customer's Facebook account. If Facebook has a valid email address, phone number, or other identifier shared with a retailer's customer records, the two data sets can be linked.³⁸ Atlas states that it does this anonymously, without passing personal data to the advertisers and with no passing of advertiser data back to the main Facebook service.³⁹ An advertising company sets a window of time during which it is willing to attribute an offline sale to an online advertising impression, and Atlas returns information about what ads—if any—a purchaser's Facebook accounts were served and when. Atlas claims this type of data tracking will enable attribution and allow marketers to better calculate the return on investment ("ROI") for advertising campaigns, and it will see how "sequencing, creative assets, placement, and timing affect conversion behavior."⁴⁰

D. Conclusion

Despite the widely-held view that ISPs have comprehensive and unique access into a user's online activity, Chapter 1 showed that ISPs actually have far from comprehensive visibility into a user's Internet activity history. This Chapter showed, by contrast, the insights that social networks gain as they receive a large amount of data about their users, and sometimes even non-users. The commercially valuable data these social networks receive is enhanced by the fact that network effects encourage more people to join social networks over time. In addition, data gleaned from social networks can be combined in various ways with other sources of user data, often across different lines of business or across devices, as discussed further in Chapters 8 and 9.

The ability of ISPs to see social network-related data is very limited. None of the traditional ISPs are among the largest social networks. In addition, Facebook and other social networks have shifted to HTTPS by default,⁴¹ so ISPs are blocked from seeing social network-related data even if they try to inspect it. The insights from social networks are uniquely available to other players in the ecosystem, but not to ISPs.

³⁷ Marcelo Ballvé and Emily Adler, "The Atlas Explainer: Where Facebook's Atlas Ad Server Fits in the Digital-Ad Ecosystem, and How It works," *BI Intelligence*, April 10, 2015.

³⁸ Cade Metz, "How Facebook Knows When Its Ads Influence Your Offline Purchases," *Wired*, Dec. 11, 2014, (<http://www.wired.com/2014/12/facebook-knows-ads-influence-offline-purchases/>).

³⁹ Note that Facebook accomplishes this by hashing the email addresses and comparing the hashed values to find matches. See Will Oremus, "On Facebook's New Ad Platform, Your Data Will Follow Everywhere," *Slate*, Sep. 29, 2014, (http://www.slate.com/blogs/future_tense/2014/09/29/facebook_atlas_ad_platform_your_data_will_follow_you_across_web_apps_devices.html).

⁴⁰ Marcelo Ballvé and Emily Adler, "The Atlas Explainer: Where Facebook's Atlas Ad Server Fits in the Digital-Ad Ecosystem, and How It works," *BI Intelligence*, April 10, 2015.

⁴¹ Jennifer Van Grove, "Facebook Migrates Everyone to HTTPS Connection," *CNET*, July 31, 2013, (<http://www.cnet.com/news/facebook-migrates-everyone-to-https-connection/>).

Search Engines

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 3: Search Engines

This Chapter examines the largest single source of online advertising revenue: search engines. A search engine is a web-based tool that enables users to locate information on the World Wide Web (“web”).¹ Search engines use automated software applications (i.e., robots, bots, or spiders) that travel along the web.² These applications gather information that is used to create a searchable index of the web. The largest search engines for U.S. users are Bing (Microsoft), Google Search (Google) and Yahoo! Search (Yahoo), none of which are in the same companies as traditional Internet Service Providers (“ISPs”).³

The importance of search to the online advertising ecosystem is shown in the annual IAB Internet Advertising Revenue Reports. In 2014, non-mobile search revenues were \$19 billion.⁴ Non-mobile search was by far the largest segment of online revenue each year from 2006 to 2014, ranging from 38 to 47 percent of total online revenue.⁵ A study by MarketingProfs found that 91 percent of online adults use search engines to find information on the web, and 54 percent of search engine users utilize a search engine at least once a day to find information online.⁶ This intensive use of search engines enables their providers to collect highly specific and personalized data.

This Chapter first describes the key data flows about search engine users. By highlighting the ability of a search engine to see both the URLs and content a user selects, it next explores a comparison of search to the theoretical capability of ISPs to see both detailed URLs and content. The discussion then highlights key attributes of search that are important to the online advertising ecosystem: insight into users’ intent; targeted advertising; efficient auction ecosystem; and links with other applications, including digital assistants such as Apple’s Siri, Google’s Google Now, and Microsoft’s Cortana. The Chapter concludes by comparing the increasing and rich data available to search engines with the lack of similar data available to companies in their roles as ISPs.

As with other Chapters in this Working Paper (such as the previous Chapter on social networks), this comparison of search engines and ISPs undermines the widely-held, but mistaken view that ISPs have comprehensive and unique knowledge about users’ online activity because they operate the last mile of the network. For the important search engine realm, companies that contain ISPs have not been leading players and providing the last-mile connection has not created important advantages. Rather, non-ISP search engines are able to gain much unique insight into online user activity, often greater than that of an ISP.⁷ As discussed in the wrap-up Chapters on cross-context tracking and cross-device tracking, the information-gathering advantages of search engines also are often combined with sources of user data from other settings, further showing that unique insights about users are more likely to come from non-ISPs than from ISPs.

¹ “What is a Search Engine?” *DesignHammer*, (<https://designhammer.com/services/seo-guide/search-engines>).

² *Id.*

³ “Top 15 Most Popular Search Engines,” *The eBusiness MBA Guide*, Dec. 2015, (<http://www.ebizmba.com/articles/search-engines>). Yahoo! Search is largely powered by Bing.

⁴ “IAB internet advertising revenue report: 2014 full year results,” *IAB*, April 2015, (http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_FY_2014.pdf).

⁵ *Id.* The IAB does not break out statistics for mobile search, which accounted for 28 percent of total online advertising revenues in Q4 2014.

⁶ Verónica Maria Jarski, “How People Search Online,” *MarketingProfs*, Sep. 21, 2013, (<http://www.marketingprofs.com/chirp/2013/11692/how-people-search-online-infographic>).

⁷ The largest ISP will only have access to a fraction of its users’ Internet traffic, as explained in this Working Paper. In contrast, Google processes over 40,000 search queries every second on average, which equates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide. “Google Search Statistics,” *Internet Live Stats*, (<http://www.internetlivestats.com/google-search-statistics/>) (last visited Jan. 4, 2016).

A. Search Engine Data Flows

1. Search queries and search results

Search engine providers gain detailed knowledge about users' online activity through their search services. Generally, search engine providers collect two different kinds of data: (1) search queries that users type into their search bars and (2) search results that users click on and visit.⁸ If there is a regular HTTP connection, ISPs can have the technical ability to collect this information, as they are the entities routing the Internet traffic. For example, if there is an HTTP connection and a user types "Example Shoes" in the search engine, the user's ISP and the search engine provider will see the URL containing the search term. Furthermore, if the user clicks on the first search result, both the ISP and search engine provider can see the destination URL that user visits. As discussed in Chapter 1, however, when there is an HTTPS connection, as is increasingly standard for search,⁹ then the search engine retains its prior ability to see user activity, but encryption blocks the ISP from seeing anything except the host domain, such as www.bing.com or www.google.com.

To better describe the data that search engine providers receive from their search engines, the next subsection will focus on Google Search, as it currently has a large market share in the search engine space, and a large amount of publicly available information exists on the data Google collects through Google Search. Google's ability to collect this data demonstrates the larger point that search engine providers – which to date are not dominated by ISPs – often have wide visibility of user data through their search engines, which is significant and different than the visibility of an ISP.

2. Google Search

Launched in 1998, Google Search currently has the greatest market share among search engines. Estimates "about Google's market share in the United States put it between 67 and 75 percent of the search market."¹⁰ In 2014, there were an estimated 2.1 trillion Google searches, and an average of 5.7 billion daily Google searches.¹¹ Google is able to collect detailed data on users through these searches, apart from any data collected by Google in other contexts.

Google has the ability to collect every search a user conducts on Google Search as the user is typing terms into the search bar.¹² Further, a complete search term is not even necessary for this tracking, as Google can read each letter a user types into the search bar.¹³ When a user is typing a word into the search bar, Google will often try to "guess" what a user is trying to search.¹⁴ Even if a user deletes her search midway, Google can record what it thinks the user was searching, which may enable an inference about what the user was actually intending to search.¹⁵

⁸ IP-Author, "The Ways Google Collects Information about Us," *Pi Datametrics*, March 20, 2009, (<http://www.intelligentpositioning.com/blog/2009/03/the-ways-google-collects-information-about-us/>).

⁹ The impact of increased use of HTTPS in search is discussed further at the conclusion of this Chapter, and in Appendix 1 to Chapter 1, listing the prevalence of HTTPS in the top 50 Internet sites.

¹⁰ Robinson Meyer, "Europeans Use Google Way, Way More than Americans Do," *The Atlantic*, April 15, 2015, (<http://www.theatlantic.com/technology/archive/2015/04/europeans-use-google-way-way-more-than-americans-do/390612/>).

¹¹ "Google Annual Search Statistics," *Statistic Brain*, June 8, 2015, (<http://www.statisticbrain.com/google-searches/>).

¹² "What does Google do with the data it collects?" *Google*, (<https://privacy.google.com/data-we-collect.html>) (last modified Aug. 15, 2015); "The Ways Google Collects Information about Us," *Pi Datametrics*, March 20, 2009, (<http://www.pi-datametrics.com/the-ways-google-collects-information-about-us/>).

¹³ "What data does Google collect?" *Google*, (<https://privacy.google.com/data-we-collect.html>) (last modified Aug. 15, 2015); Robert Epstein, "Google's Gotcha," *U.S. News & World Report*, p. 2, May 10, 2013, (<http://www.usnews.com/opinion/articles/2013/05/10/15-ways-google-monitors-you>).

¹⁴ *Id.*

¹⁵ *Id.*

If a user uses Google Search while logged-in to Google, Google is more easily able to associate searches with an identified user account. Google would therefore be able to see all the searches that were completed by a particular account. Further, when users create a Google account, Google often asks for the following identifying information: name, user name, password, birthdate, gender, phone number, and an alternative email address.¹⁶ If Google has this identifying information, then it can not only associate searches with a particular user account, but the company also can associate the searches with a known individual. This fact is significant because the company can purchase more information about a user in the data marketplace, which can be used for tailoring advertising.

If a user is not logged in while using Google Search, Google initially may not be able to associate a search with a particular identified user account. However, there are multiple ways Google can associate searches as being from the same user or device. As further detailed in Chapter 6, Google could use first-party cookie tracking to associate searches with a user or device. When an individual uses Google Search, Google installs a cookie on her computer, thus enabling Google to track the user.¹⁷ In addition, once a user logs-in to other Google services including Gmail, the user is automatically logged-in to Google search on the same device, which allows the company to cross-reference data associated with a login to the data associated with searches.¹⁸ Google might also use cross-device tracking to identify a user using Google Search on different devices. Google would have the opportunity to combine its tracking data with other sources of data, including login data, to figure out the links between the devices.

Over time, Google can build a profile that may contain inferences about many aspects of a user's life.¹⁹ In a world of big data, it is often possible to map user attributes not only directly from searches, but also by what the searches imply. A search engine can then use this data, enhanced by big data analytics, to show relevant advertisements.

B. An Analogy of Search to Deep Packet Inspection ("DPI")²⁰

One long-standing concern about the role of ISPs has been DPI, the idea that an ISP could deploy technology in its network and use its position at the last mile to look deeply into the packets traversing the network and see full URLs and content.²¹ Modern search engine operation can provide comparable visibility because search engines index the web as an important part of offering the search service. Therefore, when a user clicks on a search result, the search engine provider has the ability to know more than the URL to which it sends the user. The search engine also can have visibility about the content of the destination page, where it has already indexed that content. For example, a user may type "electric cars" into a search engine. After looking through the search results, the user may then choose to view www.cars.org/electriccars. Because the search engine previously indexed that particular web page, when the user clicks on the result the search engine has already indexed – and thus has knowledge of – the content the user sees at the destination page.

¹⁶ "Create Your Google Account," *Google*, (<https://accounts.google.com/signup>).

¹⁷ "Types of Cookies Used by Google," *Google*, (<https://www.google.com/policies/technologies/types/>).

¹⁸ Robert Epstein, "Google's Gotcha," *U.S. News & World Report*, p. 2, May 10, 2013, (<http://www.usnews.com/opinion/articles/2013/05/10/15-ways-google-monitors-you>).

¹⁹ "What Does Google Do With the Data it Collects," *Google*, (<https://privacy.google.com/how-we-use-data.html>); Ryan Tate, "Big Google is Watching: Meet Your Creepy Google Dossier (and Mine)," *Gawker*, Nov. 5, 2009, (<http://gawker.com/5397993/big-google-is-watching-meet-your-creepy-google-dossier-and-mine>).

²⁰ The main function of ISPs is not to provide search engine functions, but rather to connect their customers to the Internet. Therefore, based on interviews, to date, ISPs do not generally index the web. The discussion here about the analogy to DPI thus would not generally apply today to ISPs.

²¹ See, e.g., Google's description of its search engine discovers, crawls, and serves web pages (<https://support.google.com/webmasters/answer/70897?hl=en>).

The original concerns about DPI involved the combination of URL (an ISP routed a user to a page) and content (an ISP might see deeply into the packet in the pre-encryption age). Search engine visibility to content similarly concerns the combination of the detailed URL (the search engine routes the user to a detailed link) and content (the search engine routinely indexes the content). The search engine provider can do this combination more often to the extent it has access to detailed URLs from other contexts, such as cookies used by an advertising network.²²

This discussion does not take a position on any regulatory or other action that would be appropriate for search engine content indexing. Indexing the web and sending users relevant content is a core activity of search, which is itself a major feature of the modern Internet. Instead, we offer a descriptive statement: Modern search engines develop a wide-ranging ability to see both URLs and content in connection with users' searches. As ISPs face the technical blocks on visibility discussed in Chapter 1, such as widespread encryption, other parts of the ecosystem have access to categories of data (combined URLs and content) that have been offered as reasons to regulate ISPs. Once again, examination of the potentially "unique" role of ISPs shows that other players in the ecosystem have similar, and sometimes greater, visibility into users' online activities.

C. The Utility of Search for Targeted Advertising

The information gathered by search engines is highly useful for targeted advertising purposes. The range of information a user provides – such as interests, shopping habits, health, and mental state – can be extensive.²³ As just discussed, once the search engine provides the results, it also has access to both URLs and content for the user's destination site. We highlight four ways search engines generate information relevant to advertising: insight into users' intent; targeted advertising; efficient auction ecosystem; and links with other applications, including digital assistants such as Apple's Siri, Google's Google Now, and Microsoft's Cortana.

1. Insight into users' intent

Search data provides advertisers with nuanced insight into each user's intent. One of the advertisers' goals is to reach consumers at the moments when they are most likely to make their purchasing decisions.²⁴ Advertisers have marked out these moments when consumers are more open to influence in a diagram named "the purchase funnel." In the purchase funnel, consumers begin with a number of potential brands in mind (the wide end of the funnel).²⁵ Advertisements are then directed at consumers as they reduce the number of brands and move through the funnel.²⁶ At the end of the funnel, consumers emerge with one brand and product that they decide to purchase.²⁷ Search is important in the context of the purchase funnel because it shows a consumer's proactive move. This move indicates a more advanced position in the purchase funnel, which is evidence that the consumer may intend to purchase the item. Search engine providers collect the search terms that users type in the web browser and can sell them to advertisers, who then serve advertisements designed to move the consumer further into the funnel toward a purchase. The more specific the search, the greater the value to advertisers because it

²² For instance, as discussed in Chapter 8 on cross-context tracking, Google's privacy policy enables it to combine information from different lines of business, such as Google Search, Google Analytics, and DoubleClick.

²³ Molly Wood, "Sweeping Away a Search History," *The New York Times*, April 2, 2014, (http://www.nytimes.com/2014/04/03/technology/personaltech/sweeping-away-a-search-history.html?_r=0).

²⁴ David Court, Dave Elzinga, Susan Mulder, and Ole Jørgen Vetvik, "The Consumer Decision Journey," *McKinsey & Company*, June 2009, (http://www.mckinsey.com/insights/marketing_sales/the_consumer_decision_journey).

²⁵ *Id.*

²⁶ Liz Serafin, "Digital Media & The Purchase Funnel: What to Use When?" *Geary LSF*, March 24, 2014, (<http://www.gearylsf.com/digital-media-the-purchase-funnel-what-to-use-when/>).

²⁷ "The Purchase Funnel," *Marketing-made-simple*, (<http://marketing-made-simple.com/articles/purchase-funnel.htm>).

closely shows the user's intent. For example, a user might decide he wants to get a dog for his household. He might then search "dogs." A dog breeding company that advertises might be somewhat interested in the user, but overall, searching this term does not indicate user's specific intent; his intent might be numerous things (e.g., researching dogs for leisure, buying dog products). Alternatively, if the user searches "German Shepherd breeders," his search is more telling of his intent: The user quite possibly may intend to buy a German Shepherd. The advertising company could perform analytics to find the likelihood that users who search this term will purchase a German Shepherd in the next six months. Those analytics help the advertising company learn what price it is willing to bid for the ad. The detailed search for "German Shepherd breeders" shows movement by a user further down the purchase funnel.

2. Targeted advertising

Search is useful because it contains responsive search terms that can immediately target advertisements to the user, which can be associated with a persistent identifier to show advertisements later based on those same terms. The process typically works by advertising buyers purchasing "keywords" that are representative of these search terms.²⁸ A simple example would be if a user searched "Brand A running shoes," "Brand B running shoes," and "running shoes size 9." In these searches, the keyword "running shoes" could trigger a real-time ad; in addition, the cookie associated with the user might allow the buyer to reach the user at a later date. Once the data is collected, the search engine provider could sell this data to advertisers, and the advertisers could then target running shoes advertisements to the user, based on the assumption that the user likely desires to buy running shoes. Advertisers would be able to serve an advertisement about running shoes precisely at the moment the user searches these keywords; therefore, the user is further down the purchase tunnel.

Generally, when search engine providers have collected a wealth of data and keywords, their ability to analyze keywords and identify patterns becomes more refined, thus enabling advertisers to serve targeted advertisements effectively. For example, in the running shoes example above, repeated searches of the "running shoes" keywords signal a high interest by the user, thus permitting advertisers to advertise accordingly. However, sometimes these patterns are not intuitive, which is why the ability to have large data sets is crucial. Outside of the search engine context, statisticians have studied consumption patterns to determine what these patterns may mean for a person's background, desires, etc. In a well-known example, statisticians at Target determined that pregnant women tend to buy larger quantities of products such as: unscented lotion; calcium, magnesium, and zinc supplements; and large bags of cotton balls.²⁹ This data permitted Target to advertise baby items specifically to likely-pregnant women who were buying these products.³⁰ Similarly, a search engine can analyze its data to find patterns that are useful for targeted advertising. In another example, if a search engine finds that searches for tuxedos, limousines, and roses mean a user is likely getting married, the search engine provider could sell that helpful information to advertisers, and the advertisers could target that individual with wedding-related advertisements.

3. Efficient auction ecosystem

This data provides search engines with an efficient auction ecosystem for purchasing advertisements. Advertisers can engage analytics professionals to help decide how much to spend on a "click" for a particular search term. In making this calculation, the advertiser will need to determine how many people coming to his website from an active search convert into being a customer, and how much profit the advertiser makes for each sale.

²⁸ "Keywords: What are Keywords & Why Do They Matter for PPC," *WordStream*, (<http://www.wordstream.com/keyword>).

²⁹ Kashmir Hill, "How Target Figured out a Teen Girl Was Pregnant before Her Father Did," *Forbes*, Feb. 16, 2012, (<http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>).

³⁰ *Id.*

For example, if the German Shepherd dog breeder (Doug) makes \$500 profit on each German Shepherd he sells, and he measures that one out of every 100 people coming to his website from an active search convert into being a customer (i.e. they make Doug \$500), then Doug will pay up to \$500 for 100 clicks, or \$5 per click.³¹ This simple example shows the general idea that search data provides advertisers with an efficient auction ecosystem for purchasing advertisements. In the real world, the advertisers would place cookies on users who clicked through to the particular web page, as well as on those who actually convert to being a customer. The advertiser would measure the number of clicks and then measure its revenue. From the revenue, advertisers would be able to determine their expected profit, and that profit divided by the conversion percentage would inform the advertiser what to bid on the various clicks.

Overall, insight into actual purchases is what leads to efficiencies. The data that search engines receive gives them insight into what items a user might purchase, close in time to the actual purchase. Searches are often done right before a purchase. For example, regarding local searches, Google has found that 50 percent of mobile users are likely to visit a store within a day of their search, with 34 percent of computer or tablet users likely to visit a store under the same circumstances.³² Further, these users are more likely to buy once they are in a store, as 18 percent of local searches lead to in-store purchases, compared to seven percent of non-local searches.³³ Even outside the local search context, Google has found that “nine out of 10 mobile search users have ‘taken action as a result of a mobile search, with over half leading to a purchase.’”³⁴ These statistics demonstrate how the ability to match buyers and sellers of targeted advertisements by using data derived from a search engine can have a profound impact on what consumers purchase. Hence, once a local specific search is made, it may be beneficial for advertisers to collect this data and serve targeted advertisements.

4. Links with other applications, such as digital assistant programs

As discussed in more detail in Chapter 5 on mobile operating systems, mobile search increasingly is being integrated with other applications on a smartphone, notably digital assistant programs such as Apple’s Siri, Google’s Google Now, and Microsoft’s Cortana. These digital assistants enable search results to assimilate data from other apps and content on the smartphone – such as maps or user contacts – when returning search results. Additionally, these digital assistants allow users to make so-called “natural language” searches. Access to other app data in the smartphone allows these programs to infer the context of search requests like “who is this?” recognizing that a song is playing and fetching the name of the artist.³⁵

D. Conclusion

Despite the widely-held view that ISPs have comprehensive and unique access into a user’s online activity, Chapter 1 showed that ISPs have far from comprehensive visibility into a user’s Internet activity history. In contrast, this Chapter showed the many unique insights that search engines gain, and how those insights have translated into a large share of all online advertising. Given their ability to do “DPI 2.0,” search engines have extensive access both to detailed URLs and to content accessed by users, and thus to a large portion of a user’s Internet activity history.

³¹ Note that this is a simple example for illustration only, and the prices are not necessarily representative.

³² Jessica Lee, “Google: Local Searches Lead 50% of Mobile Users to Visit Stores,” *Search Engine Watch*, May 7, 2014, (<http://searchenginewatch.com/sew/study/2343577/google-local-searches-lead-50-of-mobile-users-to-visit-stores-study>).

³³ *Id.* Diagram 6-B in Chapter 6, Interest Based Advertising, shows the higher value of geographically relevant searches. In the example there, the mobile advertisement is worth 12 cents rather than 10 cents for a desktop-based ad viewed from the user’s home.

³⁴ Greg Sterling, “Google: 50 Percent of Those Exposed to Mobile Ads Took Action,” *Search Engine Land*, April 26, 2011, (<http://searchengineland.com/google-50-percent-of-smartphone-users-exposed-to-ads-took-action-74760>).

³⁵ For a more detailed discussion of the cross-context implications of digital assistants, see Chapter 8 on cross-context tracking.

Going forward, the ability of ISPs to see search engine-related data is very limited. None of the traditional ISPs are among the three largest search engines in the U.S. Major search engines have shifted to encrypted search, blocking any previous ability of ISPs to see the search results and destination pages a user visits as a result of the search. Google Search now uses HTTPS by default, and³⁶ Bing has announced that its shift to HTTPS by default is underway.³⁷ When a search engine uses HTTPS, the ISP can see the root domain, such as www.google.com, but no further detail. In addition, beyond search engines, a large and growing portion of destination sites also use HTTPS; this is true for a wide range of e-commerce, financial services, and many others, as shown in Appendix 1 to Chapter 1.³⁸ Where that occurs, the ISP loses any technical ability to see the detailed URL or visit the content that the user accesses. The insights from search engines are uniquely available to others in the online ecosystem, but not to ISPs.

³⁶ Danny Sullivan, "Google to Begin Encrypting Searches & Outbound Clicks by Default with SSL Search," *Search Engine Land*, Oct. 18, 2011, (<http://searchengineland.com/google-to-begin-encrypting-searches-outbound-clicks-by-default-97435>).

³⁷ "Bing Moving to Encrypt Search Traffic by Default," *Bing Blogs*, June 15, 2015, (<https://blogs.bing.com/webmaster/2015/06/15/bing-moving-to-encrypt-search-traffic-by-default/>).

³⁸ Appendix 1 to Chapter 1 lists the top 50 websites' use of HTTPS.

Webmail and Messaging

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 4: Webmail and Messaging

Emails are an important potential source of information about users' online activity because they often reveal information that can be useful for advertising purposes. Words and phrases in an email can be a useful trigger for an advertisement. For instance, mentioning a brand name may be an advertising opportunity for the company whose brand is mentioned, or for its competitors. Beyond the content analysis of an individual email, analyzing users' emails over time is a rich source for big data analytics; it enables a detailed understanding of users' interests, and thus premium pricing for interest-based advertisements.

This Chapter describes trends in webmail providers' data collection and use, led by non-ISPs such as Google (Gmail), Microsoft (Outlook.com, previously called Hotmail), and Yahoo (Yahoo! Mail).¹ First, the Chapter discusses the commercially valuable data that webmail providers are able to collect by scanning email content. Second, it discusses reasons why webmail providers collect this data. Webmail has become a significant source of advertising activity. For example, Google's Terms of Service and Yahoo! Mail's Frequently Asked Questions both state that these companies may use webmail data in order to serve personalized advertisements. Google's Customer Match service also permits Google to build advertising campaigns based on email lists. Third, the Chapter discusses how in recent years, the great majority of webmail has shifted to HTTPS; the content between the user and the webmail provider is encrypted, and thus not accessible by the Internet Service Provider ("ISP"), even if it should wish to read that content. The Chapter concludes by comparing the rich data available to webmail providers with the lack of similar data available to ISPs.

As in the previous Chapters regarding social networks and search engines, this comparison of webmail providers and ISPs undermines the widely-held, but mistaken view that ISPs have comprehensive and unique access to and knowledge about users' online activity because they operate the last mile of the broadband access service. The most prominent webmail providers are non-ISPs. As discussed in the later Chapters on cross-context tracking (Chapter 8) and cross-device tracking (Chapter 9), webmail providers' information-gathering advantages combine with sources of user data from other settings, further showing that unique insights about users often come from non-ISPs.

Although the focus of detailed analysis in this Chapter is on webmail services, person-to-person communication by webmail has been supplemented by a range of other messaging applications such as Facebook Messenger, Jabber, Kik, Line, Skype, Snapchat, Viber, WeChat, and WhatsApp. For each of these apps, information may be collected and used for advertising based on data or metadata, and data from these messaging applications may be used for cross-context and cross-device tracking. For clarity and because it has been especially prominent in advertising to date, the details of this Chapter focus on webmail providers, but the same analysis applies to other messaging methods to the extent that the data and metadata are used for marketing purposes.

A. Data and Metadata in Webmail

Webmail providers collect data through their webmail services by scanning email content as well as metadata, such as email addresses, time, date, and file size. A content scan can include embedded photographs and attachments. Webmail providers are not limited to scanning emails within the same webmail service; rather, many webmail providers scan both incoming and outgoing emails, including emails coming from different email

¹ The term "webmail providers" includes all entities that provide webmail services. By its definition, "webmail providers" also includes ISPs that provide email accounts. Each webmail provider's practices differ and an examination of each provider's practices is necessary to determine its specific data collection practices.

providers.² Further, many webmail providers are able to read emails even when they have been abandoned and are never sent, such as “draft” emails.³

Webmail providers’ ability to scan email and collect this data provides them with large amounts of data. As of December 2012, Gmail, Hotmail, and Yahoo! Mail together had over 1 billion users.⁴ This number has increased since that time and Google alone now has over 900 million Gmail users.⁵ None of these leading providers are traditional ISPs. As an increasing number of people create these webmail accounts, or simply interact with and email individuals who have these accounts, webmail providers collect more data. Individuals reveal information about all facets of their life via email, such as their thoughts, ideas, goals, fears, etc. This information is not only current, but may reflect a user’s past or future. Webmail providers have the ability to see this data, which they can use in different ways.

