Categorizing Uses of Communications Metadata: Systematizing Knowledge and Presenting a Path for Privacy

Susan Landau
The Fletcher School and School of Engineering
Tufts University
Medford, USA
susan.landau@tufts.edu

ABSTRACT

Communications metadata can be used to determine a communication's device, identify the user of the device, and profile the user's personality and behavior. The current state of affairs is that the increase of attacks against user privacy based on using communications metadata vastly outpaces the ability of users to protect themselves. With few exceptions, protections are point solutions against a specific attack. In the current situation, the user loses.

This paper is an initial step in a multi-step research effort to reset that balance. The main contribution of this paper is a categorization of the uses of communications metadata based on their privacy impact. Because of the technical complexity of the problem, including the wide variety of electronic communications, technology can only go so far in providing solutions to the privacy problems created by the use of communications metadata. Legal and policy intervention will also be needed. This categorization is intended to provide a start in developing legal and policy privacy protections for communications metadata. Along the way, I also provide an explanation for how it is that communications metadata has become so valuable, sometimes surpassing the value of content. This work provides both an intellectual framework for thinking about the privacy implications of the use of communications metadata and a roadmap, with first steps taken, for providing privacy protections for users of electronic communications.

CCS CONCEPTS

• **Security and Privacy** → Human and societal aspects of security and privacy; • **Applied Computing** → Law, social and behavioral sciences.

ACM Reference Format:
Susan Landau. 2020. Categorizing Uses of Communications Metadata: Systematizing Knowledge and Presenting a Path for Privacy. In *New Security Paradigms Workshop 2020 (NSPW '20), October 26–29, 2020, Online, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3442167.3442171

1 INTRODUCTION

When Edward Snowden revealed that the NSA had been collecting call detail records (CDRs) in bulk, [74] Americans were shocked and angry. Playing down the surveillance impact, President Obama told the nation, "Nobody is listening to your telephone calls," [133]. He added, "What the intelligence community is doing is looking at phone numbers and durations of calls." [133] The statements were accurate, but the public understood what the president had avoided saying[1]: such surveillance was an invasion of privacy. Public reaction against the bulk metadata collection was strong and sustained.

That communications metadata can reveal valuable information is well known. Studying the pattern of communications yields vast amounts of information, and for over a century military forces have relied on communications metadata to decode their enemies' tactics. In recent decades such information has become a standard tool for investigating criminal activities. As I wrote in *Listening In* [108],

> Communications metadata can show the underlying structure of criminal and terrorist conspiracies. This metadata is everywhere: in the bits in the cell towers that say this phone was in this vicinity at this time, and in routers that say an email was sent at this moment from this physical vicinity. Even negative metadata—for instance, that a phone was turned off in a given vicinity—can benefit investigators. In France, police have used information on when and where phones are turned off to find criminals using stolen credit cards [73]. Patterns that show pairs of phones that trade off—one working only when the other is not—can highlight the presence of terrorists or drug dealers.

Over the last decade and a half, communications metadata from phone calls, emails, IoT devices, and the like have become increasingly valuable to the private sector. Legal protections for communications metadata are sparse. On the commercial side this is because users have "provided consent" to the use of the data—without understanding the full set of purposes to which it is being put. On the U.S. government side this is due to a legal history resting on the idea that data shared with a third party—and most communications metadata is shared with a communications carrier of one type or another—does not carry with it a presumption of privacy (in earlier work, my coauthors and I presented arguments, from a technical viewpoint, why that rationale no longer makes sense for IP-based communications [17]). And it is in part because the increased value of metadata has snuck up quickly and law has simply not kept up.

The move from telephony to IP-based communications provides vastly more and far richer metadata to examine. Digitization and

[1]The NSA knew full well the value of such metadata; as former NSA Director Michael Hayden noted in 2014, "We kill people based on metadata." [81]

NSPW '20, October 26–29, 2020, Online, USA

ACM ISBN 978-1-4503-8995-2/20/10.
https://doi.org/10.1145/3442167.3442171

use of the Internet makes it easier to collect and analyze this data. Thus the privacy impact stemming from the use of communications metadata is far greater than in the days when such metadata consisted of CDRs. Communications metadata can be used to determine a communication's device, identify the user of the device, and profile the user's personality and behavior. The current state of affairs is that the increase of attacks against user privacy based on using communications metadata vastly outpaces the ability of users to protect themselves. With few exceptions, protections are point solutions against a specific attack (Tor is an exception). In the current situation, the user loses.

To achieve privacy protections, we need to solve such problems as: When might such intrusions be thwarted through technical means? How will these work? Who—the manufacturer of a device, the developer of a protocol, the user of a system—will be in a position to institute such protections? When are technical solutions likely to fail? What legal and policy protections are appropriate?

Various scholars have tackled these problem in different ways. Technological solutions have so far failed to gain traction. Many researchers present point solutions to particular attacks or classes of attacks using adding noise, routing through VPNs, etc. (see, e.g., [8, 113]), while Tor provides a general solution for a class of problems: web connections. Tor is a good solution to the privacy problem of web accesses, but we have not seen other solutions that provide metadata protections to a similarly wide class of communications. This should not be completely surprising; wide protections for a large class of communications are hard to develop, and perhaps even more difficult to achieve adoption.

Meanwhile both legal scholars and the courts have tackled questions regarding under what legal rules law enforcement should have access to location information and post-cut-through digits (the latter are numbers dialed after the phone number; these might be information to the other end, such as a bank account or prescription number, or it might be a way of dialing a particular number). The issue of what other types of metadata law enforcement might have access to and under what circumstances, has received less attention.

I believe that technical protections will prove insufficient and that we will need legal and policy protections. To do so effectively we need a better understanding of the threats the use of communications metadata poses to privacy. Essentially what we need is a categorization of the types of threats posed.

I am not the first to undertake such a categorization. The Data for Development (D4D) challenge launched by the telecommunications company Orange in 2012, which provided "anonymized" CDR datasets from the Cote d'Ivoire to interested researchers, challenging them to to find solutions that would help alleviate poverty, improve health, and otherwise aid development, resulted in an explosion of research [24]. This successful effort was followed by a similar challenge a year later with data from Sengal supplied by Orange and Sonatel and with funding from several foundations. The results, many of which I discuss here, were rich and fruitful. In 2014 and 2015, Telecom Italia inititated a big-data challenge that included an aggregation of telecommunications, weather, news, social networks and electricity data from Milan and the province of Trentino. I include some of the results from that challenge here as well.

There have already been a number of useful survey articles on different types of uses of such communications metadata. Calabrese et al. surveyed how mobile network data can illuminate the action of populations within cities [34]. Blondel et al. "survey[ed] the contributions made ... on the social networks that can be constructed with [massive sets of CDRs]" [25], while Daubert et al. proposed several categories of privacy leakage for IoT devices (theirs, which is not specifically focused on communications metadata, covers narrower categories than the ones I propose) [52]. My ultimate focus is in framing the privacy issues resulting from the use of communications metadata in a way that legal and policy frameworks can use for privacy protection. Thus my focus is on classifying the privacy risks that stem from all forms of communications metadata. This was not a direction of earlier works.

I start by looking at how and why communications metadata has recently become so valuable. Next, I examine who creates metadata (it is not just the user), what the threats are, and what types of communications metadata are used. I present the categorizations, and then briefly discuss the "difference" between metadata and content (spoiler alert: where to draw the line between metadata and content must be somewhat arbitrary). I conclude with possible approaches for developing privacy protections.

The literature in this area has grown so vast that it is not possible to fully cover it. I used the following sampling technique. I started by studying the papers in the major security and privacy conferences (IEEE Security and Privacy, ACM Computer and Communication Security, USENIX Security, Privacy Enhancing Technologies Symposium) as well as NetMob for work in this area. I also studied the papers produced by the Orange Telecom Data for Development Challenge and the Telecom Italia effort.[2] From these initial papers, I followed trails of research I deemed most interesting both backwards and forwards in time. I also studied the legal literature in this area for related work. I have undoubtedly missed some relevant work, but I have tried hard to cover the main threads in this area.

2 WHY HAS METADATA BECOME SO VALUABLE?

As technology changed, the meaning of communications metadata changed as well. In this section, I briefly discuss the meaning of metadata and the history of the collection of communications metadata. I largely focus on the U.S.