By scanning emails, webmail providers are able to view the content that users send in their emails, as well as traffic data, which includes who the sender and recipient are, in addition to the date and time of the email.⁶ As discussed in Chapter 1, the pervasive use of encryption for webmail means that ISPs can see the host name, such as <https://www.gmail.com>, but not the traffic data or content of the email. Therefore, ISPs have much lower visibility into webmail information than the webmail providers.

B. Purposes of Collecting Data

1. Security

The reasons webmail providers scan email and collect data have varied over time. Initially, the purpose of scanning email was twofold: (1) to preserve the integrity of the webmail system and computers by preventing any applicable viruses, and checking and identifying any spam or malware, and (2) specifically, in the case of Google, “to identify child sexual abuse imagery” (i.e., child pornography).⁷

Regarding the second point, Google, which deployed its webmail service Gmail in 2004, actively scans content that passes through Gmail accounts.⁸ In July 2014, while scanning this content, Google became aware that a Gmail user was sending indecent images of children to a friend.⁹ Upon learning this information, Google passed

² Robert Epstein, “Google’s Gotcha,” *U.S. News & World Report*, p. 2, May 10, 2013, (<http://www.usnews.com/opinion/articles/2013/05/10/15-ways-google-monitors-you>).

³ *Id.*

⁴ Mark Brownlow, “Email and Webmail Statistics,” *Email Marketing Reports*, Dec. 2012, (<http://www.email-marketing-reports.com/metrics/email-statistics.htm>).

⁵ Craig Smith, “By the Numbers: 12 Amazing Gmail Statistics,” *DMR Digital Marketing Stats/ Strategy/ Gadgets*, May 31, 2015, (<http://expandedramblings.com/index.php/gmail-statistics/>).

⁶ Various other related terms are used, including meta-data, header data, and transactional data. The term “traffic data” is defined in the Budapest Convention and “means any computer data relating to a communication by means of a computer system, generated by a computer system that formed a part in the chain of communication, indicating the communication’s origin, destination, route, time, date, size, duration, or type of underlying service.” For discussion of the term, see “Criminal Justice Access to Data in the Cloud: Challenges,” *Council of Europe Cybercrime Convention Committee*, 2015, ([http://www.coe.int/t/dghl/cooperation/economiccrime/Source/Cybercrime/TCY/2015/T-CY\(2015\)10_CEG%20challenges%20rep_sum_v8.pdf](http://www.coe.int/t/dghl/cooperation/economiccrime/Source/Cybercrime/TCY/2015/T-CY(2015)10_CEG%20challenges%20rep_sum_v8.pdf)).

⁷ Rich McCormick, “Google Scans Everyone’s Email for Child Porn, and It Just Got a Man Arrested,” *The Verge*, Aug. 5, 2014, (<http://www.theverge.com/2014/8/5/5970141/how-google-scans-your-gmail-for-child-porn>).

⁸ Harry McCracken, “How Gmail Happened: The Inside Story of Its Launch 10 Years Ago,” *Time*, April 1, 2014, (<http://time.com/43263/gmail-10th-anniversary/>).

⁹ Hayley Tsukayama, “How Closely Is Google Really Reading Your Email?” *The Washington Post*, Aug. 4, 2014, (<https://www.washingtonpost.com/blogs/the-switch/wp/2014/08/04/how-closely-is-google-really-reading-your-e-mail>).

the details to the police via the National Center for Missing and Exploited Children (“NCMEC”) and the culprit was arrested.¹⁰ This case demonstrates benefits from scanning email; it also shows the personal and potentially embarrassing content that webmail providers may access.

2. Advertising

i. Google and Yahoo terms of use

Over time, webmail providers have expanded the scope of their scanning activities. Google and Yahoo are two webmail providers that today scan email to target advertisements. For example, Google’s Terms of Service permits Google to scan email content for tailored advertising by stating, “Our automated systems analyze your content (including emails) to provide you personally relevant product features, such as customized search results, tailored advertising, and spam and malware detection. This analysis occurs as the content is sent, received, and when it is stored.” Yahoo! Mail’s Frequently Asked Questions states that communications are scanned and analyzed to detect “certain words and phrases” and that this “might result in ads being shown to you...for products and services that are related to those keywords.”¹¹

The main idea is that webmail providers are scanning users’ emails for keywords.¹² These keywords result in advertisements. For example, if the user mentions “headaches” in emails to her sister, she might later be shown an advertisement for pain medication. This email content is an example of data – which can be private and personal – that webmail providers scan and then use for advertising purposes.

ii. Google’s Customer Match

Some companies are building advertising campaigns based around email lists they upload and compare with their own email databases. For example, in September 2015, Google announced Customer Match, a new tool that allows advertisers to target specific users across Google services (e.g., Gmail, YouTube) by uploading a list of email addresses.¹³ In order to use this tool, advertisers upload email addresses into the Audiences tab of Google’s AdWords service.¹⁴ Advertisers are able to collect these email addresses from various sources, such as store loyalty cards, mailing lists, or email receipts.¹⁵ Once uploaded, Google matches these email addresses with any Google accounts that share the same email address.¹⁶ After the connection is made, the advertiser can serve that particular user targeted advertisements when she is using Google’s services.¹⁷

¹⁰ *Id.*

¹¹ “Yahoo Mail FAQ,” *Yahoo!*, (<https://policies.yahoo.com/us/en/yahoo/privacy/products/mail/faq/index.htm>).

¹² For the purposes of this section, keywords are informative words used to indicate the content of an email.

¹³ Jason Tabeling, “How to Use Google’s New Customer Match Feature,” *Search Engine Watch*, Oct. 2, 2015, (<http://searchenginewatch.com/sew/opinion/2428080/how-to-use-googles-new-customer-match-feature#>).

¹⁴ Erin Sagin, “How to Use AdWords’ Customer Match: The Ultimate Guide,” *WordStream*, Nov. 2, 2015, (<http://www.wordstream.com/blog/ws/2015/11/02/adwords-customer-match-setup>).

¹⁵ Nick Statt, “Google Will Let Companies Target Ads Using Your Email Address,” *The Verge*, Sep. 28, 2015, (<http://www.theverge.com/2015/9/28/9410975/google-customer-match-ad-targeting-email-addresses>).

¹⁶ *Id.*

¹⁷ “Google Brings You Closer to Your Customers in the Moments That Matter,” *Google Inside AdWords*, Sep. 27, 2015, (<http://adwords.blogspot.com/2015/09/Google-brings-you-closer-to-your-customers.html>).

In explaining how Customer Match works, Sridhar Ramaswamy, SVP of ads and commerce at Google, stated the following:

Let's say you're a travel brand. You can now reach people who have joined your rewards program as they plan their next trip. For example, when these rewards members search for "non-stop flights to New York" on Google.com, you can show relevant ads at the top of their search results on any device right when they're looking to fly to New York. And when those members are watching their favorite videos on YouTube or catching up on Gmail, you can show ads that inspire them to plan their next trip.¹⁸

This tool is particularly powerful because it allows advertisers to target their advertisements in different situations. First, Customer Match permits advertisers to target customers who have purchased their goods in the past – but who have not purchased anything recently – by sending an advertisement related to something that they have previously purchased.¹⁹ Second, Customer Match allows advertisers to cross-sell to existing customers.²⁰ Specifically, advertisers can target customers by showing them supplementary or complementary products based on the products the customer has previously purchased.²¹ Third, Customer Match contains a Similar Audiences product that allows advertisers to target others using Google services who are likely interested in the same products and services as the advertiser's existing customer base.²²

Customer Match illustrates the power of data, particularly with respect to email addresses. A consumer's email address, combined with Google's data sets, permits very specific targeted advertisements to specific Google users.

C. The New Prevalence of HTTPS

1. The overall shift toward encrypted email

The technical possibility of ISPs scanning email content is decreasing due to the increased use of encryption for email services.²³ The spread of encryption has happened in stages.²⁴ First, when webmail and other email services were first deployed, they often were sent in plaintext. This meant, in theory, that an ISP could use deep packet inspection ("DPI") to see the content of emails as they passed between the user and the webmail provider's server. Second, in order to provide a more secure solution, webmail providers began to use HTTPS when the user sent an email to the webmail provider's server. In this approach, the outbound user email was encrypted on its way to the webmail provider's server, and there the email was decrypted and subject to the provider's scanning.²⁵ If the email was sent to another user of the same webmail service, then the email would be encrypted from the server to that recipient. This development provided added security, as emails were not transmitted over the Internet in the clear. Third, in recent years major webmail providers have cooperated so encryption would also

¹⁸ Robert Hof, "New Google Ads Take a Page (or Two) From Facebook," *Forbes*, Sep. 28, 2015, (<http://www.forbes.com/sites/roberthof/2015/09/28/new-google-ads-take-page-from-facebook/>).

¹⁹ *Id.*

²⁰ *Id.*

²¹ *Id.*

²² Lara O'Reilly, "Google Is Copying Two Features Advertisers Love about Facebook," *Business Insider*, Sep. 28, 2015, (<http://www.businessinsider.com/google-launches-customer-match-similar-audiences-and-universal-app-campaigns-2015-9>).

²³ This refers to when ISPs are acting in their traditional role of connecting users to the Internet, not in the role of providing webmail services.

²⁴ Peter Swire, "From Real-Time Intercepts to Stored Records: Why Encryption Drives the Government to Seek Access to the Cloud," *International Data Privacy Law* (2012), (<http://ssrn.com/abstract=2038871>).

²⁵ *Id.*

take place when the email passed from the sender's webmail service to the recipient's email service provider.²⁶ In this approach, the email is encrypted when in transit, such as between user-provider, provider-provider, and provider-recipient. However, the email²⁷ is generally in plaintext at each provider's server, and thus subject to scanning. To date, a low fraction of overall email traffic has been encrypted at all stages from the sender to the recipient, which is often called "end-to-end" encryption.²⁸ With that sort of encryption, even the webmail provider does not have access to the content.

2. The Google Gmail example

The well-documented history of Google's encryption illustrates the shift toward encryption, which more thoroughly blocks even the theoretical possibility of scanning by ISPs. When Google first introduced its Gmail services, it did not enable HTTPS encryption by default due to slow email speeds. In a blog post entitled "Making Security Easier," a Google spokesman defended Google's position to not do more thorough encryption, by stating:

We use [HTTPS encryption] to protect your password every time you log into Gmail, but we don't use [HTTPS encryption] once you're in your mail unless you ask for it . . . Why not? Because the downside is that [HTTPS encryption] can make your mail slower. Your computer has to do extra work to decrypt all that data, and encrypted data doesn't travel across the Internet as efficiently as unencrypted data. That's why we leave the choice up to you.²⁹

After receiving criticism, Google enabled HTTPS encryption for Gmail users in 2008 so there was an option to encrypt webmail traveling between a user's web browser and Google's servers.³⁰ Still, there was not encryption when the recipient used a different email service. Citing statistics such as "approximately 40 to 50 percent of emails sent between Gmail and other email providers aren't encrypted,"³¹ in June 2014, Google addressed this security concern by introducing a Chrome extension that would encrypt webmail messages.³² In a blog post, Google's Stephan Somogyi described the new technology by stating, "'End-to-End' encryption means data leaving your browser will be encrypted until the message's intended recipient decrypts it, and that similarly encrypted messages sent to you will remain that way until you decrypt them in your browser."³³

²⁶ Emails sent between servers use protocols such as Simple Mail Transfer Protocol Secure ("SMTPS"), a variation on the long-used SMTP email protocol.

²⁷ "Transparency Report: Protecting Emails as They Travel across the Web," *Google Official Blog*, June 3, 2014, (<https://googleblog.blogspot.com/2014/06/transparency-report-protecting-emails.html>).

²⁸ Andy Greenberg, "Hacker Lexicon: What Is End-to-End Encryption," *Wired*, Nov. 25, 2014, (<http://www.wired.com/2014/11/hacker-lexicon-end-to-end-encryption/>).

²⁹ Christopher Soghoian, "Caught in the Cloud: Privacy, Encryption, and Government Back Doors in the Web 2.0 Era," *Journal on Telecommunications and High Technology Law* 8, 359, Aug. 17, 2009, p. 376-377 (http://www.jthtl.org/content/articles/V8I2/JTHTLv8i2_Soghoian.PDF), citing "Posting of Ariel Rideout to the *Official Gmail Blog*, Making Security Easier," *Official Gmail Blog*, July 24, 2008, (<http://gmailblog.blogspot.com/2008/07/making-security-easier.html>).

³⁰ *Id.* at 376-377.

³¹ "Transparency Report: Protecting Emails as They Travel across the Web," *Google Official Blog*, June 3, 2014, (<https://googleblog.blogspot.com/2014/06/transparency-report-protecting-emails.html>).

³² Stephan Somogyi, "Making End-to-End Encryption Easier to Use," *Google Online Security Blog*, June 3, 2014, (<https://googleonlinesecurity.blogspot.com/2014/06/making-end-to-end-encryption-easier-to.html>).

³³ *Id.* We note that the term "end-to-end encryption" generally is used to indicate that only the sender and receiver (the two ends) have access to the plaintext. In this Google deployment, the email is encrypted whenever in transit, but Google retains the ability to scan plaintext at its server.

Google is not alone in encrypting its email. Last year, Comcast began encrypting its webmail. In November 2013, Microsoft introduced Office 365 Message Encryption that allows users to send automatically encrypted emails to recipients outside their webmail service.³⁴

This move toward pervasive email encryption shows that going forward, ISPs are unlikely to be able to view email content. Even if ISPs seek to deploy DPI, they cannot access that content.

D. Conclusion

Despite the widely-held view that ISPs have comprehensive access into users' online activity, Chapter 1 showed that ISPs face major technical limits on their visibility into a user's Internet activity history. By contrast, this Chapter showed that providers of webmail and other messaging services, the leaders of whom are non-ISPs, are uniquely able to scan their users' messages to receive commercially valuable data. While this scanning was originally done for security purposes, it is increasingly being done for advertising purposes. Specifically, this scanning can be used to identify keywords present in sent and received messages, and the keywords can then be used for targeted advertising. In addition, data gleaned through webmail services can be combined in various ways with other sources of user data, such as in the cross-context tracking that will be discussed in Chapter 8 and the cross-device tracking that will be discussed in Chapter 9.

³⁴ Steven Musil, "Encrypted Messaging Coming to Microsoft's Office 365 Next Year," *CNet*, Nov. 21, 2013, (<http://www.cnet.com/news/encrypted-messaging-coming-to-microsofts-office-365-next-year/>).



How Mobile Is Transforming Operating Systems

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 5: How Mobile Is Transforming Operating Systems

When it comes to the technical capability of tracking user activity, no software or service is as comprehensive as the operating system (“OS”). Historically, for desktops and laptops, the OS generally did not use its position on a user’s device to collect or track online behavior for advertising purposes. The data flows to OS developers have substantially changed, however, in the shift to mobile computing, notably on smartphones and tablets.

This Chapter examines major shifts in data flows that arise in mobile computing, including for OS developers such as Apple, Google, and Microsoft. The Chapter begins by explaining the technical possibility of, but historical limits on, data flows to OS developers, often focused on telemetry. The Chapter then begins to explore the changes that come in mobile OS, such as the use of an “advertising ID” that is usually persistent and identical across multiple apps. These IDs often enable both the OS developer and the app developers to track usage, and they have enabled a marketplace where information about usage associated with that advertising ID can be combined across companies.

The Chapter next examines the emerging mechanisms for location tracking, which is of intense interest to marketers seeking to tailor advertisements to a user’s current physical location. Smartphone location today can generally be accurately tracked from global positioning system (“GPS”) data and from the location of WiFi hotspots the smartphone is near. Going forward, other data sources of a smartphone can reveal location information, including Bluetooth, its magnetometer, and others. In addition, location can be estimated by the generally less accurate method of trilateration, which is based on distance from cell towers. Location information from various sources is often “democratized” in the online advertising market, as the data goes to OS developers, app developers, and third parties who buy and sell location information of a device.

The Chapter then examines three OS market leaders on traditional and mobile devices (Apple’s iOS, Google’s Android, and Windows 10), including the ability to track a user’s history of Internet activity. The discussion introduces three themes:

1. *The prominence of app stores led by OS developers* such as Apple’s AppStore or Google’s Play Store. These app stores succeed by attracting app developers, who in turn are attracted to a platform that provides advertising revenue and enables ad-supported apps.
2. *The importance of personal assistants* such as Apple’s Siri, Google’s Google Now, and Microsoft’s Cortana. In order to answer a user’s questions about calendar, location, and other topics, the personal assistants integrate data from the relevant apps, *often including data stored in the OS developer’s cloud*. The growing prevalence of personal assistants and cloud computing means that operating systems gather detailed data from across the device in order to answer user queries.
3. In contrast to their historical role in gathering narrow categories of data, such as telemetry, *modern OS developers thus become far more integrated in data streams relevant to online advertising*. The precise roles of OS developers in these new data streams have not been a major focus to date of policymakers.

As with the other chapters in this Working Paper, this Chapter seeks to explain developments in an important context for gathering data about users’ online activities, in this instance the role of operating systems, especially for mobile computing. The ability of any one company to gain insight into user behavior will depend on the data it gathers across contexts (Chapter 8, Cross-Context Tracking) and across devices (Chapter 9, Cross-Device Tracking). The Chapter concludes by comparing the data tracking capabilities of modern Internet Service Providers (“ISPs”) to modern OS developers (and others in the mobile ecosystem such as app developers). As discussed here, the OS has the technical possibility of visibility into all of the actions on a user’s device, and the trends in mobile

computing are toward substantially greater OS access and use of such data for advertising-related purposes. These trends are important due to the growth in mobile advertising, which accounted for 38.5 percent of total digital ad spending in 2014, became a majority by the end of 2015, and is projected to reach 70 percent by 2019.¹

A. History of OS Collection of User Information

Historically, the leading OS providers have collected limited data about usage, but the technical ability to collect far more data exists, and there are important trends in that direction. By necessity, an OS has a complete view of all the activity on any given device. All inputs and outputs, whether from hardware or software, are handled by the OS. This means that by necessity the OS has the technical ability to see every keystroke entered, word typed, and image viewed. Even messages sent using end-to-end encryption must first be entered on a device in plaintext, and later decrypted to plaintext on a device in order to view. Therefore, any OS has the capability to both observe and record all of a device user's activity, including the Internet history and app usage activity for that device.

The OSes typically collected data for performance monitoring purposes. These telemetry or automated remote data collection services monitored a device's performance and collected data on things like major system crashes, errors, and other unusual events. Collecting this data allowed the manufacturer of the OS to update the software's performance and problems throughout the life of the OS. Telemetry data was and is vital to the life and success of an OS. Today, software testing can discover and fix many problems before a product is released, but the system is inherently imperfect. Many problems with any individual piece of software occur only in specific conditions, during interactions with other specific pieces of user software or hardware, or are just rare enough to be overlooked by quality assurance testing. Additionally, the more complex a program, the greater the number of potential conflicts and issues. The OS is inherently one of the most complicated pieces of software on a given device. Therefore, the OS is likely to have a large number of newly discovered problems throughout its lifecycle, necessitating ongoing monitoring and repair.² Traditionally a user would be automatically opted-in to sharing telemetry data, but could opt-out from sharing this data.

This limited amount of data collection could be changed, as a technical matter, to collect and monitor other information that passes through the OS's control. For example, an OS could theoretically decide to collect keystroke data for each of its users, sending these logs of all the keys pressed by a user back to the OS developer. More broadly, the OS has visibility into data such as: the URLs entered, the searches performed, the local cache of pages visited and their content, saved login information, the amount of time spent viewing particular pages, the files downloaded or uploaded, and all other browsing data. What the OS can see on a device, it can also be programmed to store and share via reports to its developer.

The OS's potential control extends to data recorded by attached hardware. Webcams and microphones, for example, are ultimately controlled by the OS. Consequently, an OS could run a program independent of the user's control that records data from these devices. Indeed, the OS could seek to hide these processes from a user, making it unclear that a new piece of software is running on the device. In modern smartphones, the range of sensors has increased, broadening the potential scope of data collection.

¹ "Mobile to Account for More than Half of Digital Ad Spending in 2015," *eMarketer*, Sep. 1, 2015, (<http://www.emarketer.com/Article/Mobile-Account-More-than-Half-of-Digital-Ad-Spending-2015/1012930>).

² Indeed, the official lifecycle of an OS can be measured as the time until it no longer receives official updates from the manufacturer. For Windows XP, Microsoft "officially" stopped providing support as of April 8, 2014, but it has agreed to provide extended support for the U.S. Navy and Army. "Federal Business Opportunities: Extension to Contract N00039-14-C-0101," April 20, 2015, (https://www.fbo.gov/index?s=opportunity&mode=form&id=cab6679781ccb5228673f053246d1654&tab=core&_cview=0). Sean Gallagher, "Navy re-ups with Microsoft for more Windows XP Support," *Ars Technica*, June 23, 2015, (<http://arstechnica.com/information-technology/2015/06/navy-re-ups-with-microsoft-for-more-windows-xp-support/>).

Despite these technical possibilities, the leading OSes have not historically monitored and collected complete usage data on a device. Such actions would be detected, raising objections from users, regulators, and others in the marketplace. For telemetry and other data collection, the data is only useful to the OS developer once transmitted back to the developer. Sophisticated users can detect those transmissions and unmask such practices. Historically, when users suspected this kind of monitoring, software developers have been pushed to respond.³ The OS was not historically a significant means of collecting personal data for advertising purposes, despite the potential wealth of information available.

B. How the Growth of Mobile Impacts Operating Systems

The shift toward mobile computing has led to important changes in the role of the OS. This section examines the role of non-cookie device identifiers in the leading mobile OSes. These alternate device identifiers can provide substantial information about user activity to OS and app developers, and to third parties to buy and sell such information.

1. Mobile device identifiers

Cookies are generally less useful for mobile devices than for laptops and desktops.⁴ Both Android and iOS assign unique advertising identifiers to devices and allow apps to access that unique identifier for each device running the operating system. These identifiers, similar to cookies, allow data about user activity to be synced across apps.

Android assigns a unique user-specific advertising ID that mobile apps can use for targeted advertising. Under the terms of the Android Software Development Kit (“SDK”), the advertising ID may be used for advertising and user analytics, but it is not allowed to be connected to personally identifiable information (“PII”) or any other persistent device ID (e.g., SSAID, MAC Address, IMEI) without user consent.⁵ There do not appear to be technical controls to enforce these terms in the SDK, so it is unclear whether app developers are complying with the ban on linking to PII or other persistent IDs.

An Android user can reset the advertising ID or opt-out of interest-based advertising, but there is no option for turning off the Advertising ID. The SDK requires app developers to respect a user’s decision to reset the advertising ID or opt-out of interest-based advertising, in which case the allowed activities are limited to “contextual advertising, frequency capping, conversion tracking, reporting and security and fraud detection.”⁶

Similarly, iOS assigns a unique “advertising identifier” to a mobile device that mobile apps can use for advertising purposes. Under the terms of the iOS SDK, the advertising identifier can only be used for serving advertising or for analytics tied to advertising, but once again there do not appear to be technical controls to enforce this limit.⁷ An iOS user can reset the advertising identifier or opt-out of interest-based advertising, but there is no option for turning off the advertising identifier. The SDK requires app developers to respect a user’s decision to reset the advertising identifier or opt-out of interest-based advertising, in which case it can be used only for “frequency capping, attribution, conversion events, estimating the number of unique users, advertising fraud detection,

³ For instance, video game distribution and management platform Steam by the Valve Corporation was the subject of a user-based investigation into DNS logs that were being sent back to Valve as part of an anti-cheating software, leading Valve to respond publicly to assure users that it was only comparing a hashed log of DNS against a blacklist of malicious DNS servers. Peter Bright, “Valve DNS privacy flap exposes the murky world of cheat prevention,” *Ars Technica*, Feb. 17, 2014, (<http://arstechnica.com/gaming/2014/02/valve-dns-privacy-flap-exposes-the-murky-world-of-cheat-prevention/>).

⁴ See Chapter 6 for a discussion of the limitations of cookies in the mobile setting.

⁵ “Android Software Development Kit License Agreement,” *Android*, (<http://developer.android.com/sdk/terms.html>)

⁶ “Google’s Advertising Identifier,” *Tune Help*, Feb. 21, 2014, (<https://help.tune.com/marketing-console/googles-advertising-identifier/>).

⁷ Apple SDK agreement.

debugging for advertising purposes only, and other uses for advertising that may be permitted by Apple.”⁸ Once again, there do not appear to be technical controls to enforce these terms in the SDK.

The advertising identifiers described above are global, that is to say that all apps on a device use the same identifier. This global identifier enables ad networks to easily match a user’s device across multiple apps, and enables further sale of the information to third parties, who can match the information provided to other information that uses the same identifier. The advertising identifiers in practice are likely more comprehensive than browser tracking cookies. Cookies are unique to the domain that sets the cookie, unlike the mobile identifiers that are the same across all of the apps.⁹

Particular mobile devices can also be identified through insertion of a unique identifier in the header of web GET and POST calls (hereinafter referred to as “UIDH” for unique identifier in the header). Similar to the advertising identifiers provided by mobile device operating systems, a network provider may insert a UIDH in the header information that is present in web calls a user transmits over the mobile network. In Verizon’s implementation of the UIDH, a user may opt-out of the advertising program that uses the UIDH and the UIDH will no longer be inserted in web headers.¹⁰

There are the now-familiar technical barriers to the comprehensiveness of tracking by means of a UIDH. The greatest technical roadblocks are those discussed in Chapter 1. First, usage is shifting to multiple mobile devices, so the prevalence of the UIDH from any one provider is limited. Second, encryption such as HTTPS blocks the insertion of a UIDH into an encrypted transmission. As discussed in Chapter 1, an ISP can see, at most, the host domain a user visits, but not the detailed URL of the sub-page visited.¹¹ Third, VPNs and proxy services similarly block insertion of a UIDH. Due to these technical developments, a large and growing portion of user sessions and bits transferred take place through one or more of these blocking technologies.¹²

Mobile device fingerprinting can also create device-specific identification, but such fingerprinting appears to be less effective for mobile devices than for laptops and desktops.¹³ For instance, mobile browsers rarely use plug-ins, a key differentiator for non-mobile fingerprinting. Screen resolutions tend to be more uniform across mobile devices, reducing the uniqueness of fingerprints. In addition, settings like font, if changed from the browser default, can be difficult to detect from a mobile browser. So, while device fingerprinting may be feasible in the mobile ecosystem, other device identifiers are easier and more reliable to use.

⁸ “How to Submit Your App When it Uses IDFA,” *Tune Help*, April 11, 2014, (<https://help.tune.com/marketing-console/how-to-submit-your-app-when-it-uses-idfa/>).

⁹ As described in Chapter 6, cookie syncing can enable the linkage of cookies from unrelated ad networks, but the effectiveness of such syncing is reduced because each cookie is unique to its own domain.

¹⁰ “Verizon Wireless’ Use of a Unique Identifier Header (UIDH),” *Verizon*, (<http://www.verizonwireless.com/support/unique-identifier-header-faqs/>).

¹¹ Because the header in the encrypted message is in plaintext only somewhere other than the ISP, the ISP cannot insert information into the header.

¹² Along with these technological limits on its prevalence, Verizon has made two significant changes to its initial use of the UIDH. First, it announced that it will not insert the UIDH after a customer opts out of the Relevant Mobile Advertising program or activates a line that is ineligible for the advertising program. “Verizon Wireless” use of a Unique Identifier Header (“UIDH”), *Verizon*, (<http://www.verizonwireless.com/support/unique-identifier-header-faqs/>). Second, Verizon announced the UIDH will be included only in traffic that is sent to companies on a white list, and the only companies on that white list will be Verizon companies (including AOL) and certain partners who will be required to only use the UIDH for Verizon purposes. “What Verizon’s Privacy Updates Really Mean,” *Verizon*, (<https://www.verizon.com/about/news/what-verizons-privacy-updates-really-mean>).

¹³ For a discussion of non-mobile device fingerprinting, see Chapter 6.

2. Mobile applications

The existence of effective mobile device IDs, such as the Android and Apple advertising IDs, provides the basis for data collection and thus monetization for mobile app developers. Such collection is consistent with the SDK offered by mobile OSes on smartphones, tablets, and other mobile devices. Some categories of user data, such as mobile device location and access to user contacts, require additional user consent and can be controlled from the device. Once the mobile app has collected user data, that data can provide revenue to the app developer – it may be used for advertising by the app developer or sold to other companies that gather information from multiple apps.

In the case of Android, mobile apps can obtain access to multiple types of data from the mobile device, including various data related to the use of ISP services:

- Unique device ID (“IMEI”) and IP address;
- Web browsing, bookmarks, and app usage history;
- Fine and coarse mobile device location, including location derived from cell towers, WiFi, and Bluetooth (discussed further below);
- WiFi history;
- Call log and SMS message history;
- User photos, videos, contacts, and calendar;¹⁴ and
- Data generated by the microphone, camera, other sensors.

Many mobile apps share this customer data with third parties as a way to support offering the app for free without imposing subscription fees.¹⁵ The consumer data from an individual app may be aggregated with data from other apps to make it more valuable to advertisers.

More recently, mobile apps have migrated from mobile devices to laptops and desktop computers. This allows app providers to correlate data across a growing number of devices.¹⁶ Evernote is one popular example.

3. Mobile location tracking

Compared to desktops and laptops, mobile devices such as smartphones and tablets can be a rich source of detailed location data. Mobile OSes can leverage a combination of methods to determine a user’s precise location, including, from generally more accurate to less accurate: Global Positioning System (“GPS”); WiFi network and Bluetooth analysis; and cell tower trilateration.

- a. **GPS.** Nearly all mobile devices include a GPS receiver, which is turned on by default. The user can turn off these receivers for access by apps, but some location tracking is always enabled for access by emergency services. GPS data can be accurate within a radius of about five to eight meters, although smartphone

¹⁴ “Android Software Development Kit License Agreement,” *Android*, (<http://developer.android.com/sdk/terms.html>).

¹⁵ Gartner Inc. forecasts that “by 2017, 94.5% of downloads will be for free apps.” See “Gartner Says Less Than 0.01 Percent of Consumer Mobile Apps Will Be Considered a Financial Success by Their Developers Through 2018,” *Gartner*, Jan. 13, 2014, (<http://www.gartner.com/newsroom/id/2648515>). See also Sarah Perez “Paid Apps on the Decline: 90% of iOS Apps Are Free, Up From 80-84% During 2010-2012, says Flurry,” *TechCrunch*, July 18, 2013, (<http://techcrunch.com/2013/07/18/paid-apps-on-the-decline-90-of-ios-apps-are-free-up-from-80-84-during-2010-2012-says-flurry/>), stating that from 2010-2012 80 to 84 percent of iOS apps were free, but by 2013, 90 percent of iOS apps in Flurry’s (mobile analytics company) network were free.

¹⁶ For example, Evernote, a once note-taking mobile application, has now expanded to numerous devices such as desktop computers and tablets, “Getting Started with Evernote,” *Evernote*, (<https://evernote.com/evernote/guide/windows/?var=3>).

location may be somewhat less accurate than in standalone GPS devices.¹⁷ The mobile OS collects location history and offers location-based advertising that uses GPS and other location data.¹⁸

Mobile apps often collect location data from the device's GPS receiver, thus gaining the benefit of the relative accuracy of GPS location. For iOS, an app must receive consent for location data during each interaction, even if the user agreed to location data during the installation and the GPS service remains enabled. Once the mobile app provider obtains location data, it may be used for advertising or shared with a third party. For Android, an app requests access to location data during installation, and after that historically has prompted for location services to be enabled if turned off during app use, but it has not otherwise reminded the user that location data is being collected.¹⁹ In Android 6.0 Marshmallow ("Android M"), however, a user may edit the permissions granted to an app at any time, including access to location data.²⁰

- b. **WiFi, Bluetooth, and other methods.** WiFi and Bluetooth receivers can provide location information to the operating systems, applications, and advertisers on mobile devices. When enabled, the WiFi receiver searches continuously for available WiFi networks in range. Every router in range of the device returns the available network name (or SSID) and the router's device identifier (or MAC Address). Companies can purchase access to commercial databases mapping each of these known MAC addresses to their location, including both public and home WiFi networks.²¹ Such databases, combined with the MAC addresses viewed by a mobile device, in effect transmit location data back to the database subscribers. The ability to subscribe to such databases essentially "democratizes" location information, so a wide range of companies can participate in location-based services and advertising.

Mobile OS providers use information about free and home WiFi networks to enhance the accuracy of mobile device location services. Android also allows mobile app providers to obtain access to the device's WiFi connections, which can be used to identify the location history of the device.²²

In addition to serving as an important source of location information, the location of known WiFi networks assists in cross-device tracking, as discussed further in Chapter 9. For example, a company can recognize that three mobile devices are consistently using the same WiFi network in the evenings. It can then look up the location of the WiFi network using one of the commercial or open source databases. This gives advertisers the ability to link household information to the three mobile devices.

¹⁷ "GPS SPS receivers provide better than 3.5 meter horizontal accuracy" and "Many users enhance the basis [civilian GPS service] with local or regional augmentations. Such systems boost civilian GPS accuracy." "GPS Accuracy," *GPS.gov*, (<http://www.gps.gov/systems/gps/performance/accuracy/>); "Mobile Location Accuracy Data Sources," *YP Mobile Labs*, 2014, (http://national.yp.com/downloads/YP_Mobile_Labs_Location_Accuracy_Data_Sources.pdf) ("A recent study found that of the mobile ad impressions that included latitude-longitude data, fewer than 34% were correct to within 100 meters of a user.").

¹⁸ Greg Rose, "Understanding Location Based Advertising," *AcquisioBlog*, Feb. 3, 2012, (<http://acquisio.com/blog/digital-marketing-beginners/understanding-location-based-advertising/>).

¹⁹ The newest version of Android, called Marshmallow, makes real-time prompts, more similar to iOS; <https://www.androidpit.com/android-m-release-date-news-features-name>. Many users, however, do not routinely update to the latest version of Android. See Chris Hoffman, "Why Your Android Phone Isn't Getting Operating System Updates and What You Can Do About It," *How-To Geek*, Nov. 11, 2012, (<http://www.howtogeek.com/129273/why-your-android-phone-isnt-getting-operating-system-updates-and-what-you-can-do-about-it/>).

²⁰ *Id.*

²¹ Robert McMillan, "After Google Incident, Data Collection Goes On," *PCWorld*, (http://www.pcworld.com/article/205062/After_Google_Incident_WiFi_Data_Collection_Goes_on.html). Stating that Apple, Google, Navizon, and Skyhook collect MAC addresses, which can be used to identify wireless routers. The WiFi data can then be linked with other data such as cell tower and GPS readings to get an idea of where the device is located.

²² "Android Software Development Kit License Agreement," *Android*, (<http://developer.android.com/sdk/terms.html>).

Bluetooth receivers and other mobile device sensors can similarly be used as a precise location tracking tool for mobile devices. New methods of location tracking by apps include the use of a device's magnetometer, LED lighting location tracking, and the use of sub-audible sound beacons.²³ By installing Bluetooth beacons in a store, for example, the owner of those beacons can track the movement of mobile devices with the Bluetooth receiver turned on.²⁴ This tracking can be precise enough to tell which products are on a shelf the device is currently in front of. Additionally, Bluetooth receivers have their own unique MAC addresses which can be recognized across multiple visits to a location. This information is not identified to a specific known user unless the user has downloaded an app that can link the beacon signal or MAC address to the user. A user can disable the device's Bluetooth receiver, making it invisible to these beacons, but it must do so in addition to disabling the general location service setting that controls GPS, WiFi, and cell tower data.²⁵

- c. **Cell tower trilateration.** Cell tower trilateration is a location technology traditionally open to mobile broadband providers, but it is generally not as accurate as GPS receivers or the newer WiFi and Bluetooth databases, and therefore is often referred to as "coarse" location data.²⁶ By knowing the cell towers closest to a mobile device, it is possible to calculate the current coarse location of the device through a process known as trilateration. This information is available to the ISP and the device's operating system, and it is used by mobile OSes to assist in determining location. Android also allows mobile app providers to obtain access to cell tower IDs, which can be used to identify the general location history of the device.²⁷ As with WiFi databases, a mobile OS provider or third party can purchase commercial databases mapping the coordinates of cell towers so they can be used to enhance location accuracy.²⁸ Cell tower trilateration is usually less accurate than GPS or WiFi location data, and it is typically accurate to within a ZIP code or neighborhood in urban areas and a greater distance outside of cities.²⁹ Our interviews indicate that GPS and WiFi location are used far more often than cell tower trilateration for mobile advertising purposes.³⁰

²³ See Dan Goodin, "Beware of Ads that Use Inaudible Sound to Link Your Phone, TV, Tablet, and PC," *Ars Technica*, Nov. 13, 2015, (<http://arstechnica.com/tech-policy/2015/11/beware-of-ads-that-use-inaudible-sound-to-link-your-phone-tv-tablet-and-pc/>). See also Jessica Leber, "Startup Uses a Smartphone Compass to Track People Indoors," *MIT Technology Review*, July 26, 2012, (<https://www.technologyreview.com/s/428494/startup-uses-a-smartphone-compass-to-track-people-indoors/>).

²⁴ "Retail & Location-Based Services," *Bluetooth*, (<https://www.bluetooth.com/marketing-branding/markets/retail-location-based-services>).

²⁵ For discussion of policy issues and a code of conduct concerning such beacons, see "About Smart Places," *Future of Privacy Forum*, (<https://fpf.org/issues/smart-places/>).

²⁶ See "Terminal Location (LBS) FAQs," Verizon, (http://developer.verizon.com/content/vdc/en/verizon-tools-apis/verizon_apis/network-api/faqs/napi_sup_lbs.html#2) (stating that cell tower trilateration has an average accuracy of less than 150 m); Christine Bauer, "On the (In-)Accuracy of GPS Measures of Smartphones: A Study of Running Tracking Applications," *11th International Conference on Advances in Mobile Computing & Multimedia*, 2013, p. 336

(https://www.researchgate.net/publication/259190145_On_the_InAccuracy_of_GPS_Measures_of_Smartphones_A_Study_of_Running_Tracking_Applications) (stating that GPS generally provides accuracy of 10 meters 95% of the time); "Mobile Location Accuracy: Data Sources," YP, 2013, (https://www.researchgate.net/publication/259190145_On_the_In-Accuracy_of_GPS_Measures_of_Smartphones_A_Study_of_Running_Tracking_Applications).

²⁷ "Android Software Development Kit License Agreement," *Android*, (<http://developer.android.com/sdk/terms.html>).

²⁸ "Antenna Structure Registration," *Federal Communications Commission*, (<http://wireless.fcc.gov/antenna/index.htm?job=home>).

²⁹ Rob Friedman, "All Geotargeting Methods Are Not Created Equal," *StreetFight*, Aug. 17, 2012, (<http://streetfightmag.com/2012/08/17/why-not-all-geotargeting-methods-are-created-equal>); "Demystifying Location Data Accuracy: The New Frontier and Biggest Mobile Opportunity," Mobile Marketing Association, (<http://www.mmaglobal.com/files/documents/location-data-accuracy-v3.pdf>).

³⁰ "Mobile Location Accuracy: Data Sources," YP, 2013, (https://www.researchgate.net/publication/259190145_On_the_In-Accuracy_of_GPS_Measures_of_Smartphones_A_Study_of_Running_Tracking_Applications).

C. Recent OS Changes

We turn to examination of three OS market leaders on traditional and mobile devices (Apple's iOS, Google's Android, and Windows 10), including the ability to track a user's history of Internet activity. The discussion highlights three themes: (1) The prominence of app stores led by OS developers, such as Apple's App Store or Google's Play Store; (2) the importance of personal assistants such as Apple's Siri, Google's Google Now, and Microsoft's Cortana; and (3) the shift for each OS toward gathering broader categories of data, with the details depending on the company's business model.

The last point requires some explanation at the start. There have been important changes in how OS developers earn revenue. Historically, Microsoft as the leading OS developer generated revenue through sale of software licenses. Computer manufacturers purchased an OS license for each machine and passed both the license and the cost along to the end consumer of the product. Individual users could buy the OS themselves to install on a new device or to upgrade an older one. Businesses could purchase packages of licenses for their multiple computers, sometimes including advanced support.

The OS today, however, is less likely to be purchased via direct sale. For both mobile and non-mobile devices, the OS is being treated as an ongoing service rather than an independent purchase. The growth of mobile, where there is no direct-sale model, has put pressure on OS developers to modify their business models. Each of the following three major OS developers examined have taken a slightly different approach to generating ongoing revenue, which in turn impacts the degree to which their OS explicitly seeks to provide advertising data on its users.

1. Apple³¹

Apple has long emphasized an integrated approach to hardware and software. Dating back to its early home computers, the OS and other software were primarily features to enhance the value of the hardware purchase. Apple had relatively little reason to collect detailed information about the individual's computing use, apart from telemetry and product improvement. With the iPhone and iPad, app developers had reasons to seek and receive more detailed information about the user. More recently, the digital assistant Siri has become a compelling new reason for far more granular and continuous collection of information by Apple about the individual's actions. Siri is turned on by default, so the emerging default is an operating system that collects detailed data about the user in order to act as a "personal assistant" for the user, and potentially for advertising-related uses.

Apple has at least two related goals in developing iOS.³² One is to add value to the iPhone and iPad as devices, spurring hardware sales. Another is to create a flourishing ecosystem for apps, sold through its mobile AppStore, where users purchase additional software for their Apple devices; Apple receives a portion of the revenue from all AppStore transactions. The lack of apps has been a major obstacle to other companies in the mobile smartphone space, such as RIM/Blackberry and Microsoft. To attract and retain app developers, Apple must make it profitable for developers to write for iOS. For mobile app developers, advertising delivery and tracking is a key feature.

³¹ This section primarily examines Apple's mobile operating system iOS. Apple's non-mobile operating system, OS X, largely tracks with the capabilities of iOS, but some additional OS X specific information is included here. Like with its mobile devices, Apple's traditional device business focuses on generating revenue from hardware sales. Apple's OS X serves the same purpose for its laptops and desktops that iOS serves for its mobile devices: the OS's design and function are a part of the hardware's value proposition and are exclusive to Apple's hardware. OS X also operates similarly to iOS, with users able to purchase and access apps from the AppStore and to make use of enhanced Spotlight search capability. OS X does not track any additional data or offer additional predictive search or advertising based on user behavior patterns.

³² iOS represents 13.9 percent of the total mobile OS marketplace and 64 percent of the overall enterprise market share in the U.S. "Smartphone OS Market Share, 2015 Q2," *International Data Corporation*, (<http://www.idc.com/prodserv/smartphone-os-market-share.jsp>). "Mobility Index Report Q2 2015," *Good Technology*, Aug. 2015, (<https://media.good.com/documents/mobility-index-report-q2-2015.pdf>).

Many developers sell their software for little to no money, and instead they rely on the serving and tracking of advertisements to supplement their revenue.³³

As discussed earlier in this Chapter, Apple assigns an advertising ID to each Apple mobile device, assisting a wide range of advertising-related activities by app developers and those who purchase information from such developers. Along with these incentives for app developers to write for the AppStore, Apple continues to develop its own iAds efforts³⁴ and leverages its role as an OS for a range of advertising purposes, including using a user's iTunes account information and purchases/downloads for iAd targeting.³⁵ Apple offers location-based advertising as part of iAd targeting. Users can opt-out of interest-based iAds on various devices and turn off location-based iAds.³⁶

The development of the digital assistant Siri is transforming the collection of user data by iOS. Siri now means that granular and continuous collection of user data is a standard practice in the OS. Siri is designed to allow users to make intuitive voice commands to their iPhone, simplifying tasks and enhancing the device's usability. The user typically activates Siri with a voice command, which leads to greater data collection and use for two main reasons.³⁷ First, the device's microphone is set to always listen; otherwise, it could not respond to a user's voice command without some other physical interaction. Second, the device learns a user's voice during a setup session to prevent a different user from accidentally activating someone else's phone.³⁸ Once activated, Siri is capable of a large range of activities.

Siri operates by learning cues from all Siri users, and also by learning behavioral patterns for individual users. Pieces of any individual Siri command are recorded and sent to cloud databases to assist Siri in understanding both the correct content (what the actual words said were) and the correct context (what the user meant or

³³ "Of 110 popular, free apps available for both Android and iOS tested, 73 percent of Android apps shared personal information with third parties, and 47 percent of iOS apps shared geo-coordinates and other location data with third parties." Jinyan Zang, et al, "Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps," *Journal of Technology Science*, Oct. 30, 2015, (<http://techscience.org/a/2015103001/>).

³⁴ Apple's iAds is set by default to offer interest-based custom ads and location-based ads to iOS 9 users. "About Privacy and Location Services for iOS 8 and iOS 9," *Apple Support*, Sep. 16, 2015, (<https://support.apple.com/en-in/HT203033>) ("Location-Based iAds: Your iPhone will send your location, including travel speed and direction, to Apple to provide you with geographically relevant iAds."); "Opt Out of Interest-Based Ads from iAd," *Apple Support*, Sep. 15, 2015, (<https://support.apple.com/en-vn/HT202074>) (To give you the best advertising experience, iAd provides ads based on your interests."). Both of these programs provide opt-out mechanisms for users, though Apple recommends against opting-out as these features provide "the best advertising experience." While iAd only accounts for 2.6 percent of total ad revenue, these features can provide significant advertising for iOS users. Lara O'Reilly, "Execs Tell Us the Writing Had Been on the Wall for Apple's Big Advertising Experiment iAd for Some Time," *Business Insider*, Jan. 14, 2016, (http://www.businessinsider.com/why-apple-is-pulling-direct-sales-support-for-iad-2016-1?utm_campaign=Feed%3A+typepad%2Falleyinsider%2Fsilicon_alley_insider+%28Silicon+Alley+Insider%29&utm_medium=feed&utm_source=feedburner) (citing an eMarketer chart titled "Net US Mobile Ad Revenue Share, by Company, 2014-2017").

³⁵ Apple states that it doesn't use web browsing information or email content for ads. Apple has, however, launched retargeting for iAds for apps. Apple says iAds does not get data from Maps, Siri, iMessage, Homekit, Health, call history, or any iCloud service like Contacts or Mail. "Apple's Commitment to Your Privacy," *Apple*, (<http://www.apple.com/privacy/>).

³⁶ "Opt Out of Interest-Based Ads from iAd," *Apple*, Sep. 15, 2015, (<https://support.apple.com/en-us/HT202074>).

³⁷ Siri can be voice activated at any time on an iPhone 6 and 6s, and any time the device is charging for older models. Lisa Eadicicchio, "8 Cool New Things You Can Do With Siri," *Business Insider*, Sep. 22, 2015, (<http://www.businessinsider.com/new-siri-features-ios-9-2015-9?op=1>).

³⁸ Kevin Tofel, "Apple Adds Individual Voice Recognition to 'Hey Siri' in iOS 9," *ZDNet*, Sep. 11, 2015, (www.zdnet.com/article/apple-adds-individual-voice-recognition-to-hey-siri-in-ios-9/!).

wanted to do) of the voice command.³⁹ These commands can handle many direct tasks, such as calling a specific phone contact, composing a text message, or playing a local music file.⁴⁰ In iOS 9, Siri can now also pull data from all the apps on a phone and monitor behavior to anticipate future commands.⁴¹ Siri is also now capable of setting contextual reminders. While reading an email, the command “remind me about this in the car” will create a notification alert the next time the phone registers as in a vehicle, bringing back up the email that was being viewed when the command was issued.⁴²

Siri is capable of integrating location data in processing commands.⁴³ Siri can offer suggestions for local restaurants based on current location, search photos based on location tags in the photo metadata, and generally consider location information when it might help process a voice command. Combined with other metadata, such as the date a picture was taken, Siri can perform a powerful search across the device’s data.

Siri is particularly powerful in combination with Apple’s Spotlight search tool.⁴⁴ Siri and Spotlight combine to offer a mix of detailed, tailored, and predictive searches across an Apple device’s data. The search function will auto-populate with not just frequent contacts, but those a user has upcoming appointments with; suggested apps based on those a user uses most frequently at certain times, locations, or in other contexts; and news and places of interest near the device’s current location.⁴⁵ Siri adds functionality in part through its collection and processing of large amounts of user commands. That data processing allows Siri and Spotlight to use natural language processing to compute naturally worded questions into executable commands. Saying “show me all the emails I ignored last week” will now return a list of ignored email from the past seven days, rather than at best a web-based search of the phrase spoken.

A personal assistant such as Siri assists with countless useful tasks. It also means that information collected within the OS becomes far more closely linked to advertising activities than was historically the case. Location-based responses are one example – “suggest a restaurant I would like in this neighborhood” or “tell me what sales are happening at the local mall.” To the extent there was a historical separation between data collected by an OS and the advertising world, that separation appears to be eroding.

2. Google Android

For its Android OS, Google relies more heavily on advertising than does the Apple iOS. While iOS is available only on Apple’s proprietary hardware, Android is installed on multiple different types of hardware from multiple vendors. Android is part of the overall Google business plan to generate a large part of corporate revenue through various advertising services.⁴⁶ These services leverage data collected across many different Google services to build better advertising delivery and tracking, including through Android.

³⁹ “By using Siri or Dictation, you agree and consent to Apple’s and its subsidiaries’ and agents’ transmission, collection, maintenance, processing, and use of this information, including your voice input and User Data, to provide and improve Siri, Dictation, and dictation functionality in other Apple products and services.” Apple’s iOS software license agreement. (<http://images.apple.com/legal/sla/docs/iOS8.pdf>).

⁴⁰ “What’s new in iOS,” *Apple*, (<https://www.apple.com/ios/whats-new/>).

⁴¹ *Id.*

⁴² *Id.*

⁴³ Apple’s iOS Software License Agreement, *Apple*, (<http://images.apple.com/legal/sla/docs/iOS8.pdf>).

⁴⁴ Notably, Apple is also making the Spotlight API publicly available for third-party apps, which will allow them to perform the same deep-linking search across all of a device’s apps and data.

⁴⁵ “What’s New in iOS,” *Apple*, (<https://www.apple.com/ios/whats-new/>).

⁴⁶ Google advertising services include, but are not limited to, AdWords, AdWords Express, AdSense, AdMob, and DoubleClick. “Google Business Solutions,” *Google*, (<https://www.google.com/services/>).

The greater reliance on advertising similarly occurs for its app Play Store. Compared with Apple, more Google Play Store apps offer free versions that rely solely on advertising and/or in-app purchases to provide revenue for the developer.⁴⁷ These developers use ad tracking and delivery both to generate direct revenue and to provide an added value to higher-cost but advertising-free versions of their apps. Like Apple, Google also receives a portion of the revenue from Google Play Store transactions, and therefore it has an incentive to make it easier for developers to leverage advertising in their apps, particularly for free apps.

Android seeks to gain a foothold in many different types of devices.⁴⁸ To that end, contrary to Apple, Android allows any smartphone hardware manufacturer to use and customize the Android OS on their devices. Google has offered the Google Nexus line of phones as a demonstration of the “stock” Android OS experience.⁴⁹ In comparison, other manufacturers have always been allowed to include their own customizations to the stock Android OS.⁵⁰ These customizations can include different visual themes or layouts, additional pre-installed software,⁵¹ and manufacturer-specific software or hardware features.⁵² Google’s flagship phones therefore served as an example of the stock Android capabilities and design, not to encourage exclusive purchase of Google hardware, but to demonstrate Android design and features to encourage consumers to purchase any Android phone.

Google runs a Display Ad network that allows its own apps and third-party apps to track users using the Advertising ID provided by Google Play Store.⁵³ When using Google’s own apps, Google may also use information from a user’s Google profile to target ads.⁵⁴ When using Google’s mobile advertising services, advertisers can target users based on some Google Play Store information such as account information, apps a user has downloaded, and how often those apps are used.⁵⁵ Google provides an opt-out of targeting for signed-in users on Google sites and on third-party sites beyond Google, as well as an opt-out for signed-out users receiving Google Ads across the web.⁵⁶ The Google opt-out applies to Google Search, Gmail, Maps, and YouTube.⁵⁷ The opt-out bars retargeting,

⁴⁷ Yoni Heisler, “Why Developers STILL prefer iOS over Android,” *BGR*, April 15, 2015, (<http://bgr.com/2015/04/15/ios-vs-android-developers-revenue-apps/>).

⁴⁸ Google’s strategy has led to an 82.8 percent market share of the total mobile OS marketplace. “Smartphone OS Market Share, 2015 Q2,” *International Data Corporation*, (<http://www.idc.com/prodserv/smartphone-os-market-share.jsp>).

⁴⁹ Google Nexus was once manufactured by Motorola while it was a subsidiary of Google. However, since selling off Motorola, Google has returned to contracting out the development and manufacture of the Google Nexus line, with the two versions of the most current model being manufactured by LG and Huawei. Sarah Silbert, “Nexus Phones Will Never See Huge Sales – But Here’s Why They Don’t Need To,” *Fortune*, Sep. 30, 2015, (<http://fortune.com/2015/09/30/google-nexus-smartphones-about-innovation-not-sales/>).

⁵⁰ Sarah Mitroff, “Android Skins: What You Should Know,” *CNet*, Nov. 10, 2014, (<http://www.cnet.com/news/android-interface-guide/>).

⁵¹ Note that the developers of these pre-installed apps are included as a contractual arrangement with the device manufacturer in return for some undisclosed sum. Seth Porges, “It’s Time to Put an End to Smartphone Bloatware,” *Forbes*, July 30, 2015, (<http://www.forbes.com/sites/sethporges/2015/07/30/its-time-to-put-an-end-to-smartphone-bloatware/>).

⁵² E.g., Motorola’s “Moto” software, including soft touch alerts. “Software and Apps by Motorola,” *Motorola*, (<http://www.motorola.com/us/Software-and-Apps-by-Motorola/consolidated-apps-page.html>).

⁵³ Jim Edwards, “Google’s New ‘Advertising ID’ Is Now Live and Tracking Android Phones—This is What it Looks Like,” *Business Insider*, Jan. 27, 2014, (<http://www.businessinsider.com/googles-new-advertising-id-is-now-live-and-tracking-new-android-phones-this-is-what-it-looks-like-2014-1>); Greg Sterling, “Google Replacing ‘Android ID’ With ‘Advertising ID’ Similar to Apple’s IDFA,” *Marketing Land*, Oct. 31, 2013, (<http://marketingland.com/google-replacing-android-id-with-advertising-id-similar-to-apples-idfa-63636>).

⁵⁴ “About Google Ads,” *Google*, (<https://support.google.com/ads/answer/1634057?hl=en>).

⁵⁵ Larry Kim, “4 Ways Google Just Made Mobile App Advertising More Awesome,” *Inc.*, April 22, 2014, (<http://www.inc.com/larry-kim/4-ways-google-just-made-mobile-app-advertising-more-awesome.html>).

⁵⁶ Jack Wallen, “Pro tip: How to Opt Out of Interest-Based Ads on Your Android Phone,” *Tech Republic*, Aug. 7, 2014, (<http://www.techrepublic.com/article/pro-tip-how-to-opt-out-of-interest-based-ads-on-your-android-phone/>); “Opt Out,” *Google*, (<https://support.google.com/ads/answer/2662922?hl=en>).

⁵⁷ “Opt Out,” *Google*, (<https://support.google.com/ads/answer/2662922?hl=en>).

ads based on visits to “other web sites,” and “demographic details on your computer’s browser”⁵⁸ The Google ads across the web target based on web surfing interests, inferred language, age, and gender.