2.1 The Meaning of "Metadata"

The term "metadata" dates from the late 1960s, but the idea is as least as old as the first library catalogue [141, p. 6]. Metadata is the "structured description of the essential attributes of an object," [69] But one person's metadata is another person's data. Yo, the April's Fools Day mobile phone app that simply says "Yo," was repurposed in Israel to say "Hodor" to warn of missile strikes—a change from metadata (notification of a communication) to an actual communication [14].[3]

What constitutes communications metadata has never been clearly legally defined. Instead it has followed technology. In the U.S., the 1979 Supreme Court decision on *Smith v. Maryland,* ruled

[2] Many of these papers appeared in [24] and other sources.
[3] Many felt the value of the communication was quite limited [14].

that call metadata—caller and callee phone number and time and duration of call, the data shared with the service provider in order to make the call—was not subject to the same legal protections as the communications content. As we later learned from the Snowden disclosures, the U.S. government expanded this definition of telephone metadata to include "comprehensive communications routing information, including but not limited to session identifying information (e.g., originating and terminating telephone number International Mobile Subscriber Identity (IMSI) number, International Mobile station Equipment Identity (IMEI) number, etc.), trunk identifier, telephone calling card numbers, and time and duration of call" [158]. Proposed regulations to update the EU ePrivacy Directive define electronic communications metadata as "data processed in an electronic communications network for the purposes of transmitting, distributing or exchanging electronic communications content; including data used to trace and identify the source and destination of a communication, data on the location of the device generated in the context of providing electronic communications services, and the date, time, duration and the type of communication" [62, p. 25].[4] Yet none of these definitions fully capture the wealth of non-content information that is transmitted for the purpose of enabling the communication.

The line between metadata and data is unclear; it really depends on vantage point [17, 28, 30, 95, 96]. Looking at a number of different fields and examples, Borgman et al. and, in greater detail, Borgman explore how the context surrounding the acquisition of scientific and technical data, information that often lies or can be derived from the metadata, can greatly inform conclusions about the data itself; this work focuses on scientific and technical research [28, 30]. But the same lack of clear distinction between metadata and content also holds true for communications.

In writing about a court case involving third-party cookie tracking, Kerr observed, "the line between contents and metadata is not abstract but contextual with respect to each communication" [95]. Kerr explained this with a low-tech example:

> Imagine a telephone call back in the early days when there were human operators at the switchboard. You would call the operator and say, 'Please connect me to Pennsylvania 6-5000.' The operator would then plug your line into the switch for Pennsylvania 6-5000, setting up a call between you and the party at that number.
>
> Now ask the question, is 'Pennsylvania 6-5000' contents or metadata? I think the answer is, well, both. Your speaking of the number was part of the contents of a communication between you and the telephone operator. On the other hand, a telephone company record that you had called Pennsylvania 6-5000 is metadata with respect to the call that was subsequently placed. Whether the information was contents or metadata depends on whether you are drawing the contents/metadata distinction with respect to the first leg of the communication (you speaking to the operator) or the second leg of the communication (you speaking with whoever answers at Pennsylvania 6-5000). [95]

Bellovin, Blaze, Pell, and I examined the distinctions between communications content and non-content in the context of IP communications [17], whose protocols are significantly more complex than that of telephone metadata. We observed that the Internet communications environment creates a situation in which whether an individual unit of data is content or non-content varies as it transits the Internet's layered structure from sender to recipient [17]. What role a piece of data had depends on where that data is when the question is being asked—and which protocol is querying. Introducing the idea of "architectural content": data that is unexamined while being transported solely within the network (that is, within points that are neither the sender nor receiver), we showed that architectural metadata at one layer of the network stack (e.g., port number) can be architectural content at another [17]. Which it is depends on the user's vantage point. In other words, for Internet communications the content/non-content distinction is an artifact of the moment and place in which the query is being made [17, 1].

This current work builds on [17] but studies a somewhat different question. Here I categorize how data that is not explicitly communications content discerns personal information about the user or their device, a question that cuts directly to the heart of user privacy. In the absence of clear definitions of communications metadata, I use the term to denote the non-content data within a communication, understanding, however, that content can sometimes be inferred from the non-content part of a communication [168].

2.2 Recent Changes in What Constitutes Communications Metadata

The widespread use of communications metadata by government and the private sector has slipped in silently, like the fog. NSA signals intelligence had always been about content: acquiring communications, decrypting communications, learning the key, breaking the encryption system. Then in the mid-late 1990s the result of the volume, variety, and velocity of communications, all of which had vastly increased as a result of the Internet, and the increasingly broad use of encryption, NSA began losing its ability to listen to content [83]. By the late 1990s, it was not just technologically sophisticated nations that were using strong encryption, so were most states. The agency was "Going Deaf" [83].

The changes in technology—mobile devices, IP-based communications—affected not just traditional NSA targets, but virtually everyone. The importance of non-state actors to U.S. national-security interests became clear with the 2000 bombing of the USS Cole and the 9/11 attacks. Immediately after the latter, President Bush authorized the bulk collection of domestic CDRs; this later came under the control of the Foreign Intelligence Surveillance Court. It was only with the Snowden disclosures that the government's bulk collection of CDRs became widely known. Public reaction led to the 2015 passage of the *USA FREEDOM Act*, which permitted government access to CDRs but ended bulk collection. However, by that time, the value of the program had waned [109].

[4] The 2002 ePrivacy Directive did not address communications metadata; the 2009 update had but one mention, which allowed a user or subscriber to withdraw permission for the electronic communications provider to process traffic data for marketing purposes [63, Article 3,Security of Processing, section 6].

What was less well known is the extent to which the U.S. military had grown dependent on communications metadata. Where once the NSA sought to tap wires and eavesdrop on satellite signals to find out what was occurring, by the mid 2000s, the agency was employing techniques, including drones, to determine who was talking to whom. In Iraq, in Afghanistan, everywhere it was listening in, the agency was mapping networks—for that was often far more valuable than learning what it was someone said to someone else [80]. The U.S. was not the only nation employing such tools; Italian authorities investigating the abduction of a cleric from the streets of Rome used CDRs to tie the kidnappers to the CIA [46]. Metadata had trumped content.

It is not just signals-intelligence agencies that consider communications metadata to be the new gold. New types of services, including those supplied by IoT, and new sources that provide vast amounts of data (e.g., surveillance footage) mean that the volume of communications metadata is increasing at an astounding rate. Communications metadata connected with IoT devices reveals peoples' presence in the home [135] and the existence of troops on a secret army base [82]. Even encrypted IoT traffic can reveal information, e.g., the hours someone is sleeping, when a home occupant is interacting with an intelligent personal assistant, etc. [7].

IP packet data is much more revelatory than the comparable CDR information. While one reason is that IP packets carry more information than CDRs—patterns of packet sizes, for example, may reveal which language a user is speaking [168]—in large part, however, the revelatory nature of IP packets stems from the fact that IP communications are of a much greater variety than telephone calls and phone texts (the communications that generate CDRs). A 2012 patent application sought to deliver different types of ads to TV set-top boxes based on the ambient noise in the room (children playing versus a romantic encounter) [156]. The source address of the IP packets delivered to that TV set top box would reveal the type of activity occurring in the room [17]. The availability of this private information is a result of the ad ecosystem in IP-based content delivery; the phone network has no equivalent.

The transformation in communications involved more changes than simply from telephones to IP-based communications. As communications changed from landline telephones to IP-based communications, then cellphones, then smartphones, the ratio of bits of content to bits of metadata changed. In 2006, the average cell call was 3.03 minutes long in the U.S.; by 2011, it was 1.78 minutes [20]. In either case, the CDR for such a call would constitute less than 1% of the call's bits. By contrast, already in 2010 the metadata for a tweet was enormously large; at that time it had thirty items including the tweet's unique ID, creation data, the ID of the tweet, screen name, and user ID to which it was replying, the author's user name, screen name, biography, and url, whether the user is "protected," (an account is "protected" if only users approved by the owner follow it [155]), the number of users the user is following, the number of Twitter lists on which the author of a Tweet appears, number of favorites this user has, ... [101]—far more, and far more informative, than what is contained in a CDR. This comparison is not quite apples to oranges in that the telephone service provider holds other records about the subscriber than the CDR.

Finally, the fact that we can search metadata far more efficiently than in decades past makes it much easier to use. Patrick Fitzgerald, who as an Assistant U.S. District Attorney investigated the 1993 World Trade Center bombing, had to sift through several books of outputs of CDRs in order to show that plotters had indeed been connecting with one another [66]. Once data was digitized, such searches became both faster and more effective. Not only are simple searches—did Party A speak with Party B before he bought explosive detonators? and then again before he bought nitrate?—significantly easier but so are far more complex ones—find groups of size four or less who communicate solely with each other. Machine learning techniques have used metadata successfully in other domains (information retrieval and information processing) to speed automatic searches. For example, Hu et al. used font size to do automatic extraction of titles from general documents [84]); it is likely that ML techniques may exhibit the same success with communications metadata.

In short, there is more communications metadata, it is far richer than previous forms of communications metadata, and this data is far easier to search than it was several decades ago. This creates a perfect storm for privacy, one that is increasingly being exploited. That is the subject to which I will shortly turn, but first I'll examine who creates metadata and what the threats are.

3 WHO CREATES METADATA?

It would appear that since communications metadata stems from the user's communications, the user is the creator of her communications metadata. The situation is more complex. Communications metadata is usually created by the user, but it can also be created through actions of the communications providers or the communications recipient. We'll look at each of those in turn.

Communications protocols determine which data is received from the user and what type of format. One of the oddities of communications protocols means that sometimes the user not only does not participate in creating the metadata from their communication, they may not even be able to determine what the communications metdata is. A striking example of this is in the setup step of mobile communications. This particular issue may be the cause behind the "technical irregularities" NSA found in its collection of domestic CDRs under the USA FREEDOM Act [109].

As we explain in [109], "A mobile phone has two addresses: the MSISDN and the IMSI number. The MSISDN is the external identifier, the number a user gives when we ask for their cell number. The IMSI is the phone's identity on its "home" network and is not shared except when the phone roams." When the phone roams—and mobile phones do—there are additional numbers used to identify the phone, including a Temporary Mobile Subscriber Identity (TMSI); this is provided to the phone by the network in which the phone is roaming in order to prevent eavesdroppers from tracking the phone. There is also a Mobile Station Roaming Number (MSRN), which is provided for the purpose of setting up a call. Sometimes the records at the two Mobile Switching Stations involved in call setup (the caller's and the callee's), store different phone numbers for the same call: one has records in terms of the callee's MSISDN or MSRN, while the other, the callee's IMSI [109]. That these records differ is usually unknown to the people on the call (indeed, NSA apparently did not originally account for this issue). Even if the

idea is known, the callers do not have access to the MSRN used to set up the call; that information is not shared with the callers.

There are also complexities in communications protocols that prevent a communicator from knowing the metadata around their communication. How a protocol is implemented can determine whether information is content or metadata. In [17], my colleagues and I observed that the two common ways to retrieve a webpage, GET and POST, result in different ways of including the query information in the communication [17]. A GET command puts the query information into the redirecting url, but a POST command places the query into the message body. Thus under a GET command, the query information is metadata, but under a POST, it is content. As my coauthors and I noted in earlier work, "The user, however, has no control as to whether GET or POST is used—and indeed, almost certainly cannot even discover which command has been issued." [17, p. 68].

Communications metadata can also be created in transit. To monitor traffic flow, network administrators need to be able to measure the data coming in and exiting, and some information about what it's doing. There are a number of network flow tools, of which Cisco Netflow was the first; it collects and summarizes source and destination IP addresses, source and destination ports, and transport layer protocol (TCP, UDP) [44]. This type of network flow information is used to provide situational awareness, address and manage security incidents, and search for threats on a network [100, p. 12, pp. 51-52, pp. 59-60, pp. 80-81]. The collection—and thus the analysis—is based entirely on communication metadata; there is no peering into packet contents [100].

Finally, the communications recipient can transform data it has received and produce information not available to the sender. That is due to the receiver's ability to combine the communications metadata with other information. An example is aggregation of incoming packets to a website, which can reveal whether the site is under a denial-of-service attack.

Understanding where and how communications metadata is created provides an entry point into considerations about controlling use of communications metadata [105, 166], an idea to which we will return.

4 WHAT AND WHO THREATEN PRIVACY?

Communications metadata can be valuable to casual eavesdroppers, industry, including service providers and Internet companies, local, state, national and also foreign governments. I will comment briefly on each in turn.

Casual eavesdroppers are ubiquitous at cafes and airport lounges. More sophisticated ones use more permanent types of systems such as IMSI catchers and radio antennas. They are hard to foil, except by the use of temporary identifiers such as the TMSI and the one in the random Bluetooth identifier produced by the Google/Apple Exposure Identification system (GAEN) for COVID-19 [6]. But unlike with content, communications metadata is rarely useful unless captured in significant chunks. Casual eavesdroppers are unlikely to be able to do sufficiently significant capture of communications metadata to pose a serious privacy threat.

There are two different sets of industry players that have access to communications metadata: the service providers and the Internet companies. Both need the metadata in order to deliver the needed service of a webpage, an email, etc.

Sometimes service providers use communications metadata for other purposes. This can include recognizing P2P traffic [97] (there are more details in the discussion on Identifying Device Activity). But it can also include tracking users. Verizon Wireless was fined $1.35 million in 2016 for the supercookies it was using to track user visits to webpages [94].

Many Internet companies, however, freely make use of customer communications metadata for purposes other than the delivery of packets. Sometimes the purpose is greater usability: by showing how users respond to configurations, Internet companies are able to improve the usability of a smartphone screen or a webpage. But here, unlike the use of communications metadata for delivering a communication, there is tension: the use of the communications metadata for improving usability can be invasive. Indeed, depending how much user notification and choice there is regarding the reporting of user actions, users may feel spied upon rather than served.

Consider, for example, the actions described by Leith regarding browser collection of user metadata, "From a privacy perspective Microsoft Edge and Yandex are much more worrisome than the other browsers studied. Both send identifiers that are linked to the device hardware and so persist across fresh browser installs . . . Edge sends the hardware UUID of the device to Microsoft, a strong and enduring identifier that cannot be easily changed or deleted. Similarly, Yandex transmits a hash of the hardware serial number and MAC address to back end servers. As far as we can tell this behaviour cannot be disabled by users." [112].

Another concern is the length of time that the metadata is stored. The first purpose of the metadata—enabling a communications service—means that the metadata is necessary in the moment and unnecessary a short time afterwards (since the metadata provides useful debugging information if something goes wrong, it must be stored for some amount of time). For others purposes, such as improving usability and aiding in the planning of new products, metadata are useful in the aggregate. They split on the direct utility to the user—and thus on the user's sense of whether the reporting on usage is valuable to them or is viewed as spyware.

Telephone companies kept business records, including CDRs, for decades, using them for long-term planning including the development of new products. But there was a substantive difference in privacy risks to individuals between then and now; recall, for example, Patrick Fitzgerald's investigative efforts with the CDRs related to the first World Trade Center bombing [66], a quarter century ago. The computing capabilities of the time meant that the privacy risks were minimal in comparison to the threat posed by present-day storage and use.

Governments' use of communications metadata pose an even greater privacy threat to users. They may be able to access the information held by the private sector as well as information collected by law enforcement and national security. How large the threat is differs by whether the communications metadata is acquired by a domestic government, which can act against its own people, or a foreign state, whose actions will then be against the other nation—although sometimes these actors can also be working against individuals within a foreign state.

With this laying out of the action, I now turn to categorizing the different types of privacy threats posed by the use of communications metadata.

5 WHAT COMMUNICATIONS METADATA CAN REVEAL

No nations have yet put restrictions on the commercial use of communications metadata. Before that, I will briefly discuss the E.U. and U.S. situations.

5.1 A Brief Overview of E.U. and U.S. Law Regarding Communications Metadata

Although the E.U. General Data Protection Regulation (GDPR) has placed strong restrictions on data collection and retention, its impact on the use of communications metadata is minimal. GDPR protects the privacy of all personal data and appies to all data processing of EU persons (natural persons) who are within the European Union, but its restrictions do not come into effect unless personal data is involved. So application of GDPR depends on what constitutes personal data; this definition has depended on the courts.

The 2006 E.U. Data Retention Directive, passed in the wake of the 2001 terrorist attacks against the United Stataes, required that service providers store communications metadata for up to two years. In 2014 the directive was struck down by the Court of Justice of the European Union (CJEU) in *Digital Rights Ireland,* Joined Cases 293 & 594/12 (2013). CJEU's ruling was based on disproportionality: bulk collection of communications metadata versus actual need for bulk data in investigations. The Court's ruling showed that CJEU viewed access to communications content as involving privacy concerns, access to communications metadata did not [32]. Thus communications metadata did not fall under the umbrella of GDPR. Later, in a 2016 ruling, *Patrick Breyer v. Germany* [137], the CJEU concluded that IP addresses are personal information. Thus IP addresses are subject to GDPR protections. But the broader class of communciations metadata—anything from packet length (which can reveal words being spoken [168]) to timing of communciations (which can reveal the details of people's daily schedules [115])—does not currently fall within GDPR protections. The 2017 E.U. E-Privacy Directive does address privacy of communications metadata, but its focus is narrow, not taking in many of the new technologies that I discuss. As of this writing, the directive has not yet passed.

In the U.S. the legal distinction between the rules governing the collection of communications and communications metadata goes back half a century to the *Katz* and *Smith* decisions.