Some user controls exist. Generally, Google depends on an app having access to user location to enable location-related ad targeting. Android apps can require location by default, although updates in Android “Marshmallow” allow users to turn off access to location app-by-app through their Android settings. Apps can continue getting location in the background even when not “open” or being used. Older versions of Android require users to click-through consent for an app to use location, but they do not allow users to turn off location per app. As a result, Google apps running on Android have location if they declare the permission and if location is turned on for the device. Google provides a location history dashboard where users can choose to delete all history displayed, or just the data for a specific day or location.

Although Android can be obtained open source and used without any data being sent to Google, device manufacturers who want to use the Google Play Store to make apps available, or who want the latest Android updates, must agree to place various Google apps or services on the phone.

Similar to Siri, Android offers a digital assistant service called Google Now (“GN”), which leverages voice recognition and behavioral patterns to predict user behavior. While the technical details differ, GN operates functionally similar to Siri, with user commands being sent off the device for processing to ensure accurate transcription and contextual understanding, so that the returned action is responsive to the user’s intent. The company demonstrated many of these features in a May 28, 2015, keynote address.⁵⁹ Google’s investment in machine learning has led to a strong contextual awareness for GN, allowing the service to understand that the voice command “what’s his real name?” is referring to the performing artist of the song currently being played.⁶⁰ GN analyzes the data on the screen, even if it is a third-party app, and processes the content, the context, and any additional useful data points like time of day, weather, and location. Google also archives all of a user’s voice commands, which is linked to the user account and reviewable online.⁶¹ While users can opt-out of having their account and identity associated with their GN voice data, they cannot opt-out from the overall collecting and archiving of commands for continued improvement of the service. GN also provides predictive suggestions based on observed user patterns and contextual evidence. For instance, adding a calendar event based on an email receipt for concert tickets, or restaurant recommendations for the user’s preferred cuisine and price range based on current location, which restaurants are currently open, and how busy the restaurants currently are.

3. Microsoft Windows 10

Microsoft’s recent OS activities show the OS’s transformation over time in the role of collecting user data. Historically, Microsoft earned substantial revenue through the sale of licenses for Microsoft Windows. Microsoft collected limited information, notably in the event of a system failure, through means such as the Windows Error Reporting introduced in Windows XP.

Windows 10 marks a change in the overall business strategy for Microsoft’s OS division, moving away from the direct sale of license model and toward an ongoing, services-based model. Windows 10 applies to a range of

⁵⁸ *Id.*

⁵⁹ Chris Welch, “Google Now Gets Smarter With ‘Now on Tap’ And Ability To Work Inside Apps,” *The Verge*, May 28, 2015, (<http://www.theverge.com/2015/5/28/8677147/google-now-on-tap-announced>).

⁶⁰ *Id.*

⁶¹ Brad Reed, “Google Records and Stores Everything You Ask Google Now – Here’s How to Find and Delete It,” *BGR*, Oct. 13, 2015, (<http://bgr.com/2015/10/13/google-now-tips-and-tricks-audio-history/>).

devices, including personal computers, tablets, and smartphones. When Windows 10 was first available, current users of older versions of the OS were offered a free upgrade to its basic “Windows 10 Home” version, hoping to encourage wider early adoption of the newest OS and its updated features, including the Microsoft digital assistant Cortana. Combined with the move toward a subscription model for Microsoft Office 365, Windows 10 is representative of a decision to encourage broader adoption of the Windows platform while focusing on other ways to generate revenue besides direct OS sales. Licensing revenue continues: Windows 10 has a premium version available through upgrades, and Windows 10 Professional, geared towards businesses, also continues to generate direct sale licensing revenue.

Windows 10 features a number of new privacy settings, which by default allow Windows 10 to collect data on contacts, calendar details, text and touch input, location, and more.⁶² Some of this data is collected for traditional telemetry and crash reporting purposes. Microsoft’s privacy statement notes that these reports may “unintentionally contain personal information,” but that in those instances any data collected will not be used to “identify, contact, or target advertising.”⁶³ While previous editions of Windows allowed users to opt-out completely from sharing these telemetry and crash reports, Windows 10 is designed so that all users share “basic” information in these reports.

Windows 10 now incorporates the digital assistant program Cortana, as well as the new browser Microsoft Edge, which combine to create contextual, predictive search functionality. Any information a user enters in the Edge navigation bar is automatically sent to Bing, which in turn offers search recommendations that update with each character entered.⁶⁴

Microsoft also offers interest-based advertising on an opt-out model. These ads may be targeted based on search queries, interests, favorites, usage data, or current location. Microsoft does not, however, use the content of emails, chats, video calls, voice mail, documents, photos, or other personal files for advertising targeting.⁶⁵

These predictive capabilities integrate with Microsoft’s digital assistant Cortana, included in Windows 10. Like Siri and Google Now, Cortana offers a powerful speech recognition-based tool to enhance OS usability. Cortana collects device location, calendar data, Windows app data, email and text message contents, call records, contacts, and other data to offer tailored and predictive suggestions to improve user productivity. For instance, users can sign into Facebook using Cortana, which grants Microsoft access to a user’s Facebook information that can be used to generate personalized recommendations.⁶⁶ Cortana collects user voice commands for processing as a part of its input personalization features, and to make sure it responds properly to the user’s voice.

Windows 10 also by default includes targeted advertising based on collected user data.⁶⁷ Location, search query, browsing history, interests, favorites, and historical usage and behavioral data are all used by this service, including data collected through use of the Cortana digital assistant. Microsoft’s own relatively detailed privacy policy illustrates the granular issue of how much OS data collection will be used, and in what settings, for targeted

⁶² *Id.*

⁶³ “Microsoft Privacy Statement,” *Microsoft*, (last modified Oct., 2015), (<https://www.microsoft.com/en-us/privacystatement/default.aspx>).

⁶⁴ Note that Microsoft does not collect data on private browsing sessions in Microsoft Edge. *Id.*

⁶⁵ *Id.*

⁶⁶ “Microsoft Privacy Statement,” *Microsoft*, Oct. 2015, (<https://www.microsoft.com/en-us/privacystatement/default.aspx?Componentid=pspMainHowWeUsePersonalDataModule&View=Description>).

⁶⁷ Users may opt-out of targeted advertising. Microsoft does not use the content of emails, chats, video calls, voicemails, documents, photos, or other personal files for targeted advertising. “Microsoft Privacy Statement,” *Microsoft*, Oct. 2015, (<https://www.microsoft.com/en-us/privacystatement/default.aspx?Componentid=pspMainHowWeUsePersonalDataModule&View=Description>).

advertisements. While this policy does specifically note that Microsoft does not examine the content of a user's interpersonal communications or personal files, all other data (including that collected by Cortana) can be used for Microsoft's targeted advertising. Hypothetically, a user may be watching a movie trailer and ask Cortana "Who is that?" Cortana can then return search results for an actor in the movie trailer, including that actor's filmography. If the actor is also in another movie that is currently playing at the user's local movie theatre, Microsoft's targeted advertising could serve an ad for tickets to the second movie. However, if the user were to send an email to a friend saying "Did you see this movie trailer? I can't wait to see more of that actor!" that data could not be used to return the same targeted advertisement.

4. How changes in mobile affect advertisers

Mobile advertising is an increasingly large portion of the total digital advertising ecosystem, and it is projected to continue growing in the near future. In 2014, mobile ad spending accounted for 38.5 percent of total digital ad spending.⁶⁸ Mobile ad spending was projected to have reached a total of \$30.45 billion by the end of 2015; a significant increase from the \$19.15 billion spent in 2014.⁶⁹ That \$30.45 billion spent on mobile advertising will account for 52.4 percent of total digital ad spending, and is projected to reach 69.9 percent by 2019.⁷⁰

The different identifiers available from a mobile device, combined with location data, allow advertisers to more accurately target ads and collect a broader set of data on devices. For example, a mobile advertising company can coordinate information across multiple apps, gaining insight into a greater percentage of the device's Internet activity history. Similarly, these trackers can give information on how often an app is accessed, how long it is used, as well as when a user clicks on a mobile ad. This rich data set is valuable to data analysts, as well as to ad publishers seeking to increase user interaction.

Mobile device tracking is valuable, as well, because mobile devices are rarely shared among users. Unlike a traditional desktop, which may be shared among the various members of a family, mobile devices are rarely shared. Therefore, once the device is identified, by default, so is the device's user.

Location information also allows for very accurate advertising targeting. For example, an advertising company can now not only buy impressions for an 18- to 24-year-old male interested in football, but it can also serve ads to any device at a specific football game whose user fits those demographic specifications. Location data can also augment search-based advertisements, highlighting ads for nearby businesses during a relevant search. Bluetooth beaconing can even be used to target ads to users in a specific part of a store, or who spend a predetermined amount of time in front of the product. A targeted coupon at that moment may be enough to result in a purchase, which is only available due to the location data from the user's device.

D. How ISPs Compare to Operating Systems

Compared to the full access of an OS, no ISP has the capability to see as much of any single user's Internet history and other device activity. The data traditionally collected by the OS through telemetry services is sent via an encrypted connection to the OS developer, rendering it unreadable to the transmitting ISP. The ISP also has no ability to directly access telemetry data from the OS, as it lacks permission and software access to those sensors and their data. OS providers also have the ability to use data about app purchase and usage for advertising, and use unique advertising device identifiers to correlate activity across apps for advertising purposes.

⁶⁸ "Mobile to Account for More than Half of Digital Ad Spending in 2015," *eMarketer*, Sep. 1, 2015, (<http://www.emarketer.com/Article/Mobile-Account-More-than-Half-of-Digital-Ad-Spending-2015/1012930>).

⁶⁹ *Id.* Note that the dollars spent in the U.S. on mobile advertising have seen significant growth in each of the past two years, from \$10.67 billion in 2013, to \$19.15 billion in 2014, and a projected \$30.45 billion by the end of 2015.

⁷⁰ *Id.*

As discussed in this Chapter, each OS has implemented its own digital assistant. Each of these digital assistants is increasingly integrated into the OS's advertising ecosystem, enabling more granular targeting of advertisements based on usage and search patterns. On mobile devices, these digital assistants are also increasingly able to pull data from the usage of different apps on a single smartphone or tablet, resulting in a larger compilation of the device's history of Internet activity available for advertising services.

In comparison, as discussed in Chapter 1, the ISP faces technical blockades against access to much user Internet activity. Users increasingly jump among ISPs and use multiple devices. Encryption is becoming far more prevalent, including data relating to personal assistants such as Siri. In short, where the OS maintains a privileged position on any device, no ISP possesses the technical capability to perform the same level of observation or tracking of what takes place on the device.

In response, ISPs do have two potential advantages in relation to the mobile advertising ecosystem, location and known identity, but each provides less advantage than one might suspect. For location, ISPs have access to cell tower triangulation for location data. As discussed in this Chapter, however, mobile OSes and mobile apps also have access to that data. Mobile OS and mobile apps also have access to additional sources of location data, including GPS and WiFi, which are far more granular and often more widely accessible than cellphone trilateration. In urban areas, mobile providers report trilateration accuracy to a ZIP code or neighborhood, considerably less precise than a common GPS range of 10 feet. The location of known WiFi MAC addresses is similarly accurate to GPS, and there is growing use of Bluetooth beacon and other accurate data. Multiple interviews reported that the other location information is commercially far more significant today than cell tower-based location.

Another possible commercial advantage is that mobile broadband providers typically receive their subscribers' identity information, and thus could map that identity to tracking information from the device. This level of insight, however, is far from unique in the mobile ecosystem. For both iOS and Android, for example, the initial setup of a phone asks for the user's Google or Apple account, which auto-populates the user's identity. Each of these accounts is usually also tied to payment information, such as for app purchases, allowing for a high degree of accuracy in the identification. Other companies similarly gain identity information, including social networks that are used as sources of authentication due to the granular information they obtain about users.⁷¹

In short, ISPs' potential advantages on location and identity turn out, upon inspection, to be no better and often worse than the information available to other players in the ecosystem.

⁷¹ In addition, mobile broadband providers typically offer family plans. Although the mobile provider may know account identity information for the person who set up the account, sometimes known as the account holder, the provider may not know the identity of the individuals who use the different devices that are part of the same account. For example, Jane Smith may have an account with a mobile broadband provider that includes smartphones for herself, her spouse, and two children, as well as three tablets. In this case, although the mobile ISP accepts payment from the user and has an ongoing account, the accuracy of any identity associated with a particular device is less certain.



Interest-Based Advertising ("IBA") and Tracking

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 6: Interest-Based Advertising (“IBA”) and Tracking

This Chapter discusses the major ecosystem of online advertising that historically has been based on cookies rather than logged-in content of search, social networks, webmail, or other contexts. It provides new Diagrams for both the non-mobile and mobile online advertising ecosystem.

This Chapter begins with an introduction to cookie-based online behavioral advertising (“OBA”).¹ In this Chapter, we use the term to cover the set of technologies and practices in the modern advertising ecosystem, with emphasis on URLs – the host names and full URLs visited, rather than the actual content viewed. We include in our discussion the broader concept of “Interest-Based Advertising” (“IBA”), which the Network Advertising Initiative (“NAI”) defines as the delivery of advertisements based on users’ interests (i.e., interest categories) and users’ previous interactions with the advertiser(s) serving the advertisement.² This definition includes the increasingly common practice of adding demographic and other offline information to cookie-based and other online information.³

In order to make the complexity of this ecosystem more understandable, Diagram 6-A introduces the key actors involved in placing an online advertisement: publishers, supply side platforms (“SSPs”), ad exchanges, demand side platforms (“DSPs”), and marketers. Diagrams 6-B and 6-C provide what we believe are useful new graphics for how advertisements are bought and sold for cookie-based and mobile advertising. The Chapter next more briefly references techniques including cookie syncing and retargeting that together enable players in the IBA ecosystem to have visibility about much of a user’s URLs, content, and overall Internet activity.

The Chapter concludes with a discussion of the relationship of Internet Service Providers (“ISPs”) to this IBA ecosystem. The leading players in the IBA ecosystem have very substantial visibility into users’ Internet activity, including access to content through collection of URLs visited, and that visibility is not dependent on data collection by ISPs. ISPs have not historically been top-tier participants in IBA. In terms of the overall themes of this Working Paper, IBA is one significant context for data collection about users’ Internet activity, and thus a source of the cross-context tracking and cross-device tracking discussed in later Chapters. Visibility about user activity from the IBA ecosystem, and its use in cross-context and cross-device tracking, comes primarily from non-ISP actors.

A. Introduction to the Online Advertising Ecosystem

The online advertising ecosystem is famous for its complexity. In our experience, the most cited diagram is referred to as the “Display LUMAscape.” The Display LUMAscape attempts to map out the online advertising ecosystem, grouping over 100 online advertising companies into 23 clusters.⁴ In Diagrams 6-A and 6-B, we provide a simpler and more conceptual approach.

¹ The Federal Trade Commission defined OBA in 2009 as “the tracking of a consumer’s online activities over time – including the searches the consumer has conducted, the web pages visited, and the content viewed – in order to deliver advertising targeted to the individual consumer’s interests.” “FTC Staff Report: Self-Regulatory Principles for Online Behavioral Advertising,” *Federal Trade Commission*, Feb. 2009, (<https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-staff-report-self-regulatory-principles-online-behavioral-advertising/p085400behavadreport.pdf>).

² The Network Advertising Initiative has adopted IBA terminology. See “Understanding Online Advertising: What Is It?” *Network Advertising Initiative (NAI)*, (<https://www.networkadvertising.org/understanding-online-advertising/what-is-it>).

³ IBA includes information from audience matching, location data, cross-context tracking, cross-device tracking, and other sources of offline data. Offline data can also be purchased by data brokers. This data is then used for advertising purposes.

⁴ Luma Partners, “Display LUMAscape,” (<http://www.lumapartners.com/lumascapes/display-ad-tech-lumascape/>).

Diagram 6-A describes five groups in the online advertising ecosystem. Understanding the ecosystem starts with two main groups, the “publishers” and “marketers.”

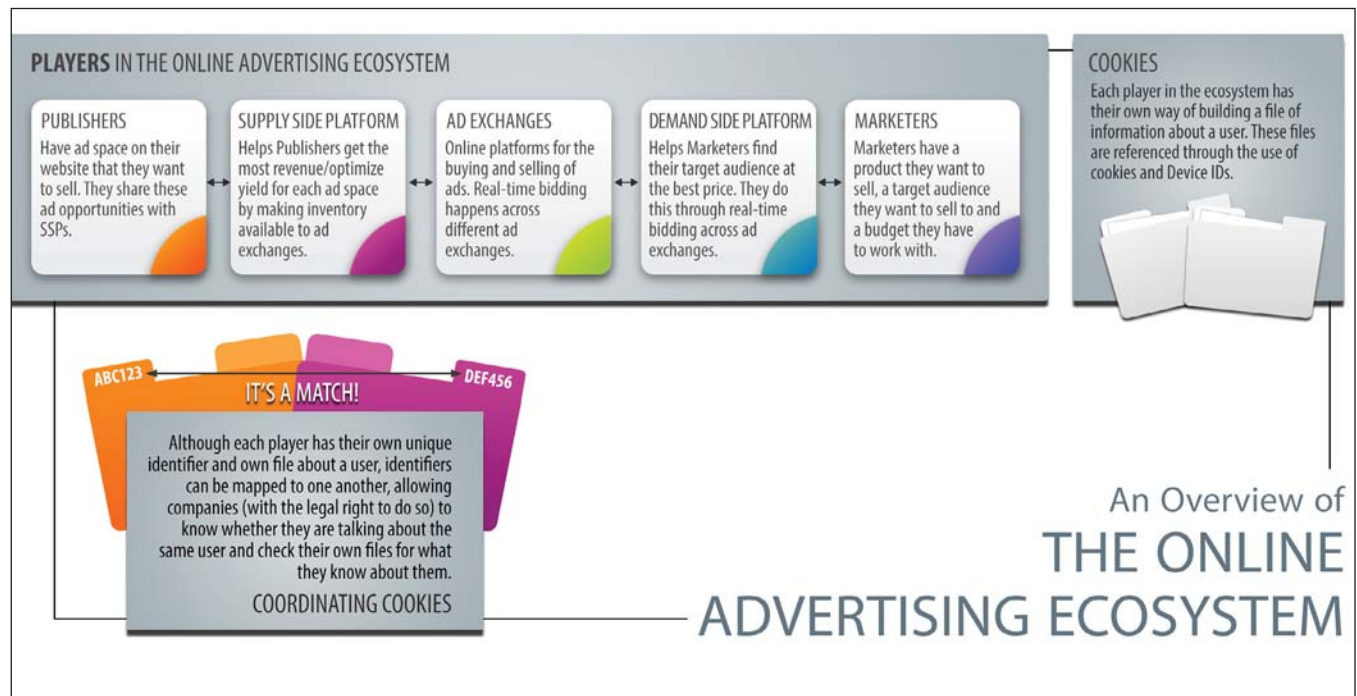


Diagram 6-A

- **Publishers.** Publishers are sellers of advertising space inventory. Publishers have a website – they publish content similar to the way a print newspaper or magazine publishes content. They have ad space on the website that they would like to sell. They are the supply-side for advertisements – they have a supply of places (an inventory) where advertisements can be displayed next to their content. In our example, the publisher, OnlineNews.com, has a vacant space for a display advertisement.
- **Marketers.** Marketers are buyers of advertising space inventory. Marketers have a product they want to sell (market). They create advertisements they want consumers to see, and they buy advertising space from publishers to do so. They are the demand-side for advertisements. In our example, ExampleShoes.com wants to buy ads that will be seen by female shoe lovers.

The overview Diagram 6-A shows three groups that match sellers of ads (publishers) with buyers of ads (marketers):

- **Supply-side platforms.** Supply-side platforms or “SSPs” are paid by publishers to get the most revenue for their supply of advertising space inventory. A publisher such as OnlineNews.com is an expert at writing interesting news stories, but not necessarily expert at selling advertisements. Therefore, publishers hire SSPs to maximize their inventory’s value.
- **Demand-side platforms.** Demand-side platforms or “DSPs” are paid by marketers to get the most value for their advertising purchases. A marketer such as ExampleShoes.com is expert at selling popular shoes but not necessarily expert at buying ad inventory. A DSP can help ExampleShoes.com find the best ad inventory for ExampleShoes.com, as well as the best price.
- **Ad exchanges.** Ad exchanges provide an online platform for matching the bids from DSPs with the asks from SSPs. Today, ad exchanges offer real-time bidding so sellers can get the highest available price and buyers can pay the lowest available price for an advertisement.

1. IBA example (non-mobile)

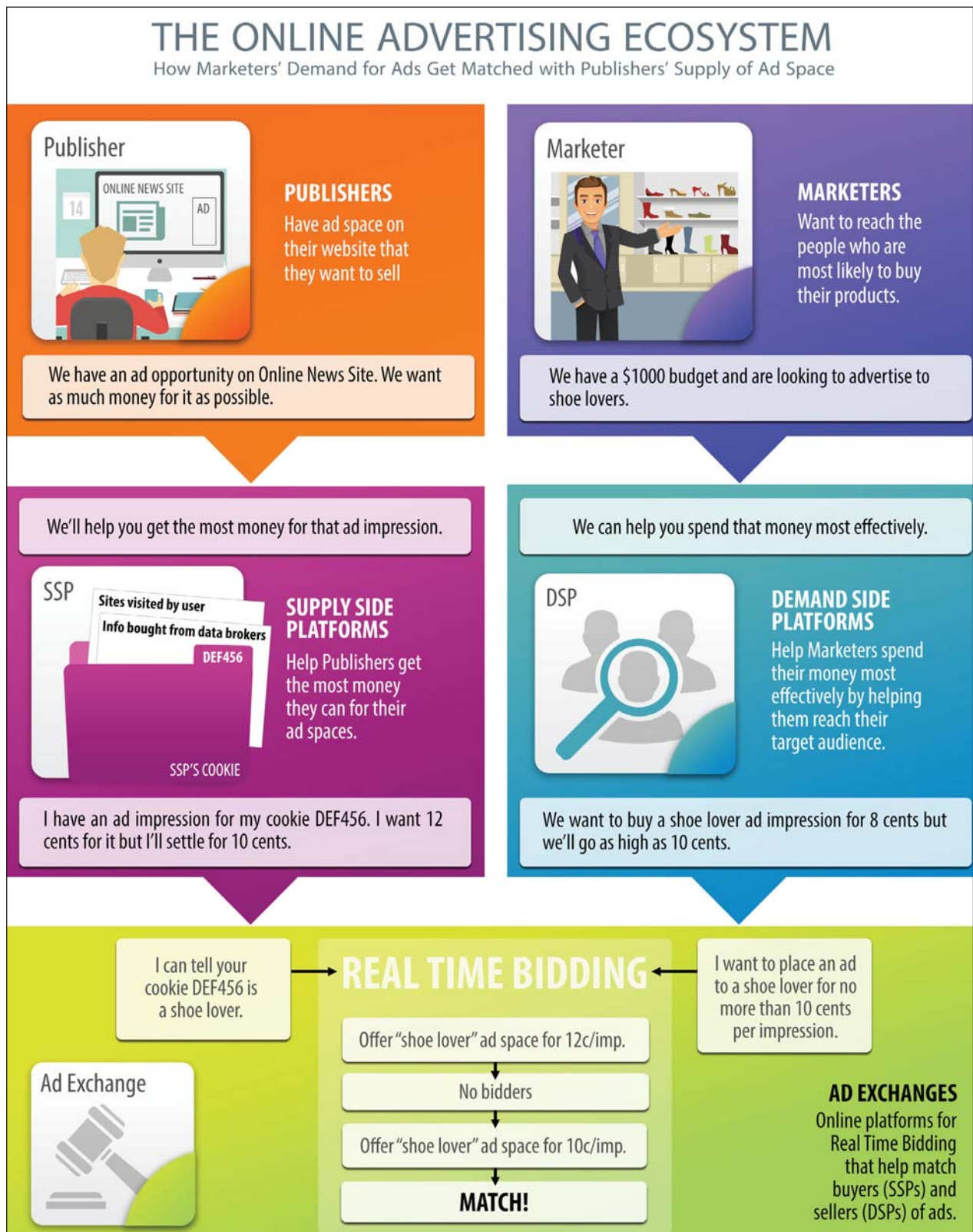


Diagram 6-B

Now we discuss Diagram 6-B's example of how Example Shoes' demand for advertising space is matched with OnlineNews.com's supply of ad inventory.

On the publisher end, our story starts when a user first visits OnlineNews.com. Online News hires an SSP to help sell its ad inventory.⁵ The SSP sets a cookie (Cookie DEF456), and that cookie generates a record of visits by the browser of a specific device, such as Firefox, on a particular laptop computer.⁶

Modern SSPs have similar relationships with numerous publishers. The next time that device (using that browser) visits one of the SSP's sites, information gets added to the cookie. Over time, the SSP develops a profile linked to that cookie, such as "shoe lover" in our example.

To summarize the supply side of the advertisement, Online News gets a visit from a device that the SSP has identified as a "shoe lover" (along with other interests linked to that cookie). The SSP, based on its profile linked to Cookie DEF456, seeks the best price for an ad impression. In our example, the SSP would like to sell the ad impression for 12 cents, but it is willing to settle for 10 cents if necessary to sell the advertisement.

On the marketer end, our story starts when Example Shoes assigns a budget of \$1,000 to buy online advertising for its products. Example Shoes hires a DSP to help spend that budget most effectively.⁷ The DSP has expertise in where it can buy advertisements at the best price and with the best match to users' interests. In our example, the DSP wants to buy a shoe lover ad impression for 8 cents but is willing to go as high as 10 cents.

We now understand the buy side (the DSP acting on behalf of Example Shoes) and the sell side (the SSP acting on behalf of Online News). An advertising exchange brings buyers and sellers together. There is real-time bidding for ad impressions, measured in milliseconds. In our simplified example, the SSP opens with an offer to sell the ad impression for 12 cents. There are no bidders, so the SSP drops the price to 10 cents. At that price, the DSP is willing to bid, and there is a match – the Example Shoes ad is shown to the person whose computer has the cookie DEF456.

⁵ SSP technology is sold by companies including AOL, AppNexus, Google, OpenX, PubMatic, Right Media, and Rubicon Project. Jack Marshall, "WTF is a Supply-side Platform," *Digiday*, Jan. 22, 2014, (<http://digiday.com/platforms/wtf-supply-side-platform/>).

⁶ The SSP's cookie in the text is commonly referred to as a "third-party cookie," in contrast to a "first-party cookie" that Online News might set itself, and which is not in the diagram. Note, information about the websites visited might be stored in the cookie or, more often, in a server gathering information for the company that sets the third-party cookie. For one introduction to cookies, See Joanna Geary, "Tracking the Trackers: What are Cookies? An Introduction to Web Tracking," *The Guardian*, April 12, 2012, (<http://www.theguardian.com/technology/2012/apr/23/cookies-and-web-tracking-intro>).

⁷ DSPs include DataXu, Google's Invite Media, MediaMath, Turn, and X+1. Jack Marshall, "WTF is a Demand-side Platform," *Digiday*, Jan. 8, 2014, (<http://digiday.com/platforms/wtf-demand-side-platform/>).

2. IBA example (mobile)

We next turn to the differences for advertisements shown on mobile devices. On the demand side, the system works essentially the same – Example Shoes wants to buy advertisements, and its DSP helps it do so efficiently. The supply side is somewhat different, however. As discussed in Chapter 5, there are two major differences: Mobile devices often use different identifiers besides cookies,⁸ and they can also use location data.