After decades of unsettled law concerning the legality of telephone wiretapping, in the U.S. in the 1967 case of *Katz v. United States,* 389 U.S. 347, the Supreme Court recognized Fourth Amendment protections for the the content of telephone communications. The 1968 *Crime Control and Safe Streets Act* was passed, establishing the requirements for obtaining a federal wiretap warrant (the requirements for state warrants vary, but must be at least as strict as the federal requirements). A decade later, in *Smith v. Maryland,* 442 U.S. 735 (1979), the Court decided that as data shared with a third party, calling information, specifically the dialed digits of a phone number, did not warrant the same level of legal protections as communications content. At the time of *Smith,* phone calls were relatively simple: the phone number calling and the phone number called were call metadata and subject to the *Smith* ruling; the voice communication was call content and fell under the protections of *Katz.* But that simple situation was about to change.

The first technical change came about because of competitors to AT&T's long-distance service. Such companies couldn't directly offer long-distance service; that would change with the 1982 end of AT&T's monopoly. Instead the companies did so through having callers dial a toll-free number, then put in an account code followed by the actual number being called. Such post-cut-through digits confused the delineation between call metadata and call content. They were the number being called but conducted within a structure—the phone call itself—that had been traditionally seen as carrying call content.

Phone mobility added further complications: under what legal structure could law enforcement collect this information? In 2018, the Supreme Court ruled in *Carpenter v. United States* that police must obtain a warrant in order to obtain cell site location information.

This, then, is a brief summary of the current legal frameworks around communications content and metadata is two parts of the world. Let's turn now to how changing technologies may impact these.

5.2 Changing Technologies and Changing Communications Metadata

Katz and *Smith,* of course, applied to the Public Switched Telephone Network (PSTN); the Internet was not a public network at the time. Because the architecture of the two networks are so different, the laws and regulations that grew up around collecting call metadata largely fail to fit Internet communications (see, e.g., [17]). The PSTN was designed to transmit voice communications, and the engineering choices, from the centralized network to real-time delivery systems, were done to maximize voice quality. There were mulitple goals for the Internet, an important one being supporting multiple types of communications services [45]. As we know, this led to a very different architecture than the PSTN (Voice over IP, VoIP, in which voice communications are now carried over IP-based networks, further complicates matters). It is often said that the PSTN is made up of dumb terminals and a smart network, while the Internet is a dumb network with smart endpoints (see, e.g., [45, 151]), and there is a fair bit of truth in this decription.

Because the Internet has "smart" endpoints, IP-based communications have much richer metadata than those of the PSTN. This stems from two facts: (i) the endpoints' capability to support many types of services—far more than the PSTN's voice, fax, and SMS, and (ii) the computers that are the communications endpoints may have a wide variety of types of information to transmit.

There is another aspect of the Internet connecting smart endpoints–computers—that makes for richer communications metadata. Smartphones are an example of this; they increasingly have many embedded sensors; the metadata from these may be used to identify devices, people, and behavior traits. For example, a mobile phone can be recognized using the sensors' communication "fingerprints" [27, 85].

Mayernick and Acker observed that "If we are to engage in meaningful discussions about our digital traces, or make informed decisions about new policies and technologies, it is essential to develop theoretical and empirical frameworks for characterizing the role of metadata within networked communication infrastructures." [122]. If we are to understand the privacy and security risks[5] arising from the leakage of information from communications metadata, we of course need to make sense of what is being revealed. We need to categorize the various types of privacy and security risks arising from the use of communications metadata. I do so in the following subsections.

Now others have also examined the different ways that communication metadata can be revelatory. Ziegeldorf, Morchon, and Wehrle considered privacy threats emanating from IoT devices; their categorization includes identification, localization, profiling (including "lifecycle transitions"), and data linkage [169]. I look more broadly than IoT devices, and the set of classes is, I believe, the fullest to date of the types of information that communications metadata—communications writ broadly to include sensors and IoT devices—reveals.

Communications metadata is revelatory about societal groups as well as individuals. This delineation is not sharp, for information about a group also is likely—though not certain—to illuminate information about an individual within that group. Nonetheless, dividing information that is inferred into two broad categories—information about groups and information about individuals—is useful. For one thing, the information is used differently. For another, such information is often collected differently. And—perhaps most importantly—who collects the information is different. While almost all information about an individual is collected either by a company or the persons government, information about groups are often collected by other nations. This is Intelligence collection and is unlikely to be regulated by law. This categorization has a natural divide into revelations about group characteristics and revelations about individual ones.

5.3 Discerning Characteristics about Groups

Group characteristics include information about the structure of society and subsets of society. This information can be used in planning in various ways; much of it is of particular interest to government. For example, information about how organizations are structured can help enemy forces by revealing information about what is called "order of battle," the types of arrayed forces, their strength and command structure, their equipment. Other information, such as about how people move within a locale or a nation, can help a state govern by revealing commuting patterns or even potential patterns of introduction of illnesses, such as malaria [154, 160]. Over the last decade, due to a combination of the ubiquity of mobile phones and the available data sets, mobiles CDRs have been the single richest source for discerning characteristics about group behavior. In this subsection, I will discuss the following group characteristics revealed by communications metadata: order of battle, communities of interest, organizational structure, and societal characteristics and behavior.

[5]Information leakage tends to be a privacy threat, but it almost always also poses a security risk as well.

Illuminating Order of Battle

Knowing what forces an enemy has at hand and how they are deployed makes a critical difference in a battle's outcome. Prior to electronic communications, such information about an opposing force involved going behind enemy lines and was not easy to discover. But over the last century, the metadata from radio communications and, more lately networks, has made the "order of battle" much easier to obtain. Here I examine how accessability to this information has changed over that period.

The arrangement of troops prior to a military engagement makes all the difference between success and failure, and military history is full of examples where knowing the enemy's order of battle is crucial. Spies can provide such information. The advent of radio, which enables commanders to direct even while not right in the center of battle, also enables others to learn this information without putting themselves at risk; this includes the enemy. Encryption, of course, thwarts such eavesdropping, but not completely. Encryption prevents understanding the communications content, but communications metadata is very rarely protected.

The first apparent use of metadata to discern order of battle occurred in 1904 during naval warfare in the Russo-Japanese war. Japanese warships intercepted messages from the Russian czar's fleet, and although the Japanese hadn't decrypted the communications, interception of the traffic gave them sufficient information about the enemy's whereabouts to defeat them [110, p. 27] [111, pp. 21-22].

A decade later, the French made great use of traffic analysis on the Western front. As German troops crossed the border into France at the beginning of World War I, the French began listening in to German military radio communications. Using signal strength to determine distance, within two weeks the French mapped the location of the German stations [90, p. 300]. The French collected the names given to the stations—the "call signs"—which may reveal the military organization to which the station belongs, traffic volume, and senders [90, p. 300]. This enabled them to built a model of the deployment of the German troops. Such traffic analysis became a standard tool for the military.

In World War II traffic analysis was critical, enabling, for example, the U.S. military to track Japanese forces in the Pacific theater [90, p. 578] [29, pp. 18-19], in the deployment of the Third Army Corps in France post D-Day [29, pp. 25-30], and also warning German forces of various Allied efforts [29, pp. 32-35]. In the early days of the Vietnam War, tracking volume of radio traffic on a civil network in North Vietnam gave the U.S. several months notice of build up of North Vietnamese troops in South Vietnam [29, p. 39].

The rapid pace of technological innovation in computer and communications equipment that began in the 1990s changed the landscape. It changed what the military bought—increasingly consumer-grade equipment [106, pp.434-436]—and it changed what soldiers bought and used. And that all of a sudden made the order of battle *much* easier to determine. One need only look at the example of Strava's fitness app, which publicly shared exercise routes. Unwitting soldiers shared their running data with the company, which then posted the metadata of their routes—routes sometimes around *secret military bases* [82]. A century ago determining an enemy's order of battle took concerted effort. Now the ubiquity of connected

devices has, in many instances, simplified the process for a nation interested in knowing how its adversaries have laid out their forces.

Exhibiting Community of Interest

Those who read spy novels believe that mining communications metadata for small groups of people who communicate only amongst themselves is extremely useful for turning up terrorist nests. In fact, such a technique was used by Hezbollah [22]. But such a technique has a variety of applications, which I will examine here.

After the 9/11 attacks, President George Bush authorized the bulk collection of domestic CDRs from U.S. telephone providers. One of the motivations was the failure to discover that calls made between two hijackers and a Yemeni al-Qaeda safe house were actually between San Diego and the known number [76]. This program eventually came under Section 215 of the USA PATRIOT Act. Former NSA Director Michael Hayden later said that, "If we had the 215 program at the time, we would have thrown that selector [the Yemeni phone number] at that mass of American phone bills and phone connection and said, 'Did anybody here talk to this number in Yemen?,' and ka-jink! The San Diego number would have popped up." [76, pp. 25-26]

The bulk collection effort ended in 2015 and was replaced, under the USA FREEDOM Act, by U.S. government access to the CDR records stored at the telephone providers. Its original instantiation allowed up to three "hops" from a selector, although collecting three hops was unusual and up to two was the norm. The value of the bulk collection program includes being able to find "alternate identifiers" for targets of interest as well as finding "connectors" between groups. Such connectors are quite important in understanding conspiracies[6].