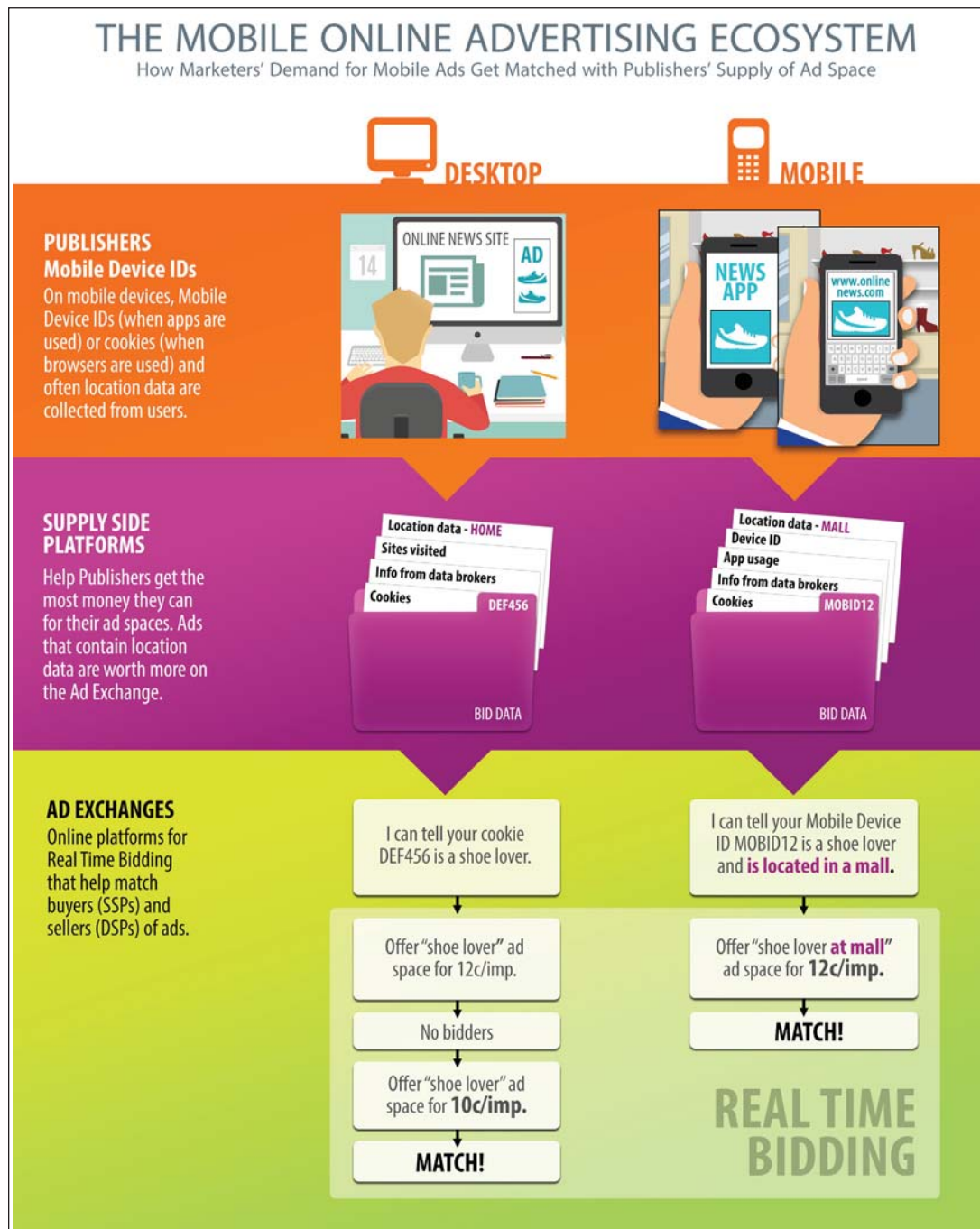


Diagram 6-C

⁸ Cookies are often less effective on mobile devices than on non-mobile devices. Cookies are associated with a browser, and browsing is less prominent in the app-heavy mobile ecosystem. Apps often connect to a content server in order to populate the mobile device with web content and advertisements, and apps do not support cookies.

Diagram 6-C illustrates the mobile equivalent of the example discussed above about cookies. The supply side is different essentially because the SSP has different information about the user. One difference is that there is a unique “advertising ID” for each device, and the SSP may gain insights about a user’s interests from app usage and other information linked to that ID when the user has not limited or reset that ID. A second difference is that certain information is uniquely available in the mobile context, including more specific location information and app usage. For cookie-based advertisements for traditional devices such as desktops, there may be some degree of geo-location. By contrast, as discussed in Chapter 5, GPS, WiFi hotspot, and other location information is far more detailed and prevalent for mobile devices. For a mobile device, the SSP essentially has all of the information that was available about a desktop user (except for some detail of cookie information), but it often has this additional location data, app usage data, and other information. For both mobile and non-mobile, SSPs and DSPs sometimes enhance their profiles (and retain them) with offline data, cross-device data, and other information sources.

The effect of the location and other information is indicated at the bottom of Diagram 6-C, in the Ad Exchange. For the desktop example, the advertisement is sold for 10 cents per impression, as discussed above. For the mobile example, the SSP and DSP may consider the device’s current location or historical location data in bidding on an available ad impression. If the device is currently next to an Example Shoes store, that may be a reason to serve an Example Shoes ad. Similarly, if the device’s profile includes a location history indicating they often visit an Example Shoes competitor, Example Shoes may choose to serve an ad for a special sale or coupon. In the Ad Exchange, the mobile advertisement may go for a higher price, such as 12 cents per impression for an ad targeted at the user close to an Example Shoes store while in a shopping mall.

B. Features of the IBA Ecosystem

Matching advertising to users based on users’ interests is one of the main features of IBA. The way information flows in IBA can offer attribution and accountability about what advertisements are in fact delivered. When designing enhanced and new IBA technologies, network security is also an important consideration. Given the scale of online advertising, nearing \$50 billion a year,⁹ the complex interactions of the IBA ecosystem can become a tempting target for malicious actors. Network security is thus an essential component to serving effective ads at a good price. Another feature of IBA has been to comply with the Digital Advertising Alliance (“DAA”) self-regulatory principles,¹⁰ the Network Advertising Initiative Code of Conduct,¹¹ and any other relevant laws and corporate obligations.¹²

There are multiple techniques in today’s IBA ecosystem to keep track of user activity for the purposes just mentioned, including advertising, attribution, and network security. We list these here, with footnotes for further reference, but we do not attempt to explain each of these techniques in detail:

⁹ Christopher Heine, “Mobile Ads Skyrocketed 76% in 2014, Making Digital Advertising a \$50 Billion Business,” *Ad Week*, April 22, 2015, (<http://www.adweek.com/news/technology/mobile-ads-skyrocketed-76-2014-making-digital-advertising-50-billion-business-164222>).

¹⁰ The DAA is a non-profit organization that collaborates with businesses, public policy groups, and public officials to establish and enforce “responsible privacy practices across industry for relevant digital advertising, providing consumers with enhanced transparency and control.” In order to achieve this mission in the context of OBA, the DAA has developed self-regulatory principles that are set forth in the Self-Regulatory Principles for Online Behavioral Advertising. See “Self-Regulatory Principles for Online Behavioral Advertising,” *Digital Advertising Alliance (DAA)*, July 2009, (<http://www.aboutads.info/resource/download/seven-principles-07-01-09.pdf>); and “About the Self-Regulatory Principles for Online Behavioral Advertising,” *Digital Advertising Alliance (DAA)*, (<http://www.aboutads.info/obaprinciples>).

¹¹ See “About the NAI,” Network Advertising Initiative (NAI), (<https://www.networkadvertising.org/about-nai>); NAI Code of Conduct, *Network Advertising Initiative (NAI)*, (<https://www.networkadvertising.org/code-enforcement>).

¹² This is not intended to be an exhaustive list of all of the various principles, codes of conducts, pledges, or promises that pertain to OBA. Rather, these are intended to be examples.

1. Cookies.¹³
2. Cookie syncing.¹⁴
3. Cookie resyncing.¹⁵
4. Digital fingerprinting.¹⁶
5. Referer headers and pixels.¹⁷
6. Retargeting.¹⁸
7. Cookie respawning.¹⁹

C. Limited Role of ISPs in the IBA Ecosystem

This Chapter has provided discussion of the various players in the IBA ecosystem and how these players are able to use cookies and other technologies to collect demographic information on users, target advertisements, and track the success of advertisements served to users.

One consequence of the IBA ecosystem is that entities are often able to use their access to URLs to then find the content that corresponds to that URL – knowing the detailed URL allows the entity in effect to click on the link and see the content. Entities that do this can then often associate that URL and content with other contexts and devices, giving these entities even higher visibility into a user’s Internet activity.

ISPs have historically not been the principal players in the online ecosystem documented in this Chapter. As explained in Chapter 1, ISP visibility into users’ Internet activity is technologically limited. The greatest insights from IBA come from non-ISPs, who often are leaders as well at cross-context and cross-device tracking.

¹³ See “Fact and Fiction: The Truth About Browser Cookies,” *Lifehacker*, Feb. 2, 2010, (<http://lifehacker.com/5461114/fact-and-fiction-the-truth-about-browser-cookies>).

¹⁴ See “SSP to DSP Cookie Syncing Explained,” *Ad Ops Insider*, May 1, 2011, (<http://www.adopsinsider.com/ad-exchanges/cookie-syncing/>).

¹⁵ *Id.*

¹⁶ See Adam Tanner, “The Web Cookie Is Dying. Here’s the Creepier Technology that Comes Next,” *Forbes*, June 17, 2013, (<http://www.forbes.com/sites/adamtanner/2013/06/17/the-web-cookie-is-dying-heres-the-creepier-technology-that-comes-next/#7538c093e45c>).

¹⁷ See Phil Gross, “Cookies, Tags and Pixels: Tracking Customer Engagement,” *Visual IQ Newsletter*, Vol. 2, Issue 9, Sep. 2012, (<http://www.visualiq.com/resources/marketing-attribution-newsletter-articles/cookies-tags-and-pixels-tracking-customer-engagement>).

¹⁸ See “What is Retargeting?” *AdRoll*, (<https://www.adroll.com/getting-started/retargeting>).

¹⁹ See “Cookie Respawning,” *Techopedia*, (<https://www.techopedia.com/definition/18555/cookie-respawning>).



Browsers, Internet Video, and E-commerce

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 7: Browsers, Internet Video, and E-commerce

The total number of contexts where advertising data collection and use is relevant is not static. The emergence of new platforms and technologies, new uses for existing technologies, and changes in user interest can all create new contexts for data and reduce the impact of existing ones. The types of data collected, and the general uses for that data are similar across contexts despite the different types or amounts of data those contexts may incorporate.

To that end, this Chapter takes a snapshot view of three additional contexts and the advertising data available in each of them. First, this Chapter discusses desktop and mobile web browsers, and how several leading web browsers (e.g., Apple Safari, Google Chrome, Microsoft Edge, and Mozilla Firefox) collect and use data. Second, this Chapter examines Internet video distribution, including how different online video platforms can collect the content of videos consumers watch. Finally, this Chapter explores e-commerce and the types of data available to and used by online retailers for both first party and third party advertising.

This Chapter focuses on each of these additional contexts separately, and on how that context alone impacts advertising. As explained in Chapter 8, cross-context data tracking is an important method for increasing the value of access to data in any single context. This Chapter focuses on the data flows for these three contexts to better understand how any business offering a relevant service might acquire and use advertising data.

A. Browsers¹

Web browsers are an integral part of the modern Internet. For non-mobile devices, web browsers remain the dominant means of access to Internet content. Even on mobile devices, minutes spent on web browsers continue to increase despite a larger fraction of user time spent in mobile apps.² Consequently, a user's web browser sees a large fraction of her total Internet activity. All of the major browsers now support HTTPS by default. Even for HTTPS traffic that is encrypted, a web browser has technical access to both the full URLs a user visits and the specific content of those URLs.

¹ The features discussed in this section are true for both mobile and non-mobile versions of these web browsers, and as such no specific discussion is made of mobile web browsing. For a discussion of how mobile software can collect different types of data compared to non-mobile devices, see Chapter 5: How Mobile is Transforming Operating Systems.

² Minutes spent in mobile browsing in 2015 increased 53 percent from minutes spent in 2013, despite users spending only about 8 percent of their total time on mobile devices in web browsers. See, "The 2015 U.S. Mobile App Report," *Comscore*, Sept. 22, 2015, (<http://www.comscore.com/Insights/Presentations-and-Whitepapers/2015/The-2015-US-Mobile-App-Report?>).

This section will examine a variety of browser technologies that rely on the collection and use of URL and content histories for advertising, analytics, and other purposes. Most browsers have privacy policies that explain their personal data practices.³

1. Telemetry

Telemetry reporting is typically used to measure performance usage information from user devices to better optimize how browser software runs. However, this information is necessarily collected during use of the software and transmitted back to the developer, which leaves the technical possibility that telemetry reporting can be used to collect data on the URLs a user visits or the content a user views. Browsers typically offer consumers a choice as to whether they want to share telemetry data.

When Chrome's "opt-in" telemetry feature is enabled, Chrome creates a persistent identifying token for the device and sends that token along with the telemetry reports to avoid duplicate reports and increase accuracy.⁴ These reports may contain URL and personal information depending on what the browser was doing at the time of the crash being reported.⁵ Likewise, in Firefox, users must enable telemetry reporting, so that only users who are comfortable sharing data do so.⁶ Firefox says that it collects "non-personal" information for telemetry.⁷ For Safari's telemetry feature, Apple indicates it collects "anonymous technical data," but only with its users' explicit consent.⁸ Microsoft too seeks users' consent for telemetry related to the use of its services.⁹ Microsoft's policy

³ For example, the Mozilla Firefox privacy policy states that it only uses information acquired for purposes for which the user has given permission, and that the main purpose is to improve Mozilla products and services. "Mozilla Privacy," *Mozilla*, Apr. 15, 2014, (<https://www.mozilla.org/en-US/privacy/>). Similarly, the policy only permits the sharing of that information with third parties for limited purposes: Mozilla Firefox will share user data with third parties when they have asked for and receive consent; when releasing anonymized data to further open web initiatives; and for processing or providing services with contractual agreements that those third parties will abide by Mozilla's standards for data handling. *Id.* According to Google Chrome's privacy notice, the information that Google receives from Chrome is processed in order to operate and improve Chrome and other Google services. "Google Chrome Privacy Notice," *Chrome*, Sept. 1, 2015 (<https://www.google.com/chrome/browser/privacy/>). The policy indicates that Chrome may share with third parties certain aggregated, non-personal information, but that Chrome tries to avoid sending information that personally identifies the user. *Id.* Microsoft's privacy policy, which applies generally to all Microsoft services, indicates that the company uses personal data to provide services; communicate with the user; and make ads relevant. "Microsoft Privacy Statement," *Microsoft*, Jan. 2016 (<https://www.microsoft.com/en-us/privacystatement/>). The policy indicates that Microsoft will share personal data with the user's consent and as necessary to complete any transaction or provide a service the user has requested or authorized. *Id.*; see also "Microsoft Edge and privacy: FAQ," *Microsoft*, (<http://windows.microsoft.com/en-001/windows-10/edge-privacy-faq>). Apple's privacy policy, which applies to all of Apple's services, indicates that Apple uses the personal information it collects, among other purposes, to update users to improve products, services, content and advertising, and for loss prevention and anti-fraud purposes; to verify identity, assist with identification of users, and to determine appropriate services; to send important notices to users. "Apple Privacy Policy," *Apple*, Feb. 1, 2016 (<http://www.apple.com/privacy/privacy-policy/>). The policy indicates that Apple will share personal information only to provide or improve its products, services and advertising. *Id.*

⁴ Google Chrome Privacy Whitepaper," *Chrome*, Dec. 4, 2015, (<https://www.google.com/chrome/browser/privacy/whitepaper.html>)

⁵ When enabled, this automatic telemetry reporting also reports some amount of data related to general user activity; however, that data is anonymized and randomized according to principles of differential privacy to protect the data from being able to infer any single user's activity. Úlfar Erlingsson, "Learning Statistics with Privacy, aided by the Flip of a Coin," *Google Research Blog*, Oct. 30, 2014, (<https://googleresearch.blogspot.de/2014/10/learning-statistics-with-privacy-aided.html>).

⁶ "Share telemetry data with Mozilla to help improve Firefox," *Firefox Help* (<https://support.mozilla.org/en-US/kb/share-telemetry-data-mozilla-help-improve-firefox?redirectlocale=en-US&redirectslug=send-performance-data-improve-firefox>).

⁷ *Id.*

⁸ "Choose Whether to Share Diagnostic Data," *Apple*, (<http://www.apple.com/privacy/manage-your-privacy/>).

⁹ "Microsoft Privacy Statement," *Microsoft*, Jan. 2016, (<https://www.microsoft.com/en-us/privacystatement/>). Microsoft offers several different telemetry levels associated with Windows and does not allow Windows users to opt out of its Basic level; this level does not appear to include reporting specifically related to browser use. *Id.*

statement indicates that diagnostic and usage data is transmitted to Microsoft and stored with one or more unique identifiers that can help Microsoft recognize an individual user on individual devices and understand the device's services issues and use patterns.¹⁰

2. Private browsing

Most browsers offer some form of protected browsing mode where browser history and cookies from these protected sessions are not saved locally.¹¹ For example, Chrome's private browsing mode, "incognito mode," does not store basic browsing history information such as URLs, cached page text, IP addresses of pages linked, or transmit any pre-existing cookies to sites that users visit.¹² When a user activates Chrome's incognito mode, cookies are temporarily stored and transmitted to websites, and deleted when the user closes the browser or all open incognito windows.¹³ For Edge's InPrivate mode, users' browsing information, such as cookies, history, or temporary files, similarly are not saved on the user's device after the user's browsing session has ended.¹⁴ Firefox offers Private Browsing with Tracking Protection that actively blocks advertisements, trackers, and social share buttons.¹⁵ This browsing mode also shows users where the blocked assets on a web page are, allowing them to better understand which websites that they visit use these trackers.¹⁶ Safari offers Private Browsing, which does not store the webpages users visit, downloaded items, or changes to users' cookies or other website data.¹⁷

3. Integration of search and other functionality

Browsers offered by companies that offer other products can integrate those other products into the browser. For example, the "omnibox" is Google's term for the combined web address and search bar used in Chrome. Unless disabled, any terms typed into this bar will generate predictions of what a user is searching for through the default search engine set in the browser. This feature sends the user's IP address and Google cookie to Google to tailor these predictive results. When the user also has the Chrome default search engine set to Google Search, Chrome will provide suggestions without user input into the omnibox, based on the user's overall browsing history. When using Google Search, the logs of these requests are stored temporarily, and any selected suggestion sends back to Google the original search query, the option the user selected and the position of the selection. This data is used to help increase the effectiveness of the feature, and is logged temporarily as part of Google Search's ongoing database.¹⁸ Google offers a logged-in capability that allows data collection across the wide range of Google services, including Gmail, Google Maps, and YouTube.

¹⁰ *Id.*

¹¹ Third parties can still track user activity during these sessions, but the cookies used will close when the browser closes, and will not be persistent.

¹² "Google Chrome Privacy Notice," *Google*, Sept. 1, 2015 (<https://www.google.com/chrome/browser/privacy/>)

¹³ *Id.*

¹⁴ "Microsoft Edge and Privacy: FAQ," *Microsoft*, (<http://windows.microsoft.com/en-us/windows-10/edge-privacy-faq>).

¹⁵ "Tracking Protection in Private Browsing," *Firefox Help* (<https://support.mozilla.org/en-US/kb/tracking-protection-pbm>).

¹⁶ "Firefox Tracking Protection," *Mozilla*, (<https://www.mozilla.org/en-US/firefox/42.0/tracking-protection/start/>).

¹⁷ "Use Private Browsing windows," *Apple*, Sept. 30, 2015 (https://support.apple.com/kb/PH21413?viewlocale=en_US&locale=en_US).

¹⁸ Omnibox suggestion request logs are retained for two weeks, after which 2% of the data is randomly chosen, anonymized, and retained for future analysis. URLs are not included in the 2% retained sample, but are collected in the original logs. All other data collected is logged and anonymized under the same rules as data collected through Google.com's Search Engine. "Google Chrome Privacy Whitepaper," *Chrome*, Dec. 4, 2015, (<https://www.google.com/chrome/browser/privacy/whitepaper.html>).

Other browsers offer similar features. Edge automatically sends the information a user types into the browser address bar to Bing and offers search recommendations as each character is typed, even if another default search provider is selected. In addition, Edge aggregates browsing history data to predict which pages users are likely to browse next and proactively loads those pages in the background. Browsing data collected in connection with these features is used in the aggregate and can be turned off at any time.¹⁹

Likewise, when Safari Suggestions is enabled, search queries, selected Safari Suggestions, and related usage data will be sent to Apple. User location information and information about subscription services may also be sent to Apple. According to Apple, location, search queries, and usage information collected in this manner and sent to Apple will only be used by Apple to make Safari Suggestions more relevant and to improve other Apple products and services.²⁰

Firefox leverages user machines instead of the cloud to handle features like built-in search. When a user performs a search through the Firefox browser bar, as opposed to navigating to the search engine's website to perform the search, the user's device handles the request for data and returns the results. This design means that no Firefox cloud server acts as intermediary for these activities, which reduces Firefox' technical visibility into user behavior.

4. Form autofill

Most browsers offer a form autofill function that learns and stores the data that users commonly submit to forms.²¹ Typically, this feature collects the field names, structure of form, and, if the form is submitted, the data the user entered. Later, when the user encounters new forms, the browser will then suggest the same information it has learned from previous forms to automatically fill out the form for the user. Autofill can also learn payment card data. In Chrome, Firefox, and Edge, this feature is turned on by default.

5. Data for advertising

Browsers offer advertising-enabling features in different ways. For example, Firefox Sync allows users to sign in to allow users to share browser data like cookies, saved passwords, and bookmarks across their devices.²² To do so, users must provide an email address and create a username, which gives Mozilla the ability to accurately map multiple devices to a single user. Firefox only uses that data to provide the underlying Sync service, and encrypts user data transferred between devices so that it is never stored on Firefox servers in plaintext.²³

Chrome also offers a logged-in sync capability for users with multiple devices. When a user signs in to Chrome on a device with their Google account, her browser history and settings are harmonized across the other devices the user has signed in to.²⁴ All data collected and sent through this feature is subject to the general Google privacy policy, which allows its use for advertising purposes.²⁵

¹⁹ "Microsoft Privacy Statement," *Microsoft*, Jan. 2016, (<https://www.microsoft.com/en-us/privacystatement/>).

²⁰ "Safari Suggestions and privacy," *Apple* (<http://help.apple.com/safari/mac/9.0/#/sfrid73436cb>).

²¹ "Google Chrome Privacy Whitepaper," *Chrome*, Dec. 4, 2015, (<https://www.google.com/chrome/browser/privacy/whitepaper.html>).

²² "Mozilla Services Privacy Policy," *Mozilla*, (<https://services.mozilla.com/privacy-policy/>).

²³ *Id.*

²⁴ "Google Chrome Privacy Whitepaper," *Chrome*, Dec. 4, 2015, (<https://www.google.com/chrome/browser/privacy/whitepaper.html>).

²⁵ *Id.*

Likewise, Safari users can log into iCloud to access their browsing history across multiple devices, such as iPhone, iPad, iPod touch, and Mac.²⁶ Under Apple's general privacy policy, Apple can use the information it collects for advertising.²⁷

Microsoft too allows users to sign into an account to provide for personalized experiences across multiple devices.²⁸ For Microsoft, synced information includes, among other things, browser history, favorites and websites the user has opened.²⁹ Because the information collected and sent through this feature is subject to Microsoft's general privacy statement, it can be used for advertising purposes.

6. ISP lack of visibility of browser activity

In comparing the insights available to Internet Service Providers ("ISPs") and browser providers, the initial point is that ISPs do not provide their own web browsers, and therefore do not collect comparable data to the entities that develop browser software. The second point, as discussed in Chapter 1 and elsewhere, is that all of the major browsers now support HTTPS by default. ISPs today thus get no access to content or detailed URLs related to browsing – ISPs are blocked from seeing the search terms, telemetry, log-ins and other data that browsers collect and transmit to their developers. In contrast to ISPs, browser providers retain the capability to access content or detailed URLs related to browsing, even when HTTPS is used.

B. Internet Video

Video content accounts for a majority of the Internet traffic in North America.³⁰ In 2014, Internet video traffic accounted for 64 percent of all consumer Internet traffic, and is projected to grow to 80 percent of all consumer Internet traffic in 2019.³¹ While the Internet is by no means the exclusive means of distributing video content to consumers, as is true throughout this Working Paper, our focus is on what information can be gathered about users' Internet activity when viewing video content online, rather than through video content delivered as traditional pay TV services or offered as an extension of such services.³² Internet video may be consumed through direct website browsing or through video applications viewed on: traditional laptops and desktops; mobile devices; dedicated video streaming devices;³³ multi-use streaming devices;³⁴ or smart televisions.

Online video content can therefore pass through the products and/or services of numerous software providers, hardware providers, operating system developers, and online services. These entities all have differing levels

²⁶ "Set Up and Use iCloud Tabs," *Apple*, Dec. 12, 2015, (<https://support.apple.com/en-us/HT202530>).

²⁷ "Apple Privacy Policy," *Apple*, Feb. 1, 2016 (<http://www.apple.com/privacy/privacy-policy/>).

²⁸ "Microsoft Privacy Statement," *Microsoft*, Jan. 2016, (<https://www.microsoft.com/en-us/privacystatement/>).

²⁹ *Id.*

³⁰ See "Cisco Visual Networking Index: Forecasting and Methodology, 2014-2019 Working Paper," Cisco, May 27, 2015, (https://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html); see also "Global Internet Phenomena: Africa, Middle East & North America," Sandvine, p.3, Dec. 2015 (reporting that, for peak period Internet traffic, 5 of the top 10 fixed access applications delivered video content, with Netflix accounting for 34.7% and YouTube accounting for 16.9% of the total peak traffic) (<https://www.sandvine.com/trends/global-internet-phenomena/>).

³¹ "Cisco Visual Networking Index: Forecasting and Methodology, 2014-2019 Working Paper," Cisco, May 27, 2015, (https://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html).

³² The focus on Internet activity is consistent with the scope of the FCC's Open Internet proceeding, which concerns rules for Broadband Internet Access Service, or "BIAS" providers.

³³ Examples of dedicated streaming devices include, but are not limited to, AppleTV, Amazon FireStick, Google ChromeCast, and Roku.

³⁴ Examples of multi-use streaming devices include, but are not limited to, Microsoft Xbox One, Sony PlayStation 4, TiVo set-top boxes, and certain cable provider-issued set-top boxes that connect to the Internet.

of visibility into a user's video content choices determined by both technical access and legal permission. This section will describe the available data flows for Internet video and the different levels of visibility for various entities that handle Internet video. The section briefly explains why ISPs have far from comprehensive or unique visibility into a user's Internet video content.

1. How Internet video services collect user data

When an Internet video connection is delivered through a browser, it can use the same technology as often used by web pages to follow user behavior. When users visit a site that hosts first-party video content, such as a news site with embedded video clips or programs, the site can use cookies and engage in the other IBA practices previously discussed.³⁵ When these videos also include, or are accompanied by, third-party video advertisements, the same chain of advertisers that are present in other third-party advertising online also receive tracking data.

When the party hosting the online video content uses a logged-in account, they gain the same insights as for other logged-in advertising scenarios. Services such as Amazon, Hulu, Netflix, and YouTube allow users to log in to an account for additional functionality.³⁶ Identifiable information linked to the account, such as payment card and demographic data, can be added to information derived from the content of video consumed. That data can also be collected over time and linked to the same account, allowing for a historical tracking of the types of content consumed and how often different types of video content are consumed.³⁷ When this login information is linked to another permanent login, such as a social network account, the other linked entity may also receive select tracking and content data.³⁸

For mobile and dedicated streaming devices, Internet video is delivered through dedicated apps rather than browsers. For the video content provider, tracking is the same whether the user is logged-in through a browser or through the content service's app. App usage information can be collected by browsers and related systems, such as the third party device hosting the app, although specific viewing history and video content may not be visible.³⁹ The video service provider can also encrypt the data connection to the app the same way it is encrypted in a browser.⁴⁰

³⁵ As with other Internet content, YouTube connections by are now encrypted by default. Google Online Security Blog, "Ads Take a Step Towards HTTPS Everywhere," *Google*, Apr. 17, 2015, (<https://googleonlinesecurity.blogspot.com/2015/04/ads-take-step-towards-https-everywhere.html>).

³⁶ See, e.g., "Amazon.com Help: Amazon.com Privacy Notice," *Amazon*, Mar. 3, 2014, ("You can choose not to provide us certain information, but then you might not be able to take advantage of many of our features") ([https://www.amazon.com/gp/help/customer/display.html/ref=footer_privacy?ie=UTF8&nodeId=468496#GUID-A2C397AB-68FE-4592-B4A2\)7550D73EEFD2__SECTION_A110DAC3F6BC4D5D9DD59797104B1E5](https://www.amazon.com/gp/help/customer/display.html/ref=footer_privacy?ie=UTF8&nodeId=468496#GUID-A2C397AB-68FE-4592-B4A2)7550D73EEFD2__SECTION_A110DAC3F6BC4D5D9DD59797104B1E5)), "Netflix – Watch TV Shows Online, Watch Movies Online," *Netflix*, Aug. 20, 2015, ("We use the information we collect to provide, analyze, administer, enhance and personalize our services and marketing efforts . . .") (<https://www.netflix.com/PrivacyPolicy>), "Privacy Policy," *Hulu*, Feb. 18, 2015, ("Our use of this information may include . . . customizing the Content (including advertising) you view [and] customizing recommendations . . .") (<https://www.hulu.com/privacy#UseOfInfoWeCollect>), See, e.g., "Privacy Policy – Privacy & Terms – Google," *Google*, Aug. 19, 2015, ("We collect information to provide better services to all of our users . . . [such as] which YouTube videos you might like") (<https://www.google.com/intl/en/policies/privacy>).

³⁷ Examples of this are sites that allow users to log-in with a Facebook, Twitter, or Google Account.

³⁸ See, e.g., "Privacy Policy," *Fox*, Oct. 7, 2015, ("By logging in with or connecting your FOX Services account with a social media service, you are authorizing us to share information we collect from and about you with the social media service provider . . .") (<https://www.fox.com/policy>).

³⁹ Integrated video search has recently been introduced to iOS9, in addition to Android. This permits the iOS or Android device user to search within multiple applications if the application developer has enabled that functionality. But it does not expose detailed video content usage details to the device, the ISP, or to the ISP ecosystem.

⁴⁰ As noted in Chapter 1, YouTube encrypts by default today, and Netflix is shifting to encryption by default.

2. How ISPs compare to Internet video providers

ISPs are not major providers of Internet video services and have limited access to Internet video content data. Video content is delivered in the same types of packets as any other online data, so the same technical restrictions on ISP visibility exist. When video content is sent over an encrypted HTTPS connection, as is increasingly the norm, the ISP may only determine that the user is receiving some video content from a specific site or application, but cannot see the encrypted content.⁴¹ When the user is connected through a VPN service or other third-party DNS look-up service, even the domain name, such as VideoWebsite.com, will be hidden from the ISP. Similar limits on ISP visibility apply when Internet video content is viewed in-app, rather than in-browser.⁴²

In short, ISPs, by virtue of their role of providing the last mile connection to the Internet, do not have privileged or unique visibility into a user's Internet video content or viewing practices.

C. E-commerce

E-commerce provides another valuable context for targeted advertising online. Online retail websites create often long-standing relationships with their customers, build a history of interactions, and thus gain insights into user behavior and purchasing decisions. Like their brick-and-mortar siblings, online retailers can use this data for their own internal marketing and business decisions, or where lawful can sell it to other parties in the advertising ecosystem to generate additional revenue. As such, it is important to examine how data flows work for e-commerce today, and how that data impacts the current advertising ecosystem.

1. How e-commerce advertising data is collected

When e-commerce websites sell goods and services and collect payment information from customers, they gain reliable and identified information that can be used for marketing purposes. These e-commerce sites collect marketing data in two major ways:

- 1) Identified information related to credit cards, names, billing and shipping addresses, and phone numbers; and
- 2) Supplementary data for these customers, often purchased from third party data brokers.

In relation to the discussion of interest-based advertising in Chapter 6, e-commerce companies are often marketers, purchasing advertisements about their products. They also generally are publishers of advertisements, and may provide first-party advertisements (about themselves), third-party advertisements (about others), or both on their sites.

⁴¹ Netflix, which as previously noted accounts for a third of peak consumer Internet traffic, has announced it will switch to HTTPS. See, Dan Goodin, "It Wasn't Easy, but Netflix Will Soon Use HTTPS to Secure Video Streams," *ArsTechnica*, Apr. 16, 2015, (<http://arstechnica.com/security/2015/04/it-wasnt-easy-but-netflix-will-soon-use-https-to-secure-video-streams/>).

⁴² Katie Benner and Conor Dougherty, "Publishers Straddle the Apple-Google, App-Web Divide," *N.Y. Times*, Oct. 18, 2015 (http://www.nytimes.com/2015/10/19/technology/publishers-straddle-the-apple-google-app-web-divide.html?_r=0).

i. Data provided during purchase

Online purchases transmit generally reliable and identified personal data to the seller's website, regardless of the method of payment.⁴³ A purchase made by credit or debit card, for example, typically provides the purchaser's card number, billing address, shipping address, email address, and phone number. Because this data is tied to a purchase, and to the purchasing card, it is more reliable than other information used for online advertising. Whereas a user registering for a free online service may give a false name or address, an online transaction will be rejected if the provided data does not match the registered data for the purchasing credit or debit card account.

Other methods of online payment provide similar levels of information. Direct transfers from a bank account, electronic checks, and PayPal transactions all include verifiable account numbers, the account holder's name, billing address, shipping address, and email address. The seller may also require the buyer's phone number and email, to allow for contact in the event of a problem with the transaction. With direct transfers or electronic checks, the information is identified and reliable, as the provided information must match the records on file for that account. The same is true of PayPal, which does not shield personal information from the seller, but rather makes it quicker for the buyer to make purchases online.⁴⁴

Seller websites also necessarily see the rest of the details associated with the buyer's purchase, such as which items they bought, reviews they have left, how frequently they purchase from the seller, wish list or registry items, and items they have saved in their online shopping carts for later purchase. While e-commerce sites may have differing technical implementation for these options, under usual practice they will know at least the identity of their buyer, and what that buyer has purchased.

ii. Appended data

An e-commerce site can use the data it collects during purchases as the basis for linking to offline activity and buying additional detailed data from data brokers.⁴⁵ If an e-commerce site has a customer named John Smith at a certain address they can then purchase more data to append to their existing profile of Mr. Smith. In addition to information that comes from an e-commerce site's own brick-and-mortar stores, appended data can include other purchase histories from other stores (both online and brick-and-mortar stores), credit history information, detailed demographic information, and more.⁴⁶ Once purchased, the e-commerce site can use that data to better target its own marketing campaigns, and to provide more insight into their ongoing communications with the user.⁴⁷

2. How e-commerce data affects advertising

The combination of data acquired from an e-commerce site's first-party traffic and third-party appended data enables better targeted marketing. Since e-commerce sites acquire reliable and identified data from payment

⁴³ This section focuses on credit card, debit card, and PayPal transactions, which accounted for 48%, 30%, and 12% of online purchases in 2014. Tamara E. Holmes, "Online payment statistics," *Nasdaq*, Mar. 11, 2015, (<http://www.nasdaq.com/article/online-payment-statistics-cm453646>). This Working Paper does not examine other types of payment transactions, such as direct transfers or Bitcoin.

⁴⁴ "Learn How to Pay and Buy Online," *PayPal*, (<https://www.paypal.com/us/webapps/mpp/pay-online>).

⁴⁵ "Appended" data is defined as data purchased from third party data brokers based on known identifiers for the purposes of enhancing a proprietary database.

⁴⁶ "Little Blue Book: A Buyer's Guide," *Oracle bluekai*, Dec. 2014, (https://docs.oracle.com/cloud/latest/marketingcs_gs/OMCDA/pdf/Misc/bluekai-little-blue-book.pdf).

⁴⁷ The purchase of appended data is generally subject to a license agreement, with the data broker maintaining ownership of the original data. See, e.g., "Data Use Agreement," *Acxiom*, (<https://www.myacxiompartner.com/u/dialogs/OrderLicenseTerms.aspx>), "Experian Online Data License Terms and Conditions," *Experian*, (<http://www.experian.com/small-business/legal-terms.html>).

information, appending means they can create even more reliable and identified data profiles for customers. These data profiles can be used to target both first-party and third-party marketing campaigns. First-party targeted marketing for e-commerce sites may take the form, for instance, of recommended purchases or prompts to chat with a sales associate about a special offer. The technique of targeted first-party marketing became familiar to many users when Amazon recommended titles based on previously purchased books, and this technique has not been a major target of privacy regulatory scrutiny to date.⁴⁸

E-commerce sites also use these data profiles to provide targeted third-party marketing. For example, ComfyShoes.com can make use of third-party retargeting services to purchase banner advertisements on other websites that advertise the products a user had searched for previously on ComfyShoes.com. Sites may offer discounts through these advertisements or other incentives for the consumer to return and complete the purchase of a product they have investigated earlier. For instance, if a user searches for a new pair of boots at ComfyShoes.com, and then leaves to read articles on WorldNews.com, Comfy Shoes might place an ad on WorldNews.com featuring the boots the user had looked at, in an attempt to retarget the user to come back to ComfyShoes.com. Third-party retargeting can also be used to encourage customers who have not made a purchase recently to return to the e-commerce site by offering special deals and advertising products a specific consumer is likely to buy based on their data profile. While the advertisements themselves are placed in a third-party context, the interaction is derived from data acquired during a first-party interaction, which seeks to re-establish that first-party relationship with the user to encourage a purchase.

The effectiveness of these data profiles increases with the frequency of purchases an individual consumer makes on a particular e-commerce site. E-commerce sites combine identified first-party relationships with data about those individual's demographics, to deliver more targeted goods, services, and features based on big data analytics. Extensive patterns of purchases yield more data to be analyzed related to an individual account, and can result in more effective recommendations and targeted marketing compared to one-time purchases.

E-commerce sites may also act as data brokers themselves, licensing their first-party marketing data to other entities in the advertising ecosystem. While there are competitive incentives to keep this type of customer data private, companies with robust e-commerce sites may still sell information related to demographic markers, likelihood of purchase of big ticket items, and other targeting markers for their customers as a way to generate additional revenue.⁴⁹

3. Why ISPs are not major e-commerce sites

ISPs have a small share of the overall market for e-commerce. ISPs offer first-party e-commerce interactions with users, such as ways for new customers to sign up for services and purchase or lease related equipment such as smartphones. ISPs do not gain information about the detailed and numerous transactions available to leading e-commerce sites.

⁴⁸ "Data Brokers: A Call for Transparency and Accountability," *Federal Trade Commission*, May 2014, (<https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>).

⁴⁹ Depending on the nature of an e-commerce site's business, it may be subject to one of the regulatory regimes in American law. For example, the Gramm-Leach-Bliley Act governs financial services firms that sell goods or services through their e-commerce sites, and HIPAA similarly governs health care providers and insurers. Therefore, different e-commerce sites may face different legal restrictions on how they can collect and use data depending on the nature of the business.

As explained in previous Chapters, ISPs also have limited insight into the activity of subscribers on e-commerce sites generally. Payment information is nearly always transmitted over encrypted connections, blocking access by third parties including the ISP. Similarly, increasingly prevalent HTTPS connections for e-commerce sites remove potential visibility for the ISP into a subscriber's search and activity on those sites. These technical and market facts limit the possible insights any ISP can gain by virtue of providing the underlying Internet connection. In short, ISPs have no systematic advantage in the e-commerce context over other entities, are not leading e-commerce players, and the nature of their services provides them with limited e-commerce data.

D. Conclusion

For each of the three contexts – browsers, Internet video, and e-commerce—ISPs face similar factors as in the previous Chapters regarding social networks, search, webmail and messaging, and the IBA system generally. ISPs do not offer their own web browsers, and so do not have an opportunity to collect data comparable to web browser developers. Even when those browsers share data with their developers, those connections are almost always encrypted, blocking the technical ability for ISPs to see that data. Moreover, ISPs do not have unique or privileged access into user Internet video content. Lastly, ISPs are also limited players in the e-commerce context, due to the infrequent nature of their sales interactions with customers, and have no unique or privileged access into a user's e-commerce transactions generally.

Cross-Context Tracking

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 8: Cross-Context Tracking

This Chapter describes how players in the online advertising ecosystem take information from varying sources, or varying “contexts,” and bring them together to gain greater insight into users’ online activity. The previous Chapters have highlighted important contexts where information is gathered about users’ online activity, such as through Internet Service Providers (“ISPs”), social networks, search, webmail, operating systems, mobile device information, interest-based advertising (“IBA”), web browsers, Internet video, and e-commerce. Each Chapter has highlighted the data that can be collected from each context, thereby providing ISPs and even more so, non-ISPs, with visibility and insight into users’ online activity. Some of these contexts, such as IBA, provide greater insight, especially about the URLs a user visits. Others, such as social networks and webmail, provide insight into the content a user is viewing, receiving, or sending. Some, like search, can provide both.

The term “context” is useful for a number of reasons. First, the term is intuitive – readers can understand search as a different “context” than e-commerce or social networks. Second, the term draws on the writings of Professor Helen Nissenbaum, including her widely-read 2010 book “Privacy in Context: Technology, Policy, and the Integrity of Social Life.”¹ Third, the Obama Administration explicitly adopted the idea of “respect for context” as one of the seven principles for its 2012 Consumer Privacy Bill of Rights.² In this Working Paper, we do not seek to define precisely what constitutes a “context” or a “respect for context.”³ Instead, given the significant academic and U.S. government support for attention to context, we highlight how useful the term can be for discussion of online privacy and advertising practices.

What we call “cross-context” tracking is what occurs by combining two or more of the types of data discussed in the previous Chapters. In previous Chapters, we have generally described each context – such as a social network, search engine, or webmail – as though the data were collected only in that single context. In fact, the same company often plays a role in multiple contexts, such as when an operating system company also has a search engine, or a social network company has an advertising network.⁴ Until now, the Working Paper has followed the one-context-at-a-time approach to make it easier to describe and explain the actual data flows and accompanying possible privacy issues. This approach clarifies what does and does not result from a particular position in the Internet ecosystem, such as an ISP.

¹ Helen Nissenbaum, “Privacy in Context: Technology, Policy, and Integrity of Social Life,” Stanford Univ. Press (2010).

² The White House, “Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy,” Feb. 2012, (<https://www.whitehouse.gov/sites/default/files/privacy-final.pdf>). The Consumer Privacy Bill of Rights defined “Respect for Context” as “Consumers have a right to expect that organizations will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.” *Id.* at 1.

³ Professor Nissenbaum has recently examined multiple possible definitions of “context” and “respect for context,” and this Working Paper does not seek to choose among the possible definitions. Helen Nissenbaum, “Respecting Context to Respect Privacy: Why Meaning Matters,” Science and Engineering Ethics (2015), (<http://link.springer.com/article/10.1007/s11948-015-9674-9>).

⁴ We have followed the same approach in defining ISPs as companies acting in their role as ISPs, even if the same corporation acts in other contexts, such as having an advertising network.

CROSS CONTEXT CHART

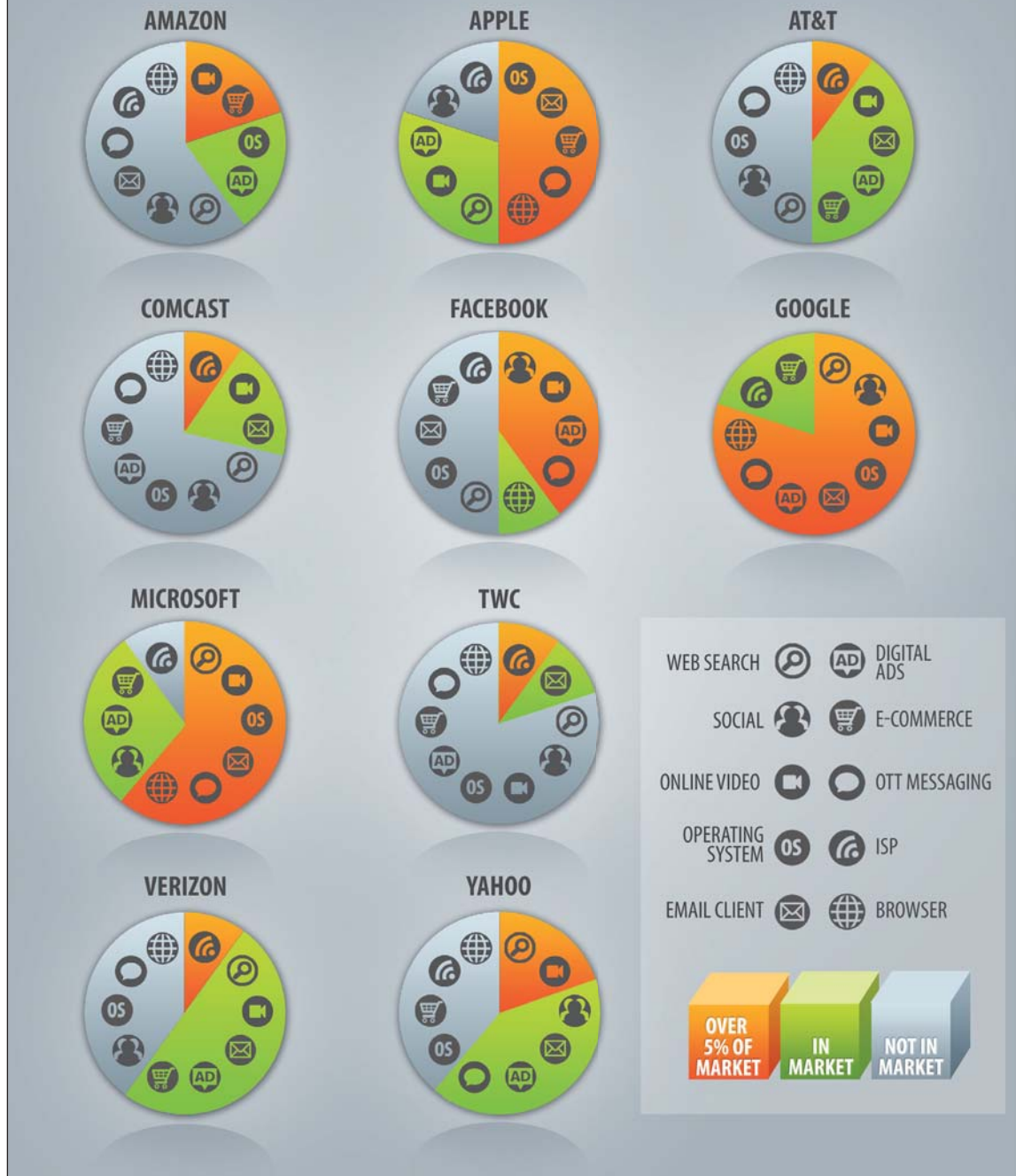


Diagram 8-A

The Cross-Context Chart of Diagram 8-A moves beyond the one-context-at-a-time approach to show leading players' multiple contexts in today's online ecosystem. The essential point of the Cross-Context Chart is simple – the same companies are often the leaders today in multiple contexts. Unique insights into users' online activity come substantially from cross-context tracking, the ability of a company to take advertising actions based on information collected in more than one context.

The Cross-Context Chart illustrates these points, rather than claiming that each company is indisputably categorized in the correct way. For each of 10 contexts, the chart indicates for each company whether it is “over 5% of market,” “in the market” (having a significant commercial business), or not in the market. Diagram 8-A contains citations for each company’s categorization for each context; our choice of companies and contexts was shaped by where we could find useful statistics. Other researchers could undoubtedly find specific places to disagree about whether a company is “over 5% of market” or “in the market,” or whether a company is “in the market” or not. Although researchers might thus disagree about details of the Cross-Context Chart, we believe the basic point is clear and correct – leading companies gain insight by combining information about users gathered in multiple contexts.

The history of computing shows what counts as important context changes over time. Context convergence is a standard feature of computing. A simple example of how contexts have converged over time is computer hardware. In the past, various parts of computer hardware had to be purchased separately, but today multiple functions are built into one microprocessor from Intel or another chip company. Similarly, today’s mobile operating systems (“OS”) bring together numerous software functions that used to be separate, from calculators and calendars to email and maps. The discussion here does not assert that our list of key contexts is the only list possible or will be the best list in years to come. Instead, the Cross-Context Chart documents 10 contexts that appear prominently today: web search, social, online video, operating system, email client, digital ads, e-commerce, over-the-top messaging, ISP, and browser.⁵

This Chapter describes two ways that a cross-context tracking company can build a context map: Through logged-in browsing (sometimes called deterministic) and not logged-in browsing (sometimes called probabilistic). Each method uses different techniques for mapping a user’s actions in different contexts and impacts whether the user is identified or unidentified. This Chapter then provides three examples of cross-context tracking: The combination of search with web browsers, the combination of a social network with an advertising network, and the combination of a diverse suite of services under a unified privacy policy. Lastly, this Chapter discusses why ISPs in particular are at a disadvantage in creating context maps compared with others in the ecosystem.

The importance of cross-context tracking is increasing for at least two reasons. First, the growing use of encryption limits the visibility of a user’s activities to all except those who are party to the encrypted communication. Along with providing security of transmission, *encryption also enforces separation of different contexts*. Where a company seeks to view more of a user’s online activity, the increase in encryption will be accompanied by more cross-context mapping. Second, the discussion throughout the Working Paper has highlighted reasons that diverse contexts can gain important insights into user Internet activities. Sometimes it is hard to remember how recently we have seen the rise of activities that now seem central to the Internet: Google was incorporated in 1998,⁶ Facebook opened to public users in 2006,⁷ and smartphone growth took off with the introduction of the iPhone in 2007.⁸ Search, social networks, and smartphones are now major contexts that are relevant to online advertising, yet cross-context mapping for these activities only emerged when the contexts became so important.

⁵ If the reader disagrees with this precise list, it does not take away from two points: (1) there can be different companies or units in a company that are leaders in one context, but not in other contexts, and (2) cross-context tracking is an important component of the current online advertising ecosystem.

⁶ Stephanie Buck, “Happy Birthday Google: Making Sense of the Web for 13 Years,” *Mashable*, Sep. 4, 2011, (<http://mashable.com/2011/09/04/google-happy-birthday-13-years/#smf2OjybGmqB>).

⁷ Michael Arrington, “Facebook Just Launched Open Registrations,” *Mashable*, Sep. 26, 2006, (<http://techcrunch.com/2006/09/26/facebook-just-launched-open-registrations/>).

⁸ Brian X. Chen, “June 29, 2007: iPhone, You Phone, We All Wanna iPhone,” *Wired*, June 29, 2009, (http://www.wired.com/2009/06/dayintech_0629/).

As with the other Chapters in the Working Paper, this discussion of cross-context tracking undermines the widely-held but mistaken view that ISPs have comprehensive and unique knowledge about user online activity because they operate the last mile broadband access service, which connects end users to the Internet. As many of the previous Chapters discuss, non-ISPs tend to be market leaders in most of the contexts that provide unique insight into users' online activity. As illustrated in Diagram 8-A, many of these non-ISPs also tend to be market leaders or have services in multiple contexts, providing them with effective visibility into users' online activity.

A. Cross-Context Tracking Data Flows

We examine two principal categories of cross-context tracking: Logged-in (deterministic) and not logged-in (probabilistic).

1. Logged-in cross-context tracking (deterministic tracking)

Now, more than ever, individuals use multiple services while logged-in to a particular account. For instance, users typically must log-in to send webmail or access a social network page; while logged-in, users might access multiple other services from the same company. When the user logs-in, a cross-context tracking company has the ability to keep track of that user across the multiple services. The ability of a company to connect a user in disparate contexts is sometimes called "deterministic" tracking because the identical log-in provides a basis for determining that it is the same user.⁹

This cross-context tracking can both enhance user functionality and serve an advertising function. When the individual is logged-in, the cross-context tracking company can tailor its services. For instance, a user might ask for map directions to a friend's house; when the user begins typing the address, the company might autocomplete the request, drawing on an earlier email or the contacts program. The user gets the proper address more easily, speeding the journey.

Logged-in cross-context tracking also serves an advertising function, as the cross-context tracking company is able to connect all of the services an individual uses to that particular user's account. In some instances, the user may not be strongly identified, such as for an easily-created webmail account. Often, however, the user provides name and credit card or other information such as a phone number, which can generally be linked to a named individual. When the user's actual identity is confirmed, then the online information holder can append offline information about the user. As discussed further below, the cross-context tracking company has more visibility into users' online activity and meets multiple advertising goals when it offers a suite of services to a single user.

2. Not logged-in cross-context tracking (probabilistic tracking)

The cross-context tracking company can also build a context map without login information, based on inferences from other data available to the company. This approach is sometimes called "probabilistic" tracking because the links are based on assessments of probabilities rather than deterministically from the same login information.¹⁰

Not logged-in context maps are still built around an individual user or device, but without the attached user account. Instead, the cross-context tracking company will look at all the data it collects and can access and use a proprietary algorithm to determine with some degree of certainty which activity in different contexts it believes

⁹ Ricardo Bilton, "Cross-Device Tracking, Explained," *DigiDay*, Aug. 21, 2015, (<http://digiday.com/publishers/deterministic-vs-probabilistic-cross-device-tracking-explained-normals/>).

¹⁰ Allison Schiff, "A Marketer's Guide to Cross-Device Identity," *Ad Exchanger*, April 9, 2015, (<http://adexchanger.com/data-exchanges/a-marketers-guide-to-cross-device-identity/>).

belong to the user.¹¹ This sort of probabilistic tracking is common in the online IBA context, where companies expend considerable efforts synching cookies and otherwise associating the Internet usage of one device or user with other online activity they believe is done by the same user.

One role of not logged-in context maps is to extend the insights that a company has from a user's logged-in activities. For example, suppose the company has logged-in information from webmail and search, but knows little about what that user does elsewhere on the Internet. In that instance, the cross-context tracking company has a business incentive to team up with a demand-side or supply-side platform, or other company that might have more insight into the URLs for that individual. By acquiring additional data, or improving its big data analytics, the cross-context tracking company can improve its ability to target advertisements.

Overall, both deterministic and probabilistic cross-context tracking provide companies with new visibility and insight into users' online activity. Often these companies are non-ISPs, as non-ISPs have large market shares in most of the contexts discussed throughout this Working Paper. These tracking methods greatly expand the data collection and analytic capabilities of non-ISPs, because non-ISPs can track the same user as she uses different services and platforms, thus giving the non-ISP an expansive view of that user.

B. Examples of Cross-Context Tracking

We next briefly describe three examples of cross-context tracking: Unified search and web browsers, the combination of a social network with an advertising network, and the combination of a diverse suite of services under a unified privacy policy.

1. Unified search and web browsers

Unified search and address bars are one example of cross-context tracking, combining the two data flows of search and web browsing. Historically, the address bar, where a user enters a URL, was separate from a search engine. For example, a user might have used the web browser address bar to type in www.Retailer.com and separately go to a search engine to research stores in their city that sell a specific product. The address bar would take the user to www.Retailer.com, while the search engine would take her to a list with the most relevant websites, videos, etc. If the user used two different companies for the address bar and the search, then the companies would generally not have been able to aggregate the data and see that the same person was both visiting the retailer's website and searching for stores in that city selling the product. The web browser's and search engine's visibility into the URLs the user searched and visited would have been segmented.

The trend in recent years has been to combine web browsers and search so users can both visit URLs and search specific keywords.¹² This unification can be attractive to users, who can seamlessly switch between URLs and search. It also can benefit the advertising efforts of a company that receives information about users through these previously separate services. Indeed, the trend may be pronounced enough that users may increasingly consider search and browsing to be a single context.