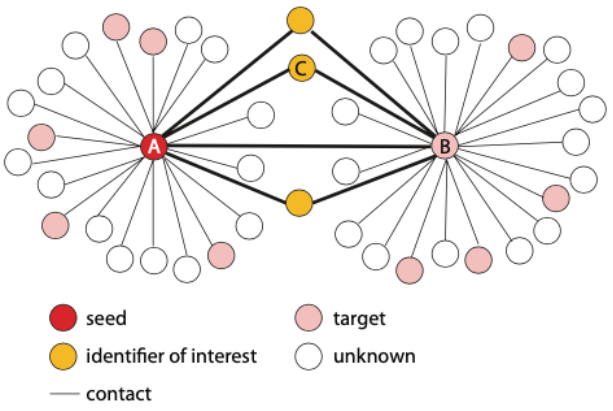

Figure 1: "Connector" in Contact Chaining (diagram from [131, p.43])

Figure 1 gives an illustration of C's role as a connector. A National Academies study on the issue[7] described how bulk collection would enable discovering C, "[I]f all calls between A or B and C occurred before either A or B was identified as a target, later collection targeted on A or B will not find C by way of A or B . . . Bulk collection provides useful 'history,' because it does not limit collection to only the targets known at the time of collection." [131, p. 43]

When Snowden disclosed the U.S. government's bulk collection of domestic CDRs, many believed that the set was thoroughly datamined in multiple different ways, including for "communities of interest": small groups of people who spoke only with each other [50]. In fact, although some governments do use such analysis to find terrorist organizations, the U.S. domestic CDR collection appears to have been used solely for searching for identifiers within two, and occasionally, three hops of an identifier satisfying "reasonable and articulable suspicion" of ties to specific terrorist groups. It is more than a little ironic, then, that Hezbollah used exactly such techniques to uncover CIA agents and their informants in Beirut in the mid 2000s, a discovery they publicized on television [46, 22:45-25:10].

Studying CDR-based "communities of interest" helps in telephone fraud detection. Small-time fraudsters tend to be well connected to other fraudsters; having found one, investigators use the affinity graph to search for others [50]. Fraudsters who commit identity theft to open new accounts have communication patterns that resemble their old one, and that "community of interest" helps in giving them away [50].

Just as with order of battle, determining a community of interest, whether for a criminal, terrorist, or some other nation's spies, has become significantly simpler due to modern communications technologies.

Revealing Organizational Structure

As anyone who has worked in a complex organization knows, there is the "org chart," and there is how the institution actually works. Communications patterns are readily revealed by communications metadata, which also can show informal networks as well as changes to a community's structure

There are many types of organizational structures that communications metadata reveals. A fundamental one is corporate structure, both formal and informal; tracking who speaks to whom—when and for how long—during times of corporate change shows both of these. Enron, the energy corporation that engaged in massive accounting fraud in which the senior company executives engaged in deceptive practices to hide major liabilities from the company's books, went bankrupt in 2002. The Federal Energy Regulatory Commission made the somewhat surprising decision to post 1.6 million emails sent and received by Enron executives from 2000-2002 (the commission later removed some of the more sensitive and personal of these). These became a treasure trove for researchers. Studying average response time and cliques, Rowe et al. were able to determine the "social hierarchy" of an Enron division (the same technique also worked in analyzing an academic setting) [142]. Email trails also showed a sharp change in October 2001, the time of the Enron crisis [55]. During that period,

> The Board . . . diminished communication within this rank. In contrast, the Executive Management as well as the Lawyers intensified their lateral contacts and

[6]Sociologist Shin-Kap Han observed that the American revolutionary, Paul Revere, fit exactly that important connector status; he was a member of several social organizations that served as organizational bases for the revolutionary movement [79, pp. 149-150].

[7]I participated in the study.

> decreased upward reporting through the crisis. [56, p. 221]

Corporate phone calls and emails are but one way of using communications metadata to determine organizational structure. Chen et al. showed that tracking accesses of electronic health records—the logs of such accesses are simply another form of communications metadata—demonstrates how decision making is done within a medical organization [41]. (Unlike most studies in this domain, the paper by Chen et al. is not a "big-data" work.) Further work of this sort is able to determine collaborative care teams [42]. A meta analysis of such studies showed that such forms of tracking were used to study multiple issues, including care coordination, organizational effectiveness, and interaction of care teams [33].

The organization of less formal networks of people are also revealed through their communication patterns. Studying the metadata of communications in a social graph shows who the key influencers are in social media [140]. Communications patterns can demonstrate the organizing behavior of a protest group [48]. Using nine months of a Twitter "garden hose" feed[8] from the Occupy Wall Street movement (July 3, 2011 to March 12, 2012), Conover et al. discovered a difference between the types of communications within local communications and those that stretched across state boundaries (which is a proxy for communications across geographic distance). The former communications were focused more on the logistics of local demonstrations while the latter were more about strategic issues including framing language and communications with the media [48, p.1].

Communications metadata can also demonstrate more subtle interactions between people. In modeling disease spread, epidemiologists assume a perfect mixing of groups of people, but Onnela et al. has shown that geography "compartmentalizes"; network ties erode with distance [134].

When the telephone companies were monopolies (and sometimes state monopolies), researcher access to CDRs was essentially nonexistent. That has changed over the last decade. The wealth of publicly useful information that could be derived from mobile phone CDRs helped cause a rethinking. The availability of that data brought in a slew of social scientists—along with the type of questions social scientists ask. There has been a blossoming of research in understanding social structures such as organizational structure or community characteristics well beyond how people use electronic communications; the work by Onnela et al. points to an example that is not about how people use mobile communications, but about the impact of geographic distance on social networks.

Demonstrating Community Characteristics

Phone call patterns—Internet usage during the World Cup Draw [58], at the end of Ramadan, on Christmas [57]—reveal societal characteristics including a society's religion. Communications metadata can also reveal more subtle features of a society, which I discuss here.

The metadata can also reveal less obvious characteristics, such as an area's relative wealth. Eagle et al. observed that, "[D]iverse communication patterns tend to rank higher [in economic development] than the regions with more insular communication [implying] communication diversity is a key indicator of an economically healthy community." [60, p. 1030]. This 2005 study was of the U.K.; in 2011 Soto et al. applied machine learning techniques to CDRs to successfully predict socio-economic level of a Latin American city [148]. Working with datasets from Rwanda, Blumenstock et al. showed that mobile phone metadata can be used to predict distribution of wealth within a nation—and even such information as whether an individual "owns a motorcycle or has electricity in the household" [26]. Related work by Mao et al. on a dataset from Cote d I'voire showed that "social centrality in mobile communication networks—PageRank—can identify economic centers at the national and city levels." Their work showed that rich areas are more likely to communicate with rich areas, and not with poorer regions of the country [118].

CDRs for mobile phones provide geographic information, and this can be used in ways that landline phones cannot [25], e.g., to produce a relatively precise census [53, 58]. Such a count will not include detailed information on domicile size, number of toilets, televisions, and phones that a U.S. census does. But using mobile phones as a basis for the count can provide a relatively accurate measure of population size at a fraction of the cost of the way a census is normally conducted. Such a methodology can be used in other ways as well, e.g., to probe the density of a city at midday—or at midday during a pandemic.

Using the Telecom Italia data, Bajardi et al. show how to recognize immigrant communities living within a city [13]. They introduce the concept of an *entropy* function that measures the number of distinct countries within the calling patterns of a neighborhood. This functions enables Bajardi et al. to distinguish between tourist and visitor attractions with a high transient population and residential neighborhoods with a high proportion of immigrants.

In short, as Blondel et al. noted, the fact that mobile CDRs provide geographic information vastly increases their value as a sociological research tool [25].

Monitoring Societal Movement

The most salient characteristic of cellphones is that they are mobile; that fact has not escaped the notice of sociologists, epidemiologists, statisticians, transportation planners—or any other scholar whose work intersects with people and movement. As Williams et al. noted in 2015, "This exciting new type of data holds immense promise for studying human behavior with precision and accuracy on a vast scale never before possible with surveys or other data collection techniques." [165] There has been a rather large amount of research based on what one can learn from the CDRs of mobile phones; I present here a sample from the most significant areas of research.