2. Combination of a social network with an advertising network

The core definition of a social network has involved users uploading information to the service about themselves or people they know. Over time, a range of social plugins has developed so that users visiting a website today can show approval of a web page for a wide range of services such as Digg, Facebook, Pinterest, Reddit, and

¹¹ *Id.*

¹² Chrome Browser, Google, (<https://www.google.com/chrome/browser/desktop/>).

Twitter.¹³ A key feature of “social” plugins is they assist the user in signaling interest to friends and colleagues about something on the Internet.

Social plugins have become important for functions previously considered to be a different context, as described in Chapter 6 on IBA. In a study of the “Like” button and other Facebook social plugins, one study said: “The near-ubiquity of the social plugins also makes them the ideal tool for collecting the browsing activities of Web users.”¹⁴ For example, a “Wall Street Journal examination of nearly 1,000 top websites found that 75% now include code from social networks, such as Facebook’s ‘Like’ or Twitter’s ‘Tweet’ buttons.”¹⁵ These social plugins can match a user’s online identity with her Internet browsing activities and can track a user’s arrival on a page even if the plugin is never clicked.¹⁶ This collection of URLs previously had been a separate context from the significant portion of content coming from the detailed posts in a social network.

This combination of social network and advertising network became more explicit in September 2015 when Facebook announced that the Like, Share, and Send social plugins will funnel data on individuals’ Internet browsing habits into Facebook’s advertisement-targeting tools and systems.¹⁷ Unless a user opts out, a visit to a page that has the social plugins enabled will result in the browser data being captured for advertisement targeting.

3. Unified privacy policy across services

A third example of cross-context tracking is where a company creates a unified privacy policy across services that historically have been different contexts. One example is the 2012 Google announcement that the company was implementing a “new main privacy policy that covers the majority” of Google’s products.¹⁸ Google stated that a main change from the previous policy was that Google “may combine information you’ve provided from one service with information from other services.”¹⁹ This policy explicitly enabled cross-context information sharing, such as among Gmail, Google Maps, Google Search, and YouTube, thus signifying that Google, a non-traditional ISP, has visibility into users’ activities across many of its various services and platforms.

C. Impact of Cross-Context Tracking on Advertising

The cross-context tracking company may create a context map to use with its own advertising business, or as a commodity to sell to other companies in the advertising ecosystem. When placing ads, a demand-side platform or other advertising company might use a context map for a variety of advertising benefits. A detailed discussion of these benefits is in the following Chapter on cross-device tracking. The application to cross-context advertising is quite similar, with benefits such as frequency capping, attribution of the value of different ads, better targeting

¹³ See “20+ Best Social Media Plugins for WordPress,” WordPress, (<http://www.wpxplorer.com/20-best-social-media-plugins-wordpress/>).

¹⁴ Gunes Acar, Brendan Van Alsenoy, Frank Piessens, Claudia Diaz, Bart Preneel, “Facebook Tracking Through Social Plug-ins: Technical Report Prepared for the Belgian Privacy Commission,” June 24, 2015, Version 1.1, (https://securehomes.esat.kuleuven.be/~gacar/fb_tracking/fb_plugins.pdf).

¹⁵ J. Howard Beales and Jeffrey A. Eisenach, “Putting Consumers First: A Functionality-Based Approach to Online Privacy,” *Navigant Economics*, Jan. 2013, (<http://www.broadbandforamerica.com/sites/default/themes/broadband/images/mail/puttingconsumersfirststudy.pdf>) citing Jennifer Valentino-Devries and Jeremy Singer-Vine, “They Know What You’re Searching For,” *The Wall Street Journal*, Dec. 7, 2012, (<http://www.wsj.com/articles/SB10001424127887324784404578143144132736214>).

¹⁶ *Id.*

¹⁷ Stephen Deadman, “A New Way to Control the Ads You See on Facebook,” *Facebook*, Sep. 15, 2015, (<https://facebook.com/notes/facebook-and-privacy/a-new-way-to-control-the-ads-you-see-on-facebook/926372204079329>).

¹⁸ “Updating Our Privacy Policies and Terms of Service,” *Google Official Blog*, Jan. 24, 2012, (<https://googleblog.blogspot.com/2012/01/updating-our-privacy-policies-and-terms.html>).

¹⁹ Sharon Profs, “Five Ways Google’s Unified Privacy Policy Affects You,” *Cnet*, March 1, 2012, (<http://www.cnet.com/how-to/five-ways-googles-unified-privacy-policy-affects-you/>).

of ads, sequenced advertising campaigns, and delivery of ads based on simultaneous or near-in-time activities by the user.

D. The Diminishing Visibility of ISPs in Cross-Context Tracking

The discussion here shows that information gathered by an ISP is merely one of an extensive list of contexts where the advertising ecosystem may gather data about users and devices. Chapter 1 explained how technological developments have made it difficult for ISPs to discern the content and URLs accessed by users who connect to the Internet via their systems. The shift to multiple, mobile devices and the growing use of encryption are important contributors to the lower portion of a user's online activity that ISPs can see over time. These changes also heighten the importance of cross-context tracking for companies who are parties to the encrypted information, such as webmail providers or websites using HTTPS.

The Cross-Context Chart of Diagram 8-A shows that it is difficult to examine the leading contexts today and to make the case that ISPs have a unique or dominant market position due to the fact that they connect users to the Internet. For dominant contexts today – such as operating systems, search, social networks, and webmail – ISPs have not been the market leaders. In some portions of mobile and non-mobile online behavioral advertising (“OBA”), the ISP function is combined with significant market players; there is little reason to think, however, that the presence of the ISP function in a company has been an important determinant of success. For video, companies that are ISPs have greater prominence today, but there are multiple vigorous competitors in those contexts as well.

Just as ISPs have rarely been market leaders in today's most important contexts, so too it is difficult to see a reason that they have unique strengths for cross-context tracking. This Chapter has highlighted the advantage that a cross-context tracking company has if it keeps users logged-in across multiple contexts. The leading examples of this logged-in behavior are companies that are not major ISPs, thus undermining the widely-held view that ISPs have unique access to and knowledge about users' online activity.

Citations for Appendix 1: Cross-Context Chart¹

Amazon Chart Citations

Social: Amazon does not currently offer a social networking site.

Web Search: Amazon does not currently offer a general purpose search engine, even though it collects search data from Amazon users on Amazon's website.

Email Client: Amazon does not currently offer an email client service.

OTT Messaging: Amazon does not currently offer an over-the-top messaging service.

Digital Ads: Amazon accounted for 3 percent of total digital display ad revenue and 0.6 percent of mobile ad revenue. (<http://www.emarketer.com/Article/Facebook-Twitter-Will-Take-33-Share-of-US-Digital-Display-Market-by-2017/1012274>), (<http://www.forbes.com/sites/greatspeculations/2015/07/09/u-s-digital-advertising-landscape-and-key-players-part-2/#2d21b54673ed87351bd73eda/>).

Operating Systems: Amazon's "Fire" devices run on a proprietary operating system, but is not currently ranked among top mobile operating systems.

Browser: Amazon does not currently offer a web browser.

Online Video: Amazon's Prime Video has subscriptions from 13.0 percent of U.S. households as of 2015. (<http://www.geekwire.com/2015/netflix-still-king-of-streaming-video-but-amazon-gaining-market-share/>).

E-commerce: Amazon Sites are currently ranked 1st among retail websites with 188 million unique monthly visitors. (<http://www.statista.com/statistics/271450/monthly-unique-visitors-to-us-retail-websites/>).

Broadband Internet Service: Amazon does not currently offer broadband Internet service.

Apple Chart Citations

Social: Apple previously offered the iTunes Ping social network, but shut the service down on September 30, 2012. (<http://www.pcmag.com/article2/0,2817,2409675,00.asp>).

Web Search: Apple currently offers Spotlight Search, a search engine that allows users to make searches on their Apple devices. Search requests will include data from the various music, applications (e.g., Safari), contacts, and email messages on the device. (<http://searchengineland.com/ios-9-apple-siri-spotlight-search-230814>).

Email Client: Apple email services ("Mail" on iOS devices and Apple Mail) account for a total 53 percent share of email client market as of December, 2015. (<http://emailclientmarketshare.com/>).

OTT Messaging: "One service that can't be quantified so easily is Apple's in-house iMessage, which Evans refers to as 'dark matter' because Apple guards its usage statistics closely." Evans notes: "It's probably big, with over 400 million iPhones in use today, but we don't know how big." (<http://www.businessinsider.com/whatsapp-vs-texting-statistics-2015-1?r=UK&IR=T>).

¹ For any corrections to the items listed in this Appendix, as for any other corrections for the paper, please send an email to comments@iisp.gatech.edu.

Digital Ads: Apple did not rank for total digital display ad revenue, and accounted for 2.8 percent of total mobile ad revenue in 2015. (<http://www.emarketer.com/Article/Facebook-Twitter-Will-Take-33-Share-of-US-Digital-Display-Market-by-2017/1012274>), (<http://www.forbes.com/sites/greatspeculations/2015/07/09/u-s-digital-advertising-landscape-and-key-players-part-2/#2d21b54673ed87351bd73eda/>).

Operating Systems: Apple accounts for 5.1 percent of desktop operating systems, and 43.3 percent of mobile operating systems. (<https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0>), (<http://www.statista.com/statistics/266572/market-share-held-by-smartphone-platforms-in-the-united-states/>).

Browser: Apple's Safari browser accounts for 27.04 percent of browser market share. (<http://gs.statcounter.com/#all-browser-US-monthly-201512-201512-bar>).

Online Video: Apple's iPod & iTunes accounted for 1 percent of total video site visits. (<http://www.statista.com/statistics/266201/us-market-share-of-leading-internet-video-portals/>).

E-commerce: Apple.com sites are currently ranked 4th among retail websites with 80 million unique monthly visitors. (<http://www.statista.com/statistics/271450/monthly-unique-visitors-to-us-retail-websites/>).

Broadband Internet Service: Apple does not currently offer broadband internet service.

AT&T Chart Citations

Social: AT&T does not currently offer a social networking site.

Web Search: ATT.net offers search operated by Yahoo! (<https://att.yahoo.com/>).

Email Client: AT&T email service is not currently ranked in the top 10 for email client market share. (<http://emailclientmarketshare.com/>).

OTT Messaging: AT&T does not currently offer an over-the-top messaging service.

Digital Ads: AT&T partners with Yahoo! for digital ads on att.net (<https://policies.yahoo.com/us/en/att/privacy/adinfo/index.htm>).

Operating System: AT&T does not currently offer an operating system.

Browser: AT&T does not currently offer a web browser.

Online Video: AT&T currently offers video streaming services through the Elation joint venture. (<http://www.ellation.com/about>).

E-commerce: AT&T currently offers some direct hardware sales online, but does not currently rank among top U.S. retail websites. (<http://www.statista.com/statistics/271450/monthly-unique-visitors-to-us-retail-websites/>).

Broadband Internet Service: AT&T had 15.8 million fixed broadband subscribers, and approximately 25 percent of total wireless subscriptions. (<http://www.leichtmanresearch.com/press/111715release.html>), (<https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/strategy-analytics-press-release/2015/06/30/us-wireless-market-to-add-100-million-subscribers-by-2020-says-strategy-analytics#.VqAIDvkrKM8>).

Comcast Chart Citations

Social: Comcast does not currently offer a social networking site.

Web Search: Comcast does not currently offer web search.

Email Client: Comcast's email service is not currently ranked in the top 10 for email client market share. (<http://emailclientmarketshare.com/>).

OTT Messaging: Comcast does not currently offer an over-the-top messaging service.

Digital Ads: Comcast does not currently own or operate its own digital ad platform.

Operating System: Comcast does not currently offer an operating system.

Browser: Comcast does not currently offer a web browser.

Online Video: Comcast is ranked in the top ten for top video content property. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-December-2015-US-Desktop-Online-Video-Rankings>).

E-commerce: Comcast does not currently offer direct retail sales online.

Broadband Internet Service: Comcast had over 22 million broadband internet subscribers and 40 percent of all broadband internet subscriptions as of 2015. (<http://www.statista.com/statistics/217348/us-broadband-internet-susbcscribers-by-cable-provider/>).

Facebook Chart Citations

Social: Facebook itself has the largest share of total visits with 45.4 percent. Instagram, which Facebook owns, is ranked 8th with 1.32 percent share of total visits. (<http://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/>).

Web Search: While Facebook offers robust search for posts and links within Facebook, including the ability to load those pages within the Facebook platform rather than through an external browser, Facebook does not currently offer the ability to search the Internet generally.

Email Client: While Facebook offers users an @facebook.com email address, it does not currently provide a full email service but instead forwards messages to a user's primary email address. (<https://www.facebook.com/help/224049364288051>).

OTT Messaging: Facebook Messenger and WhatsApp have a combined 1.5 billion active users. (<http://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/>).

Digital Ads: Facebook accounted for 25.2 percent of total display ad revenue and 16.7 percent of total mobile ad revenue in 2015. (<http://www.emarketer.com/Article/Facebook-Twitter-Will-Take-33-Share-of-US-Digital-Display-Market-by-2017/1012274>), (<http://www.forbes.com/sites/greatspeculations/2015/07/09/u-s-digital-advertising-landscape-and-key-players-part-2/#2d21b54673ed87351bd73eda/>).

Operating System: Facebook does not currently offer an operating system.

Browser: Facebook's mobile app offers an in-app browser for users, but does not currently account for an appreciable share of mobile browsers. (<http://www.androidcentral.com/how-turn-facebooks-app-browser-external-links>).

Online Video: Facebook is ranked second for top video content property. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-December-2015-US-Desktop-Online-Video-Rankings>).

E-commerce: Facebook does not currently offer direct retail sales.

Broadband Internet Service: Facebook does not currently offer broadband Internet service.

Google Chart Citations

Social: YouTube is ranked as the 2nd largest share of visits with 22.2 percent, while Google+ is ranked 10th and accounts for 0.9 percent. Therefore, combined Google social network products account for 23.1 percent of total share of visits. (<http://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/>).

Web Search: As of November 2015, Google sites are ranked first in share of search with 63.9 percent. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-November-2015-US-Desktop-Search-Engine-Rankings>).

Email Client: As of December 2015, Google email services (Gmail and Google Android mail) account for a total 25 percent share of email client market. (<http://emailclientmarketshare.com/>).

OTT Messaging: Authors could not find numbers of current active Google Hangouts users. However, given that Google Hangouts is included in the Android OS, which accounts for 52.9 percent of smartphones, and is also available in Gmail, which accounts for 15 percent of email client market share, authors believe Google qualifies as a market leader in OTT Messaging.

Digital Ads: Google accounts for 13 percent of digital display ad revenue and 35.2 percent of mobile ad revenue share. (<http://www.emarketer.com/Article/Facebook-Twitter-Will-Take-33-Share-of-US-Digital-Display-Market-by-2017/1012274>), (<http://www.forbes.com/sites/greatspeculations/2015/07/09/u-s-digital-advertising-landscape-and-key-players-part-2/#2d21b54673ed87351bd73eda>).

Operating System: Google's Chrome OS did not rank individually in total share of desktop operating systems, and Google Android accounted for 52.9 percent of mobile operating systems. (<https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0>), (<http://www.statista.com/statistics/266572/market-share-held-by-smartphone-platforms-in-the-united-states/>).

Browser: Google's Chrome and Android browser account for a combined 42.17 percent of browser market share. (<http://gs.statcounter.com/#all-browser-US-monthly-201512-201512-bar>).

Online Video: Google's YouTube accounted for 73.6 percent of total video site visits. Google Play also offers streaming video on demand. (<http://www.statista.com/statistics/266201/us-market-share-of-leading-internet-video-portals/>).

E-commerce: Google offers direct sale of some hardware as well as apps and electronic media purchases, but does not currently rank among top retail websites in the U.S. (<http://www.statista.com/statistics/271450/monthly-unique-visitors-to-us-retail-websites/>).

Broadband Internet Service: Google Fiber internet access is currently only available in limited markets. As of March 2015, Google Fiber had less than 30,000 total subscribers. (<http://www.kansascity.com/news/business/article13799168.html>).

Microsoft Chart Citations

Social: Microsoft offers a social network, Socl (<http://www.so.cl/>), but does not currently have a significant market share (not listed in 10 social networks based on share of visits). (<http://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/>).

Web Search: As of November 2015, Microsoft sites ranked 2nd in share of search with 20.9 percent. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-November-2015-US-Desktop-Search-Engine-Rankings>). This excludes search data obtained through partnerships.

Email Client: As of December 2015, Microsoft email services (Outlook, Outlook.com, and Windows Live Mail) accounted for a total 11 percent share of the email client market. (<http://emailclientmarketshare.com/>).

OTT Messaging: Microsoft's Skype has 300 million active global users. (<http://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/>).

Digital Ads: In 2015, Microsoft accounted for 1.7 percent of digital display ad revenue, and none of mobile display ad revenue. (<http://www.emarketer.com/Article/Facebook-Twitter-Will-Take-33-Share-of-US-Digital-Display-Market-by-2017/1012274>).

Operating Systems: Microsoft accounted for 89.73 percent of desktop operating systems, and 2.7 percent of mobile operating systems. (<https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0>), (<http://www.statista.com/statistics/266572/market-share-held-by-smartphone-platforms-in-the-united-states/>).

Browser: Microsoft Internet Explorer and Edge browsers account for a combined 18.76 percent of browser market share. (<http://gs.statcounter.com/#all-browser-US-monthly-201512-201512-bar>).

Online Video: Microsoft sites are ranked in the top ten for top video content property. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-December-2015-US-Desktop-Online-Video-Rankings>).

E-commerce: Microsoft offers hardware and software sales directly online, but does not currently rank among top retail websites in the U.S. (<http://www.statista.com/statistics/271450/monthly-unique-visitors-to-us-retail-websites/>).

Broadband Internet Service: Microsoft does not currently offer broadband Internet service.

Time Warner Cable Chart Citations

Social: Time Warner Cable does not currently offer a social networking site.

Web Search: Time Warner Cable does not currently offer a general purpose search engine.

Email Client: Time Warner Cable email service is not currently ranked in the top 10 for email client market share. (<http://emailclientmarketshare.com/>).

OTT Messaging: Time Warner Cable does not currently offer an over-the-top messaging service.

Digital Ads: Time Warner cable does not own currently or operate its own digital ad platform.

Operating System: Time Warner Cable does not currently offer an operating system.

Browser: Time Warner Cable does not currently offer a web browser.

Online Video: Time Warner Cable does not currently offer a stand-alone video streaming service separate from its television subscription services.

E-commerce: Time Warner Cable does not currently offer direct retail sales online.

Broadband Internet Service: Time Warner Cable has about 12 million subscribers and just over 20 percent of all broadband internet subscribers. (<http://www.statista.com/statistics/217348/us-broadband-internet-susbscribers-by-cable-provider/>).

Verizon Chart Citations

Social: Verizon does not currently offer a social networking site.

Web Search: AOL, which is owned by Verizon, accounts for 1.0 percent of Desktop search share and is ranked 5th in search. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-December-2015-US-Desktop-Search-Engine-Rankings>).

Email Client: Verizon and AOL email services are not currently ranked in the top 10 for email client market share (<http://emailclientmarketshare.com/>).

OTT Messaging: Verizon does not currently offer an over-the-top messaging service.

Digital Ads: AOL, which is owned by Verizon, accounted for 3.5 percent of total digital display ad revenue, and Millennial Media, which has been acquired by AOL, accounted for 0.3 percent of mobile ad revenue. (<http://www.fool.com/investing/general/2015/07/06/why-did-microsoft-corporation-ditch-its-display-ad.aspx>), (<http://www.emarketer.com/Article/AOL-Millennial-Face-Uphill-Battle-Capture-Mobile-Ad-Dollars/1012954>).

Operating System: Verizon does not currently offer an operating system.

Browser: Verizon does not currently offer a web browser.

Online Video: Verizon owns AOL which currently offers its own video hosting service, but is not currently ranked for total video site visits. (http://features.aol.com/?icid=gnavbar_rootvideo_main5), (<http://www.statista.com/statistics/266201/us-market-share-of-leading-internet-video-portals/>).

E-commerce: Verizon currently offers direct sales of products, but does not currently rank among top U.S. retail websites. (<http://www.statista.com/statistics/271450/monthly-unique-visitors-to-us-retail-websites/>).

Broadband Internet Service: Verizon has 9.2 million fixed broadband subscribers and approximately 32.5 percent of total wireless subscribers. (<http://www.leichtmanresearch.com/press/111715release.html>), (<https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/strategy-analytics-press-release/2015/06/30/us-wireless-market-to-add-100-million-subscribers-by-2020-says-strategy-analytics#.VqAIDvkrKM8>).

Yahoo Chart Citations

Social: Yahoo purchased social networking site, Tumblr, in 2013. (<http://www.businessinsider.com/marissa-mayer-heres-why-i-just-bought-tumblr-for-11-billion-2013-5>).

Web Search: Yahoo! Answers is ranked 9th with 1.28 percent share of total visits, and Tumblr is ranked 6th with 1.36 percent. Yahoo Sites, which are powered by Microsoft Bing, are ranked 3rd in search. (<http://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/>).

Email Client: As of December 2015, Yahoo! Mail accounts for a 3 percent share of total email client market. (<http://emailclientmarketshare.com/>).

OTT Messaging: Current numbers were unavailable for active Yahoo Messenger users, though Jeff Benaforte was “crossing his fingers for a hundred million users” in a December, 2015 Wired article. (<http://www.wired.com/2015/12/yahoo-messenger-texting/>).

Digital Ads: Yahoo accounted for 4.6 percent of total digital display ad revenue, and 3.7 percent of mobile ad revenue. (<http://www.emarketer.com/Article/Facebook-Twitter-Will-Take-33-Share-of-US-Digital-Display-Market-by-2017/1012274>), (<http://www.forbes.com/sites/greatspeculations/2015/07/09/u-s-digital-advertising-landscape-and-key-players-part-2/#2d21b54673ed87351bd73eda/>).

Operating System: Yahoo does not currently offer an operating system.

Browser: Yahoo does not currently offer a web browser.

Online Video: Yahoo is ranked in the top three for top video content property. (<https://www.comscore.com/Insights/Market-Rankings/comScore-Releases-December-2015-US-Desktop-Online-Video-Rankings>).

E-commerce: Yahoo does not currently have direct retail sales.

Broadband Internet Service: Yahoo does not currently offer broadband Internet service.

Cross-Device Tracking

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 9: Cross-Device Tracking

Advertisers today are mapping users as they move between devices in much the same way that they track users moving between contexts. As users use an increasing assortment of Internet-connected devices throughout their typical day, each device constitutes a percentage of that user's Internet activity.¹ Cross-device tracking is a tool an advertising tracking company employs for combining information about each of a user's devices in order to track as much of the individual's Internet activity history as possible. A cross-device tracking company does this by building a "device map" for each user. A device map centers on an individual user. The user can be pseudonymous (e.g., User abc123) or identified (e.g., John Smith). The device map then connects every device the company either knows belongs to that user or believes likely belongs to that user. Then, any information about browsing history and content tied to those various devices can be combined for a total of the user's Internet activity and history.

This Chapter describes the two ways cross-device tracking companies can build a device map: through logged-in browsing (often called "deterministic") and not logged-in browsing (often called "probabilistic"). Each method uses different techniques for mapping devices to a user and impacts whether that user is identified or unidentified. Drawing on the technologies discussed in previous Chapters, this Chapter next discusses how a device map improves the ability to meet goals such as frequency capping, more accurate attribution for advertisements, improved ad targeting, sequenced advertising campaigns, and simultaneous user behavior tracking. This Chapter concludes by comparing Internet Service Providers ("ISPs") and non-ISPs in creating user device maps, and notes the advantages of non-ISPs.

The importance of cross-device tracking increases as the number of devices per user increases. In the 1990s, many users connected to the Internet via one home computer, often shared within a family. By 2016, it is not unusual for an individual to connect to the Internet via a handful of devices, such as a laptop, tablet, smartphone, work computer, game console, and perhaps others. Going forward, studies about the Internet of Things predict an exponential increase in the number of devices connected to the Internet, to roughly 12 devices per capita by 2019, and far more after that.² Typical users in the near future may deploy numerous additional devices through connected cars, connected homes, wearable computers, and innovative ways that are difficult to know today. As the number of devices grows from one, to a few, to many, so too does the importance of cross-device tracking.

A. Cross-Device Tracking Data Flows

As with cross-context tracking, cross-device tracking can occur when users are logged-in (deterministic tracking) or not logged-in (probabilistic tracking).

1. Logged-in cross-device tracking (deterministic tracking)

Intuitively, it is simple to connect the devices of a user who logs in to the same company's services from multiple devices. Although there is some possibility that more than one person logs-in to that account, for most practical purposes the same login indicates the same user. For purposes of a company building a cross-device tracking device map, a user who logs-in from a new device enables the company to link any unique device identifier to

¹ In 2014, there was an average of 1.2 mobile-connected devices per capita in the U.S., and that number is expected to grow to 3.2 devices per capita by 2019. "VNI Mobile Forecast Highlights, 2014-2019," Cisco, (http://www.cisco.com/c/dam/assets/sol/sp/vni/forecast_highlights_mobile/index.html#~Country).

² The total number of networked device, both mobile and non-mobile, was 6.2 devices per capita in 2014, and is expected to grow to 11.7 networked devices per capita by 2019. "VNI Forecast Highlights," Cisco, (http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights/index.html).

the user, such as a smartphone ID or a static Internet Protocol (“IP”) address.³ The ability of a company to connect a user’s devices via login is sometimes called “deterministic” tracking because the identical login provides a basis for determining that it is the same user.

As an example, assume a single user account accesses three devices: a personal laptop, a personal smartphone, and a work desktop. Without logging-in, a cross-device tracking company would see each of these devices as three separate entries: Device1, Device2, and Device3. Once the user’s account logs-in on each device, however, the cross-device tracking company now has a single entry, “Username’s Account,” which aggregates all tracking data from Device1, Device2, and Device3. This way, the cross-device tracking company knows when “Username’s Account” is served an ad, which ad and version is served, when it is served, to which device it is served, potentially the device’s location at that point, and whether “Username’s Account” clicked on the ad.

Logged-in cross-device tracking can also create a device map around a known identity, rather than just a known account. Social network accounts often include a user’s real name, verifiable information on personal details, and other identity-tied data points. Retailers can connect devices when the same user purchases from both a mobile device and a laptop. Other companies offer a suite of services to logged-in users, so that a login to one service can be linked to the devices used to log-in for other services.

In these examples, identities may be considered deterministic even when they are not 100 percent verified. Stronger forms of authentication, such as in-person presentation of a passport, are rarely required in settings used for advertising. Nonetheless, the logged-in accounts for the same user provide a high enough level of certainty for most advertising purposes. When that percentage of certainty is high enough,⁴ the cross-device tracking company’s device map is now centered around John Smith, who owns “Username’s Account,” which is on Device1, Device2, and Device3.

2. Not logged-in cross-device tracking (probabilistic tracking)

A cross-device tracking company can also build a device map without login information based on inferences from other data available to the company. This approach is sometimes called “probabilistic” tracking, because the links are based on assessments of probabilities rather than deterministically from the same login information.

Not logged-in device maps are still built around an individual user, but without an attached username account. Instead, the cross-device tracking company will look at all the data it collects and apply a proprietary algorithm to determine to some degree of certainty which devices it believes belong to the user.⁵ For most advertising purposes, a company is willing to bid some amount for an ad with considerably less than 100 percent certainty that the same user deploys the multiple devices. In practice, the level of probabilistic certainty may be quite high. One cross-device tracking company has claimed as high as 90 percent certainty in the accuracy of its device maps.⁶

³ Jules Polonetsky and Stacey Gray, “Cross Device: Understanding the State of State Management,” *Future of Privacy Forum*, Nov. 2015, (https://fpf.org/wp-content/uploads/2015/11/FPF_FTC_CrossDevice_F_20pg-3.pdf).