A killer app for smartphones is real-time route planning. Already as early as 1994, there was research into mapping travel using cellphone data [150]. Working in Bangkok in 2007, Pattara-Atikom and Peachavnish used "cell dwell time"—the time a cell phone spent at a particular base station—to measure traffic speed; this produced data accurate to within 73% to 85% [139]. Over the following several years, there were multiple such experiments in the Bay Area and elsewhere [150], precursors to such applications as Waze, Google Maps, etc. The 2016 survey by Gundlegard et al. found that cellular network data produced the obvious results of trip extraction, trip speed, but also detection of which subway segments the user traveled; the paper also listed work using differential privacy to protect

[8]This was approximately 10% of the Twitter stream that was in machine-readable format [48].

the data of individual users [77]. Where people go, how long they take to get there, what modes of travel they use will remain a hot area of research; except in a time of lockdown, such information affects nearly everyone.

CDRs from mobile devices provide quite rich information about peoples' activities. This ranges from the mundane—recognizing patterns of city life, e.g., commuting patterns [15, 102] what times of day and what days of the week big city squares are busy [35, 157]—to the disruptive—where publicly disruptive or violent actions are occurring [78]. There has also been work to use cellphone data to track where large public events are occurring such as the start of a historical road race, "Mille Miglia" [37] or a public protest [57]. These can be characterized by decreased phone movement and call volume [57].

CDRs from mobile devices can be used to inform and improve public health. This happens in a number of different ways depending on the type of health issue and the type of public response. During major disasters such as hurricanes and earthquakes, people leave their homes. Knowing how people respond—where they go during a crisis—is crucial to providing them with aid, whether food or improving facilities (sanitation, health care, etc.). By tracking the movement of SIM cards after the 2010 Haitian earthquake, researchers produced estimates that matched the Haitian National Civil Protection Agency's, which had tracked the refugees' movement based on counting ship and bus movements—and was slower and far more labor intensive [19]. Studying the mobile CDR records during Haiti's 2010 cholera outbreak, Bengtsson et al. showed that such phone records could be used in the future for improving response efforts [18]. Tracking movement can also be used to develop public health responses in non-emergency situations, e.g., learning people' movement in sub Saharan Africa shows where resources are best placed to slow the spread of malaria [154, 160]. Similar work has been on using CDRs to model, and thus slow, the spread of dengue fever [162]. In the wake of the COVID-19 pandemic, Google began offering aggregated travel information to help public health officials understand the response to social distancing guidance [71].

Tracking cell phone movement also provides real-time information on migration in response to a crisis, often providing more timely and accurate information than can be obtained by other means [87, 116].[9]

Thus what we see is that the transit information provided by mobile CDRs provides data about human movement on multiple different time scales: hourly (for commuting data), daily (about social activity such as civic events), multi-day to monthly (e.g., about migration). This information is extremely valuable to governments for both immediate and future planning for public health and safety.

5.4 Discerning Characteristics about Individuals

In considering what characteristics communications metadata might reveal about individuals, first one has to ask: are we looking at information revealed from a single device or aggregated from multiple devices through which a user interacts.

[9] "We saw a clear increase of users arriving in Chittagong beginning approximately two days after the cyclone and continuing throughout the remaining one and a half months, during which highly temporally resolved data were available." [116].

In order of simplicity—and also privacy invasiveness—questions proceed from what device is a person using, what are they using it to do, who is using it, and what does a person's use reveal about them? I will discuss each of these in turn. I will then briefly touch on what one can learn from data aggregation of communications metadata. In contrast to group characteristics, where mobile CDRs provide extremely useful information for learning about the actions of individuals, communications metadata arises from a wider variety of devices, including IoT.

Identifying Devices

The lesson from Peter Exckersley in 2010 was that all Firefox browsers may look alike to their users, but each one was different in its own way when viewed by the web server [61]. Nor was Firefox unique; the same was true of Opera, Chrome, Safari and others. The point of Eckersely's work was in that in sharing information about how to display, the browser had shared enough characteristics about its operating environment that it could be "fingerprinted." Others have looked at other ways of fingerprinting devices; we examine some of these here.

Devices on the network are, by definition, accessible; the information they release from their presence may allow device identification. In 2005 Kohno et al. showed that clock skews read from TCP timestamps could be used to identify devices [99]. This form of identification can be used to identify a device as it connects to the Internet, enabling forensic investigations.[10] Arackaparambil et al. later presented how to use clock skew as a way to authenticate a device [10]. It is not unusual to see research that shows how to use an attribute to identify a device to be later repurposed into using that attribute for authenticating the device [9].

Eckersley showed that web browsers can be "fingerprinted" through collection of normal information that the browser shares: the browser, version number, any extensions that are being used, system hardware, browsing history, display information such as font text, size, and background colors, [1, 61]. The latter is known as Canvas fingerprinting after part of HTML5's Canvas element that includes such information for enabling attractive displays. In 2014 Acer et al. observed that 5% of the top hundred thousand websites used Canvas fingerprinting to identify devices [1]. In 2016, Laperdrix et al. showed that browser fingerprinting worked on mobile devices despite that being a more constrained environment [103].

Others used different techniques to identify a device. Using a sample set of 80 standalone accelerometer chips, Dey et al. found it was possible to fingerprint individual chips [54]. That in itself was not sufficient to fingerprint a device, but Hupperich et al. showed how to combine such fingerprints with device information revealed by the browser and data about the device's location gleaned from the IP address and hostname of the WiFi router; that was sufficient to identify the individual devices of the 724 participants in the study [86].

Amerini et al. and others have studied the question of using smartphone fingerprints as a way to authenticate the devices, thus enabling the phone to be a trustworthy security element, concluding that this direction is plausible [3]. The paper observes that the ability to recognize individual phones in this way raises privacy issues. It

[10] It can also be used for other purposes, e.g., to count the number of hosts behind a NAT.

also raises security concerns. The U.S. Department of Homeland Security is investing in research to use "continuous authentication" from mobile device components such as gyroscopes, GPS, force sensors to authenticate the user [144, p. 12, p. 37, and p. 40]. If such work is successful, it is likely to be used in many environments, not just high-value security domains.

Now we turn from identifying the device to something perhaps even more intrusive, identifying what the device is doing.

Identifying Device Activity

Mobility made phone CDRs immensely more valuable, enabling many of the research developments described in the previous subsection. Here, in identifying device activity, we see the power created by the design choice of smart endpoints. The fact that Internet endpoints have the capability to support multiple types of services and that the Internet's communications endpoints have a wide variety of types of information to transmit means that there is a tremendous ability to learn what activity is occurring in the underlying endpoint device. There are multiple ways to learn device activity from transactional information: communications metadata in transit can be used by network operators to understand the underlying activity; actions on the device can generate responses that, when their transactional information is examined, can reveal what is occurring on the device even if the application itself has only operated locally, and is occurring, on the device; and from simply the existence of communications from IP-enabled devices. These are but a tip of the iceberg of how communications metadata can reveal device activity. I'll briefly examine each of these types of identification mechanisms.

Network researchers have an intense interest in learning about traffic flowing over the network. Such study serves various purposes, including enabling an understanding of how traffic is changing (which is necessary in order to improve quality of service), seeing how to recognize illegitimate uses of the network (in order to develop preventative measures), etc.

Peer-to-peer (P2P) file sharing of copyrighted material presented a challenge in the late 1990s and early 2000s. Early on, port numbers were a useful way to determine the type of traffic flowing across the Internet. But to evade copyright enforcers, P2P file sharers began using alternate ports so as to hide their activity. So network researchers began studying how to recognize this traffic without being too invasive. Karagiannis et al. showed an efficient way to do so by searching for certain bit strings used in TCP layer traffic of P2P communications [91]; Sen et al. took a related approach [145]. As the work took off, Kim et al. surveyed these approaches, noting there were ones that focused on "host-based behavior" (e.g., [92]), others that looked directly at traffic flow [97]. Kim et al. compared the different methodologies and observed that despite the P2P efforts to camoflauge activity by using unusual port numbers, port number remains a valuable way to classify traffic type, as does characterizing host-based behavior [97]. The latter is not good, however, for "elephant-flow" situations, that is, the situation of an extremely large, continuous flow over an IP network.

Another source of information of device activity is the source addresses of packets delivered to a smart device, e.g., IP source addresses of packets delivered to a smartphone may reveal what apps are on a phone (or what apps are being used at a particular time) [17].

The uses of email metadata are quite straightforward, but other aspects of communications can be surprisingly revealing of activity. One such is from the metadata of VoIP communications. Patterns of packet sizes may identify the language being spoken [168]. Backes et al. have shown it is possible to identify the speaker even if the speech is encrypted [11]. This metadata can also reveal when speakers change language, a signal of a change in mood or thought.

In 2012, Moore and Clayton showed how to use email metadata to recognize the dropboxes in which spammers harvest victims' information [128]. Their idea was to respond to spammers with a fake address, then rely on the metadata email providers use for spam detection. This shows to which Dropbox account the email is being delivered.

The metadata of IoT devices is likely to reveal usage, that is, a smart coffeepot is most likely to communicate when it is being used. In 2008, Pai et al. studied the revelatory nature of transactional information in sensor networks, observing that an adversary can determine what devices are in operation (this has particular significance for military equipment), where objects of interest are (this is through frequency of communications), etc. [135]. The smart home has made this real; in 2019 Apthorpe et al. ran tests showing how "a passive network observer could . . . infer consumer behavior from rates of IoT device traffic, even when the traffic is encrypted" [7]. As we increasingly move into the world of smart devices, the leakage of personal information from the communications patterns of smart coffeemakers, smart beds, smart toothbrushes, etc. could be enormous.

Finally, I come to a largely understudied area: telemetry, information collected about device and application usage. This form of data collection is new; such data collection about wireline phones is not particularly illuminating. Telemetry of smart phones, however, can reveal the system's health, e.g., how the battery is doing. It can also reveal what a user does, how a user employs the device, and potentially help advise a system manufacturer about future product designs.

The providers of devices and services are largely circumspect about what information is collected and what is done with it. For example, Google's statement about collecting Android telemetry simply says:

> Android 9 includes the statsd telemetry feature, which solves this deficiency by collecting better data faster. statsd collects app usage, battery and process statistics, and crashes. The data is analyzed and used to improve products, hardware, and services. [5]

Telemetry data is used in part to improve user experience, both in the moment and, by studying user responses to features, in future products. However, the telemetry information can also be used to track the user. It is out of scope to study this issue in detail here, but recall Leith's study of the telemetry communications of six major browsers—Chrome, Firefox, Safari, Brave, Edge and Yandex—by Leith [112]. Leith observed that the identifiers persist over four different timespans: (i) ephemeral identifiers; (ii) session identifiers, reset on browser restart; (iii) browser instance identifiers set on installation, and (iv) device identifiers [112]. Brave uses only ephemeral identifiers; Chrome, Firefox, and Safari use session and

browser instance; Yandex uses device identifiers [112]. Leith did not give Edge or Yandex a clean bill of health.

Leith's study, which is quite thorough, is for a single application—web browsers—albeit a very important one. Much more work needs to be done examining what telemetry data is being collected from various devices, how it is used, and what the privacy implications of this collection are.

Identifying the Device or Application User

When law enforcement does investigations of criminal activity conducted over the Internet, one limiting factor in prosecution is the need to actually identify the user at the other end of the device. While this is seriously problematic for the prosecution of cases, in practice, there are multiple different ways to determine who a user is by using communications metadata. Some solutions involve using communications metadata from the user's interaction with the device, while others involve the user's interaction with applications.

If communications metadata can be used for authenticating a device, it is also very likely that this transactional information can be used for identifying a user. Explorations of authentication include the previously mentioned work by Amerini et al.; they are examining the feasibility of smartphone fingerprints for authenticating devices [3]. There is also work to authenticate via typing patterns[11], but so far password typing patterns have not proved effective as a second-factor authentication mechanism for users (and thus for identifying users) [38]. Touch patterns [120] and mobility patterns [89] are also being explored for authentication; while these methods still have a ways to go in terms of reliability, mobility patterns at least are already at the point of being a realistic threat against users' privacy [51].

Where people go is often sufficient to identify them. In 2009 Golle and Partridge showed that roughly knowing the locations of a person's home and work was often sufficient to uniquely identify them [125]; such information is, of course, easily determined by cell-site towers and IP addresses. That work was in theory; studying 1.5 million individuals over fifteen months, de Montjoye et al. showed that four spatiotemporal points derived from cell-tower locations were sufficient to identify 95% of the individuals [125]. It's not just the metadata of what we usually think of as communications that will unmask individuals in this way; de Montjoye and a different set of collaborators showed that the spatiotemporal records of three months of credit-card records for 1.1 million people could be used to uniquely reidentify 90% of the users [127].

Returning to metadata of VoIP communications, because voice communications can be compressed; this saving of bandwidth matters in real-time communications. This compression can provide sufficient information to enable recovering individual phonemes (a unit of speech corresponding to a particular sound, such as the "n" sound in "not" and the "b" sound in "bot"). Backes et al. have shown it is possible to identify the speaker even if the speech is encrypted [11].

Of course, when we ask whether some piece of information is at risk of being discovered, the question is by whom and what additional information they have. Some adversaries—and in the case of consumer devices and applications, this might include the manufacturer or developer—might be receiving additional information that allows them to conclude that (i) the actions on a device or application are all by the same person, and (ii) that person is the same as someone they have previously identified. This could occur in many different circumstances; signed-in users on phones or browsers are two obvious examples. Thus the universe of devices and applications that might be able to identify a user is significantly larger than the set of examples I have just provided.

Profiling the User

It might seem that being able to identify an individual from their communications metadata records is the most invasive form of privacy intrusion, but that is not so. A person's identity—their name—is public. Who they are, how they spend days, their personality is private. Yet communication metadata can divulge all this and more. Furthermore, some profilings of users can be socially beneficial. I will start with two of those—both of them based on metadata from Twitter accounts—and then discuss the privacy-intrusive examples.

One of the more interesting examples of profiling a user through traffic metadata is that it is possible to determine, "That's not a person; it's a Twitter bot!" Using such information as the ratio of number of followers to number of followed, and timing of tweets, Chu et al. were able to distinguish between real followers and automated ones [43]. It is also possible to use Twitter metadata to determine whether a user is an ISIS follower [98]. While such a user does not behave like a bot, neither is his behavior like a normal Twitter user. Instead, as Klausen et al. observe, features that are useful for predicting whether an account is an ISIS follower include whether the account has been following ISIS seed accounts, whether the account has been previously suspended, number of "friends" and "followers," number of tweets from the account, whether geolocation is enabled, whether the account is protected, and whether the account is verified [98].

In 2008 I commented that "transactional information is remarkably revelatory." [72]. That turned out to be an understatement. In his famous dissent in the 1928 case *Olmstead v. United States,* Justice Louis Brandeis warned:

> The progress of science in furnishing the government with means of espionage is not likely to stop with wiretapping. Ways may some day be developed by which the Government, without removing papers from secret drawers, can reproduce them in court, and by which it will be enabled to expose to a jury the most intimate occurrences.[12]

Justice Brandeis was prescient. Earlier I discussed ads that could be sent to TV set-top boxes based on the ambient noise [156]; the source address of such ads—e.g., diapers versus beer—provides an indication of the type of activity occurring within the room [17]. It

[11] This issue has a long history; the "fist" of telegraph operators was used by British wireless operators during World War I in order to recognize individual telegraph operators on German naval ships. The Americans used the same technique, not always successfully, against the Japanese in World War II [12, p. 308].

[12] Why Privacy? 169 170 Privacy: Protections and Threats of the home. Advances in the psychic and related sciences may bring means of exploring unexpressed beliefs, thoughts and emotions … [31, pp. 474]

should be no surprise then that communications metadata associated with your phone would reveal your gender, age, marital status, education, and income [113, 119, 146].[13]

It is possible to determine multiple aspects of a person's daily activities [8], from sleeping patterns [16, 147] and exercise routines [82] to their use of the TV and Amazon Echo [2] simply from the timing and location of IoT traffic. It is similarly possible to learn much about people's activities from their CDRs. Examining a set of telephone CDRs submitted over a 32-week period from 823 volunteers, Mayer et al. showed it was possible to infer such private information as that someone was planning to start a marijuana growing business and that someone else had an abortion [121].[14]

Communications metadata is also revelatory of people's thoughts. Because webpages now load from multiple sources, packet lengths reveal what webpages are visited, while time/spatial proximity of users, which is easily determined from the communications metadata of mobile phones, is indicative of a personal relationship. Indeed, Facebook has a patent application for connection recommendations based on this idea [39].

Perhaps even more striking is still early research on the use of communications metadata to reveal deep personality traits. Using measures of basic phone use, behaviors of active users, mobility, regularity, and diversity—all available as part of phone metadata—de Montjoye et al. predicted user neuroticism, extraversion, agreeableness, conscientiousness, and openness [124]. The study took place over fifteen months and involved 69 participants. A longer-term study by Viana et al.—three years involving 55 participants—appeared to confirm these results, though the Viana study also used information regarding battery charging and Bluetooth proximity that would not necessarily be included in communications metadata [159]. Because these studies are high touch—they involve contact with subjects in order to determine actual personality traits—they are expensive. Because of the cost of labeling data, we do not expect Big Data versions of such work soon. But the initial research seems quite clear that the communications metadata can certainly provide at least a "first cut" on personality features.