⁴ This Working Paper does not purport to determine what any actual cross-device tracking company would consider a sufficient degree of certainty for identity, or for any other point in this Chapter. Rather, this Working Paper discusses the capabilities of a hypothetical cross-device tracking company based on current technology.

⁵ This type of algorithmic mapping allows a cross-device tracking company without logged-in data to potentially account for any connected device that permits some level of tracking, even when no account logins are present.

⁶ “With 91.2% data accuracy confirmed by Nielsen, the company offers the largest in-market opportunity for marketers . . .” “Who We Are,” *TapAd*, (<http://www.tapad.com/about-us/who-we-are/>).

A cross-device tracking company can combine its tracking data with other sources of data to improve its device map. For example, the cross-device tracking company might purchase tracking information from a different advertising company that uses logged-in tracking. This sort of purchased data helps by providing more data and improving data analytics. First, the new purchased data helps the company become aware of links between devices that it did not previously know; for instance, the purchased data might include a tablet that the company did not know was linked to a smartphone. Second, the cross-device tracking company can improve its analytics by checking the results of its algorithm for mapping devices against the newly acquired data.

Important trends support the greater use of not logged-in cross-device tracking. First, the growing number of devices per user increases the business rationale for cross-device tracking. Whereas advertisers seek a more complete history of Internet activity, the splintering of a user's Internet activity into multiple devices reduces advertisers' ability to achieve that goal. Cross-device tracking is a remedy for that splintering. Second, a large and growing number of industry players participate in selling or brokering information about how to link multiple devices.⁷ Third, this rapid growth in available data is accompanied by advances in big data analytics, so that industry experts can make more varied and informed inferences about which devices might be linked to the same user.

There is also an underappreciated connection between logged-in and not logged-in cross-device tracking. Suppose a particular user has six devices. Based on logged-in information, a company might be able to connect some of these devices, such as three. Through not logged-in techniques, over time a cross-device tracking company might be able to add the fourth, fifth, and sixth devices to its device map. This example illustrates that, at any given moment, some of a user's devices might be identified deterministically and some probabilistically. The two approaches work together as the cross-device tracking company works toward its goal of linking all of a user's current devices.

B. Impact of Cross-Device Tracking on Advertising

A cross-device tracking company may create a device map to use with its own advertising business, or as a commodity to sell to other companies in the advertising ecosystem. A company buying advertising, which is trying to place each ad in as effective of a place as possible, might use a device map for benefits, including frequency capping, more accurate attribution for advertisements, improved ad targeting, sequenced advertising campaigns, and simultaneous tracking. The Working Paper's discussion of the various contexts of the online ecosystem enables a clearer explanation of these online advertising features and how information is used in the advertising ecosystem.

There are important non-advertising uses of cross-device tracking. For users, cross-device tracking provides convenience and functionality. Especially as many services shift to the cloud, users gain flexibility in connecting from many places and many devices while maintaining full functionality. Cross-device tracking has important security purposes for both users and companies. Such tracking is important for detecting suspected account takeover or other fraudulent activity. For instance, a cross-device tracking company may notify the user if there is a log-in attempt from an unknown device, alerting the user to possible unauthorized activity. Similarly, the company may ask additional challenge questions (e.g., mother's maiden name) or for other additional authentication when someone tries to log into a user account from a new device. These sorts of security benefits rely on cross-device tracking so that different procedures are followed when there is suspicion whether a device is authorized. The discussion here does not attempt to engage in any overall cost/benefit analysis of cross-device tracking. Instead, the discussion focuses on advertising-related uses of cross-device tracking, especially to understand the relative

⁷ "Comments to the Federal Trade Commission: Cross Device Tracking Workshop," *Future of Privacy Forum*, Oct. 16, 2015, (<https://fpf.org/wp-content/uploads/2015/10/FTC-Cross-Device-Comments-Oct-16-2015-Understanding-the-State-of-State-Management.pdf>).

visibility of ISPs in the emerging ecosystem.

1. Frequency capping

An accurate device map can boost the return on investment for block purchase of advertising impressions. When buying advertising impressions, a marketer is generally looking for some number of impressions for each user. If it purchases 1,000 impressions, the marketer might prefer that those impressions go to 1,000 different people; or, if it believes five showings of an ad is most effective, then it might prefer to reach 200 people five times each. Advertisers thus seek to set frequency caps, depending on how often they believe it is useful to show the same ad to the same user.

If the marketer prefers a single impression per user, cookies have been an effective way to achieve frequency capping. For a single device, the cookie can allow an advertiser to prevent serving the same ad twice to the same user, as long as that cookie is present. However, that user may still see the same advertisement on their other devices because the same cookie is not present on the other devices. A device map addresses this problem, preventing the ad from being served on any of the devices mapped to a particular user. So, if a single user sees the company's ad on her smartphone, she should not then see it again when using her laptop at a later time.

2. Attribution

A device map can also be useful for attributing user purchasing behavior to previous advertising impressions, an important variable in the price for an ad. Without a device map, only advertisements on the device where a purchase is made can be considered as contributors to the purchase. For example, assume a user sees an ad for a product on her smartphone at 1:00 p.m. She then goes to her laptop at 1:10 p.m. and uses a search engine to look up the product. The search engine returns a sponsored ad for the product, which the user clicks. The user then purchases the product. Without a device map, it would appear to the product's company that the advertisement on the search engine drove the sale, increasing the perceived value of the search ad and decreasing the value of the smartphone ad impression.

With an accurate device map linking the user's smartphone and laptop, however, it would be apparent to the product's company that the impression on the smartphone drove the user to purchase, thereby increasing the observed value of the smartphone ad impression. The device map allows the product's company to compare which advertisements were seen on all of the user's devices, and how close in time the advertisements were served to better evaluate the return on investment ("ROI") for each purchase.

3. Improved advertisement targeting

A device map can allow for improved data analytics on tracking and demographic information in order to serve more specifically-targeted advertisements. As a general rule, the more data available about a user or device, the more that can be accomplished through data analytics. Here, the high degree of confidence in a device map (near-certainty for logged-in cross-device tracking, substantial likelihood for not logged-in cross-device tracking) improves the capability of data analysts using these device maps.

Device maps also allow for improved form-factor targeting, helping ensure that a user sees an advertisement on the best available screen. For example, say a company wishes to serve an ad to every user who attends an event. When User1 goes to the event, the company can see that User1's smartphone is present from its location data and can serve an advertisement. However, the company might prefer that User1 sees the ad on a laptop rather than

a smartphone.⁸ An accurate device map can allow the company to mark the tracking information from User1's laptop so it can bid on the next available impression on that device. Without a device map, the company would have to serve the ad to User1's smartphone, or choose not to serve the ad at all.

4. Sequenced advertising

An accurate device map can allow a company to serve a sequence of ads to a user across multiple devices. A classic example of sequenced advertising is the Burma Shave highway billboard advertisement. Burma Shave placed a number of different billboard ads along a stretch of highway; each built off the information delivered from the previous one, knowing that the driver would see them in sequence along the highway. A sequence of advertisements can also be used to reinforce a point, or to better attract a user's attention. With a device map, a company can ensure that its sequence of ads is served in order, regardless of the device the user is on at the time. The company can also effectively restrict how soon the next advertisement in the sequence is served. While sequenced advertising is not particularly prevalent in the current online ecosystem, improved device maps can increase the value of these types of coordinated campaigns in the future.

5. Tracking simultaneity

Device maps can help companies track users' multi-screen activity. For example, a user might be watching streaming video content on her Smart TV while also using her tablet. With an accurate device map, a company can see which ads are being served by the streaming video content, and sync those ads with the ones served on the user's tablet. As device maps and tracking on non-traditional devices evolves, the ability to track and serve advertisements during multi-screen interaction will continue to grow. Moreover, because users already often engage on multiple screens, especially when viewing video content, accurate device maps will allow data analysts to determine how best to take advantage of advertising during simultaneous multi-device engagement.

6. Summary on advertising uses

In short, accurate cross-device mapping allows for better targeting of advertisements as well as better evaluation of advertisements' relative value. The ability to know the ad impressions an individual user sees prior to a purchase event, regardless of the device where the impression was served, allows a marketer to better evaluate the accurate ROI for that ad. A more accurate ROI also allows the company to better evaluate what about those higher-value ad impressions helped drive a sale, resulting in better performing advertisements in the future. As discussed in Chapter 2, Facebook Atlas provides an example of cross-device tracking. It offers advertisement tailoring to better target ads to specific customers and locations, including the capability to correlate offline purchases with online advertising impressions.⁹

C. The Limited Visibility of ISPs for Cross-Device Tracking

In considering the impact of cross-device tracking on ISPs and non-ISPs, the fundamental trend is that the visibility of ISPs is increasingly limited. As the average number of devices per user increases, the connection between one device and the Internet covers a smaller fraction of the user's Internet activity history. A home Internet connection covered most or all of many users' Internet activity history in the 1990s; the same connection today to a home desktop covers a much smaller fraction.

⁸ The size of the ad may be preferable for a larger screen, and an ad containing video or audio may be more likely to be seen over a non-mobile connection.

⁹ For a more detailed discussion of Facebook Atlas specifically, see Chapter 2: Social Networks.

Parallel to the rise of multiple devices for each consumer is the rise of multiple ISPs for each consumer. Today, any one user often has relationships with multiple ISPs. One ISP may provide their wired broadband connection at home, a second provides their mobile broadband connection, a third provides their broadband connection at work, and the user may connect daily to any number of other WiFi hotspots operated by different ISPs. With mobile devices in particular, each device often switches between ISPs over the course of a day. When at home, a mobile device may use the local network WiFi, but it uses a different ISP for its mobile broadband outside the home. Thus, each ISP often only has insight into a diminishing fraction of a user's total devices, and even then only provides Internet service to some fraction of each device's total use.

Additionally, a significant portion of users change ISPs, further splintering their total Internet activity history across multiple ISPs. According to the Federal Communications Commission ("FCC"), one out of six customers switches wireline providers every year, and 37 percent of customers switch every three years.¹⁰ Between one-fifth and one-third of total subscribers also switch their wireless carriers annually.¹¹ The ISP turnover rate is approximately twice of those with non-ISPs, both mobile and non-mobile.¹²

The comparison is striking to a service that sees a user logged-in across multiple devices, such as a social network or an online company that offers a suite of services. When a user is logged-in to a company for a high fraction of the time she is connected to the Internet, that company gains visibility to a larger number of the devices in her device map, and a larger fraction of the browsing done on each individual device. For these logged-in services, the company continues to have access regardless of where the device is or which ISP it is currently using. Such a company sees a larger percentage of the Internet activity history for both the user and each device, compared to any one ISP.

ISPs are one category among many that have partial visibility into a particular user's device map. The home ISP has an account relationship with the home user, knowing the subscriber's name and billing address, and may have their credit card information and phone number. The mobile ISP has an account relationship with the mobile subscriber. But many other companies have similar account relationships for one or a few devices: e-commerce sites (purchases through credit cards linked to each device used to purchase), game companies (subscriber relationships with the game console), social networks, cross-context companies with suites of services where a user logs in once, and many more.

In this emerging ecosystem, ISPs become a source for one part of the device map for the growing number of devices the typical user employs. The role of ISPs as ISPs diminishes as a portion of a user's overall computing. Cross-device tracking companies work with numerous sources of information to gather and analyze data in creating the device map. ISPs are merely one source of data, and their subscriber relationships provide a limited and diminishing portion of a user's Internet activity history.

¹⁰ Broadband Decisions: What Drives Consumers to Switch-or Stick With-Their Broadband Internet Provider, *Federal Communications Commission*, Dec. 2010, (https://apps.fcc.gov/edocs_public/attachmatch/DOC-303264A1.pdf).

¹¹ "Annual Report and Analysis of Competitive Market Conditions With Respect to Mobile Wireless, Including Commercial Mobile Services: Fifteenth Report, *Federal Communications Commission*, June 27, 2011, (https://apps.fcc.gov/edocs_public/attachmatch/FCC-11-103A1.pdf).

¹² Horace Dediu, "Measuring Mobile Platform Churn in the US Market, *Asymco*, July 2011, (<http://www.asymco.com/2011/07/12/measuring-mobile-platform-churn-in-the-us-market/>).

Conclusion

A Working Paper of
**The Institute for
Information
Security & Privacy**
at Georgia Tech

— May 2016 —

Chapter 10: Conclusion

This Working Paper has provided a detailed, factual description of today's Internet ecosystem in the United States, with attention to user privacy and the data collection about individuals. The ecosystem described in this Working Paper is important to ongoing policy discussions in Congress, federal agencies including the Federal Communications Commission ("FCC") and the Federal Trade Commission ("FTC"), and elsewhere. This Working Paper is intended to help provide a factual basis for making public policy decisions about the privacy framework that should apply to Internet Service Providers ("ISPs") and other companies that collect and use consumers' online data.

Diagram E-1, shown in the Executive Summary, shows a funnel for what information is available about user activity going forward for ISPs. At the top are the multiple contexts discussed in this Working Paper, where different players in the online ecosystem see detailed URLs and content about user activity. Due to pervasive encryption, virtual private networks ("VPNs"), and the other developments discussed here, technology often blocks ISPs' access to user traffic. Next, users are shifting to multiple devices and ISPs, so an ISP's connection to any one device is far less than complete, especially in the Internet of Things world we are rapidly entering. Finally, especially as WiFi hotspots become the majority of traffic, any one ISP only sees a fraction of the activity on any one device.

In light of these facts, as shown in previous Chapters, the evidence does not support a claim that ISPs have "comprehensive" knowledge about their subscribers' Internet activity, for encryption and other technological reasons. Similarly, ISPs lack "unique" insight into users' activity, given the many contexts where other players in the ecosystem gain insight but ISPs do not, and the leading role in cross-context and cross-device tracking played by non-ISPs.

This Working Paper takes no position on what rules should apply to ISPs, or to providers of services in the other contexts (often called "edge providers"). However, public policy should be consistent and based on an accurate understanding of the facts. We hope this Working Paper has provided a useful contribution to that understanding.

List of Acronyms

CAIDA	Center for Applied Internet Data Analysis
CDN	Content Delivery Network
CPNI	Customer Proprietary Network Information
DAA	Digital Advertising Alliance
DNS	Domain Name System
DPI	Deep Packet Inspection
DSP	Demand Side Platforms
FCC	Federal Communications Commission
FQDN	Fully Qualified Domain Name
GN	Google Now
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IAB	Interactive Advertising Bureau
IBA	Interest-Based Advertising
IETF	Internet Engineering Task Force
IP	Internet Protocol
IPSEC	Internet Protocol Security
ISP	Internet Service Provider
NAI	Network Advertising Initiative
NCMEC	National Center for Missing and Exploited Children
OBA	Online Behavioral Advertising

OS	Operating System
PII	Personally Identifiable Information
PKI	Public Key Infrastructure
ROI	Return on Investment
SMTP	Simple Mail Transfer Protocol
SMTPS	Simple Mail Transfer Protocol Secure
SDK	System Developer Kit
SSID	Service Set Identifier
SSP	Supply Side Platforms
TLS	Transfer Layer Security
TTL	Time to Live
UIDH	Unique Identifier Header
VPN	Virtual Private Network

Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others

This addendum describes the comments received to date on the February 29 version of the Working Paper, and lists the edits to the report in Version 2.0, dated May 2016.

Response to Comments

Since the release of the February 29, 2016 version of this working paper, we have received two sets of comments. First, Professor Nick Feamster of Princeton University published a letter to the Federal Communications Commission and a blog post to *Freedom to Tinker*. After email exchanges to clarify his comments, Professor Feamster authorized us to post the following statement that he made to us: “Upon more careful review of the paper, I have not found anything in the report that I believe is incorrect. I continue to believe that there are important additional facts that should be considered by policymakers, which were not discussed by the paper.” Our detailed response to Professor Feamster is attached and posted below.

Second, Aaron Rieke, David Robinson, and Harlan Yu authored a report for Upturn called “What ISPs Can See: Clarifying the Technical Landscape of the Broadband Privacy Debate.” Like our Working Paper, the Upturn report provides facts and does not make policy recommendations. Their report said that our Working Paper was “technically accurate in most of its particulars.” It also said our Working Paper “could leave readers with some mistaken impressions about what broadband ISPs could see” and offered their report as a “complement” to the Working Paper. Below, we remove one example (in our 120-page report), and thank the Upturn authors for their input.

The Upturn report highlights four observations:

1. **“Truly pervasive encryption on the Internet is still a long way off.”** Our Working Paper’s most-cited finding is that the use of HTTPS has risen sharply from 13% of Internet backbone traffic in April 2014 to 49% in February 2016. We expect the fraction of encrypted traffic to continue to rise, for reasons stated in the Working Paper, but offer no statistics beyond an estimate of 60% of encrypted traffic by the end of 2016.
2. **“Even with HTTPS, ISPs can still see the domains that their subscribers visit.”** We agree, and show that in text and diagrams in the Working Paper. As noted in the Working Paper, operating systems can see all the domains that the subscribers visit, with broad access available to other entities such as advertising and social networks.
3. **“Encrypted Internet traffic can still be surprisingly revealing.”** We agree that network operators in many instances can see features of network traffic beyond the domain name, such as “the size, timing, and destination of the packets” that the Upturn report highlights. In terms of that being “surprisingly revealing,” we offer three observations. First, ISP visibility into this information is far from complete, due to factors such as local caching and content delivery networks. Second, to date, online advertising efforts have focused on the content and detailed URLs visited by a user, not on data elements such as the size, timing, and destination of packets. We have not seen evidence that the latter are rich data sources for online advertising purposes. Third, the content and detailed URLs visited by a user are always available to others in the online advertising ecosystem, through the web browser and operating system, as well as the site visited by the user (and others in the online advertising ecosystem who may have access to it). Therefore, online data collection and use by edge providers would likely remain the focus of online advertising activity regardless of what legal rules apply to ISPs.

4. **“VPNs are poorly adopted, and provide incomplete protection.”** Concerning adoption, we cite current statistics showing modest adoption in the U.S., and describe VPN offerings from Facebook and Google that may increase the level of adoption going forward. A more recent development is that Opera has announced a free VPN feature for its desktop web browser and iOS mobile web browser. <http://www.opera.com/mobile/vpn> (“With Opera VPN you can block ads, change your virtual location and stop sites from tracking you around the web.”) The Google and Opera VPNs are very easy to enable – the user simply toggles on the capability, which is built into the browser. Concerning scope of protection, in our response to Professor Feamster, we agreed that VPNs can be configured in different ways, some of which do not mask the domain name from network operators. We note that the new offerings from Facebook and Google (and now Opera) do appear to provide “complete protection,” in the sense of blocking domain name as well as content and detailed URLs, although we do not know the extent of blocking of data elements such as the size, timing, and destination of packets.

In summary on the Upturn four main assertions, there is considerable convergence on the underlying facts. We note that Upturn’s analysis does not include any discussion of how web browsers, operating systems and other edge providers have the same or greater visibility into web browsing data, which is covered extensively in the Working Paper. This comparative analysis is relevant to any full assessment of how consumer online data is collected and used.

Changes to text since February 29 version

1. Edits to p.15: Appendix 1, Key Terms

Under first section titled Availability and Use: Remove the last two sentences: “For example, an Internet Service Provider’s (“ISP’s”) users may conduct many billions of Domain Name System (“DNS”) lookups each day. While theoretically this data could be collected and used, it appears to be impractical and cost-prohibitive to do so today.”

Comment: We removed this example from Appendix 1. After the Upturn report was released, we interviewed multiple ISPs about the practicality of collecting customer DNS lookups. While some interviews agreed with the statement in the original text about whether such actions are “impractical and cost-prohibitive,” others disagreed. As a result, we no longer cite the example of collecting and using data about DNS look up.

Concerning one topic discussed in the Upturn report, researchers have reported about ISP scanning of DNS look-ups for access by users to domain names indicative of malware infections. One ISP we interviewed indicated that it is relatively easy to log customer IP addresses that match well-known malware URLs. The ISP noted that such logging is very different in scale from logging all DNS look-ups.

Consistent with the general trend toward encryption discussed in our Working Paper, the IETF has created the DNS Private Exchange Working Group to develop mechanisms to provide confidentiality for DNS transactions. <https://datatracker.ietf.org/wg/dprive/charter/>. Availability of DNS lookups may therefore decline in the future.

2. Edits to p.31 – 2. Virtual Private Networks (VPNs)

Add footnote to end of sentence “Where VPNs are in place . . . domain name the user visits.”

FN: As noted in the Upturn report, a properly configured VPN will mask DNS queries, though not all VPNs are so configured (*What ISPs Can See*, <https://www.teamupturn.com/reports/2016/what-isps-can-see>). For example, Cisco’s support for its VPN client details the multiple ways DNS lookups can operate, depending on multiple factors. <https://www.teamupturn.com/reports/2016/what-isps-can-see>. While this is true, the remainder of this section will describe how a VPN properly configured to mask DNS lookup will operate.

3. **Edits to p.27: First sentence “The user’s computer next sends its request...”**

Add footnote in 4th paragraph (not including the numbered list) directly after the phrase “the ISP can see only the host name”:

FN: The ISPs can in many cases also see other information such as the size, timing, and destination of packets. The commercial usefulness of this data will depend on how useful such data is compared to going to the website visited by the user (or others in the ecosystem who may have access to it), which will have more complete information about the content and detailed URLs that the user has accessed.

4. **Various non-substantive typographical errors were corrected. Pagination remains the same as in Version 1.0.**

March 6, 2016

Addendum: Online Privacy and ISPs

After reading the Working Paper, Princeton University Professor Nick Feamster recently sent a [letter](#) to the FCC and made a subsequent [blog post](#) in Freedom to Tinker (FTT). Professor Swire's initial response is available [here](#). On the blog post, the authors are on board in many ways. Our report was over 120 pages. With that said, our Working Paper necessarily did not cover many topics that may be relevant to the FCC's consideration of broadband privacy. It is notably difficult to explain these topics in a simple and clear way that informs a non-expert audience while being precise enough to satisfy an expert audience.

This addendum addresses Prof. Feamster's blog post in some detail. It then, based on discussions between Professors Feamster and Swire, briefly summarizes some of the views of Prof. Feamster and the Working Paper authors. Based on these email exchanges, our areas of agreement about the facts are considerably greater than first appeared.

To begin, we agree with the facts for the three things Prof. Feamster calls "missing pieces" in our Working Paper:

1. "A single ISP can still track significant user activities from home network traffic and (as the user moves) through WiFi sharing services."
2. "ISPs can observe user activity based on general traffic patterns (e.g., volumes), unencrypted portions of communication, and the large number of in-home devices that do not encrypt traffic."
3. "DNS traffic sometimes goes to the ISP's DNS server after it exits the VPN tunnel. Configuring certain devices to use VPNs may not be straightforward for many users."

In terms of our Working Paper, we will discuss each of these three points.

Home and WiFi networks. First, our Working Paper, at page 25, makes the same point about how a single ISP may track an individual through WiFi sharing services: "The episodic glimpses are more continuous when the user switches from one connection (such as home) with a particular ISP, and goes to another connection (such as a WiFi hotspot) with the same ISP."

Concerning home usage, we cite to statistics about the relative shift over time from individual user activity on a single home device to multiple devices, which are often mobile. We suspect Prof. Feamster would agree that a much larger proportion of the typical user's Internet

activity today is mobile and out of the home than was true in the early days of the Internet, including sometimes (but sometimes not) hopping to a different ISP at a Wi-Fi hotspot. For what it is worth, we looked for but did not find useful statistics about how often a WiFi hotspot is the same ISP as home subscription.

Pervasive encryption. Second, for the visibility of an ISP into encrypted traffic, Prof. Feamster seems to agree with two things we worked hard on in the Working Paper: (a) statistics showing a large rise in the share of HTTPS traffic in the past two years; and (b) our Diagram 1-A at page 26 showing what is visible to an ISP with the shift to HTTPS. In Prof. Swire's experience, many non-experts had not previously had a grasp of Diagram 1-A; that diagram speaks directly to often-voiced concerns about comprehensive Deep Packet Inspection, showing why content is blocked by properly-deployed HTTPS. On the prevalence of encryption, Prof. Swire personally finds the most interesting fact in the report the rise of HTTPS traffic from 13% two years ago to 49% today (those statistics come from one data source, but we have found no reason to believe other sources tell a different story).

Prof. Feamster states that ISPs can observe general traffic patterns, such as volume, as well as unencrypted traffic, including for a growing number of home devices. We agree entirely, and nothing in our report states otherwise:

- We repeatedly explain that unencrypted information allows full visibility for an ISP, including in Diagram 1-A.
- Part of the paper's research included a tutorial on Wireshark from someone at Georgia Tech who is proficient in networking, which provided clear visualizations about how number of bits, session length, and other general data is visible to the ISP. Prof. Swire's experience and research is that these sources of data are less useful for tracking and online advertising than content or detailed URLs. We would be interested if Prof. Feamster or others could offer information about how these general sources of data in fact are useful for tracking and online advertising.
- We mention the Internet of Things briefly in the report. Prof. Swire taught Internet of Things cybersecurity in class this week, and we share Prof. Feamster's concern about the lack of encryption and severe lack of overall security for many Internet of Things devices and services. We hope a wide range of technology, policy, and business experts find ways to improve cybersecurity in that realm, and we can note that in the Working Paper. For now, Internet of Things is one portion of home use, which is one portion of an individual's overall Internet-connected activity.

The rising prevalence of encryption (which Prof. Swire has often supported in his writing) is the single strongest basis for one claim of the Working Paper – ISP access to a user's Internet activity (notably including content and deep links) is not “comprehensive” today. Today, content and deep links are blocked for roughly half of traffic, and we expect that fraction to rise. Simply put, ISPs can have “a lot” of visibility into user activity, but not “comprehensive” visibility.

Based on exchanges between Professors Feamster and Swire this weekend, Prof. Feamster has authorized us to say that he agrees with that last sentence. The genesis of our paper was the FCC public workshop last April, when some speakers claimed “comprehensive” ISP access and others denied it. We have hoped that a public service of the paper has been to

disprove the “comprehensive” description. Whatever the FCC may decide to do, we think it should base its approach on actual visibility rather than “comprehensive” visibility. For instance, if the FCC had mistakenly based a rulemaking on “comprehensive” ISP liability, the factual predicate for its action would have been subject to severe and accurate criticism.

VPNs. Third, concerning VPNs, we carefully created three diagrams and a good deal of text to explain how VPNs work, for a non-expert audience. It seems that Prof. Feamster agrees in most or all details with this description. One point Prof. Feamster makes is that VPNs can be configured so that the ISPs’ DNS server does in fact see DNS look-up information even though a VPN is in use. That configuration is different than what we learned in briefings about how VPNs operate. We would be glad to get helpful citations from Prof. Feamster and clarify that for the Working Paper how configurations differ, including statistics on what is common.

For VPNs Prof. Feamster also said: “Configuring certain devices to use VPNs may not be straightforward for many users.” Once again, we did not say anything to the contrary. It is true that VPNs have the capability to block even the host name from the ISP, so the destination of a user’s web activity can be blocked from the ISP. Our own research suggested that VPN use by individuals has not been especially common in the US, so we did not emphasize this technical limit on ISP visibility nearly as much as the pervasive encryption point. On the other hand, the new services by Facebook and Google may change that in the not-so-distant future. VPNs and proxy servers are a type of limit on ISP visibility.

Conclusion. Prof. Feamster published the FTT blog post after the FCC letter. As we think the above discussion shows, the Working Paper is broadly consistent with the three “missing pieces” Prof. Feamster highlights; in a number of instances, the Working Paper specifically makes points that Prof. Feamster does.

Discussions between Professors Feamster and Swire this past weekend, however, also discussed statements in the FCC letter about “many technical inaccuracies” and “basic misunderstandings of various Internet technologies.” Based on that email exchange this weekend, Prof. Feamster has authorized us to say: “Upon more careful review of the paper, I have not found anything in the report that I believe is incorrect. I continue to believe that there are important additional facts that should be considered by policymakers, which were not discussed by the paper.”