Communications metadata can determine subtle aspects of the nature of people's relations with close friends or family. One would expect that the frequency and time of communications shows the nature of relationships, and indeed, that is so [129]. But even more subtle aspects of human relationships can be determined. One of the most interesting pieces of information to come out of studying mobile CDRs is the emotional distancing that resulted from the 2016 U.S. presidential election. One study found that guests from "opposite" voting precincts spent up to fifty minutes less time with each other at Thanksgiving than they had in previous years [40]. The data was carefully collected to look at those people who could easily control their time at the hosts. These were visitors for the meal, not the weekend. The visitors' phones were at home Thanksgiving morning and evening, but at the hosts during part of the day [40].

The takeaway of this discussion is that communications metadata is frighteningly capable of discerning personality traits of individuals, including the nature of their relationships with others.

Location Data: Recognizing the User

Location data is the most valuable communications metadata that mobile phones provide.[15] *It can be used to identify the user, where the user is, and what the user is doing. Governments use this information—and so can other organizations.*

Location data has proved extremely useful in criminal investigations. Cell site location records have proved invaluable in any number of criminal investigations, and such records are often the go-to evidence for many types of investigations. "We find people," said one criminal analyst, "and it saves lives." [114]. Location data has proved similarly useful in intelligence investigations. An example is the 2005 assassination of former Lebanese Prime Minister Rafik Hariri, which was unraveled through evidence from cell site location records CDRs [22].

Location information is also useful in uncovering the identity of "ordinary" individuals. We have already seen that as few as four location identifiers at the granularity of cell-tower locations are enough to identify 95% of the population. Gambs et al. showed how to reidentify "anonymized" users in a geolocated dataset [67]. In 2018, Manousakas et al. showed that an anonymized mobility location graph—a set of nodes representing anonymized sites a person reguarly visited and a set of edges with weights corresponding to the probability of a transition from one site to another—can be used to uniquely identify an individual [117] (This result is reminiscent of earlier work by Cortes et al. that an individual's communication networks can be recognized by its graph structure even when anonymized; this was used in cellphone fraud detection [50].)

You "are" your set of locations. And that means that the information from communications metadata—from mobile CDRs, from email IP addresses, from ads served to your phone, from GPS—is highly personal—even while it is now widely accessible.

5.5 What does aggregating communications metadata provide?

In 2010, Judge Douglas Ginsburg, in writing the opinion for *U.S. v. Maynard*:[16], observed,

> Prolonged surveillance reveals types of information not revealed by short-term surveillance, such as what a person does repeatedly, what he does not do, and what he does ensemble. . . . [A] single trip to a gynecologist's office tells little about a woman, but that trip followed a few weeks later by a visit to a baby supply store tells a different story. A person who knows all of another's travels can deduce whether he is a weekly church goer, a heavy drinker, a regular at the gym, an unfaithful husband, an outpatient receiving medical treatment, an associate of particular individuals or

[13] Seneveratne et al. show that it is possible to determine gender from the smartphone apps [146]. Malmi and Weber show that it is possible to determine gender, age, marital status, and income [119], while Li et al. show it is possible to determine gender and education level [113]; [88]. show it is possible to predict gender from CDRs [88].

[14] As this paper focuses on issues of personal privacy, I am only addressing the IoT devices for personal use and not those in factories and business environments.

[15] Location data is available from multiple different sources: CDRs, data from a mapping application, a sequence of WiFi access points [143]. Because the focus of this paper is on privacy issues arising from communications metadata, in this discussion I focus on CDRs rather than on data from mapping applications and WiFi locations, since those are the data being transmitted—and not the metadata.

[16] *U.S. v. Maynard* was consolidated into U.S. v. Jones 615 F. 3d 544.

> political groups—and not just one such fact about a person, but all such facts. [68, pp. 29-30]

Location information is striking in the type of inferences that can be drawn about the user, but location information is far from the only change in the metadata collected by communication carriers and Internet companies. Changes in protocols, storage, and speed of communication networks have had a dramatic impact on the ability to harvest metadata and analyze the information hidden within it [117, 136]. The DHS effort on continuous authentication provides a different model of using multiple forms of telemetry for authenticating a mobile device user [144, 12, p. 37, and p. 40].

Judge Ginsburg's 2010 opinion was from a time when smartphone penetration was 20% of the U.S. population and smart home devices were a figment of engineers' imagination [149]. Since then, mobile phones have become the device of choice for web browsing, the mobile ad network, with its ability to track user location through GPS, WiFi, Bluetooth, and DNS, has emerged, and IoT devices have become common. Current aggregation of communications metadata is well described by the Rodgers and Hammerstein song, "Getting to know you, getting to know all about you." [17]

6 PROTECTING PRIVACY

The violation of user privacy through the use of communications metadata, including telemetry, has slipped in on little cat feet. At some level, users know that some types of communications information—their phone number, their IP address, perhaps the color they prefer for not-yet-clicked weblinks—is shared (indeed, the fact that users might know this is the logic behind the lower legal protections afforded to communications metadata by the *Smith v. Maryland* decision). From various well-known cases (see, e.g., [4]), users may be aware that communications metadata can lead to their identification. But few people are aware that communications metadata can also reveal their preferences and create a rather complete picture of who they are, what they buy, what they read, listen to and watch, where they go, who they meet there and what they do. This makes profiling of a user through communications metadata perhaps the most serious of the privacy invasion of this type of transactional information.

6.1 Technological Solutions

Now there are various ways to protect against the collection of transactional information. Tor is, of course, designed for Internet connections, solving the metadata privacy problem for a very large use class. Other solutions have been proposed for various communications situations. In 2003 Beresford and Stajano proposed temporary identifiers for hiding users' location information [21], a solution that was already used by mobile phones (TMSI) and is now part of the Apple-Google infrastructure for contact tracing, but is not standard practice. There's a serious tension here; as Blondel et al. observed, "A compromise between preserving the anonymity and keeping enough information in the dataset is difficult to achieve." [25]. Location information is simply too valuable for the advertising ecosystem to be anonymized except in some custom cases.

[17] This is from the musical "The King and I."

There are some other approaches. de Montjoye et al. proposed the idea of a "Personal Data Store" with computations done within the user's space returning responses to queries but not raw user data [126]. It was used in two field studies, but has not really seen uptake; to be fair, neither have other systems that propose local computation. Li et al. propose a somewhat ad hoc solution of a mix of use of VPNs and Tor, randomized MAC addresses, and dummy traffic to protect users from threats related to access of communications metadata while on WiFi networks [113]. Apthorpe et al. propose "stochastic traffic shaping" for home IoT devices; this is a method of making upload and download information for different devices have similar shapes and injecting random traffic into the Internet stream along with using a VPN to hide the protocol information (they do this via a VPN hosted on an instance of EC2) [8]. Jourdan et al. provided a solution against an adversary using communications metadata to identify a user; their solution involved using preprocessing information activity data on the phone, and then sending it off with a random pseudonym to the central database/cloud; if the recipient wants or needs to compute the data related to a particular individual, then it will be sent the appropriate set of pseudonyms [89].

But there are limitations to all of these approaches. The most effective tool, Tor, works and is widely used; indeed, it even thwarts the NSA [132]. But it slows connections, and it is often the case that Tor exit nodes are rejected at certain websites. Furthermore, Tor does not work with various plugins. Other proposals are either not actually implemented, not implemented at scale, or are too ad hoc to be actually useful for the public.

The underlying cause of the failure of technologists to provide general solutions that protect communications metadata in the way Tor does is that there are too many different solutions needed. The Internet and its smart endpoints have enabled an increasingly large number of types of communications, each with its form of metadata; more such are being advanced all the time. This conspires against the development of general tools for protecting all types of communications metadata.

That does not mean we should not pursue technological solutions for protecting the privacy of communications metadata. It is possible that a solution that affects as wide a swath of communications as Tor does can be designed, developed, and become somewhat widely used. Indeed it is possible, even likely, that the legal framework surrounding the collection of metadata will change, and such a change would make it easier to implement some of the technical solutions being proposed whose current adoption seems unlikely.

If such changes in law and policy regarding the legal protections afforded communications metadata are to occur, there are four questions that must be answered:

(1) What type of information does the use of metadata uncover?
(2) What types of personal information should be protected by legal means?
(3) From whom should this information be protected?
(4) What are the right laws and policies to do so?

The first has been the main focus of this paper. Let me now briefly discuss the next three, which will ultimately require a thorough legal and policy analysis, that is well beyond the scope of this paper.

6.2 What types of personal information should be protected by legal means?

My initial categorization of the uses of communications metadata into information about groups and information about individual devices and people was no accident. This is a natural split. First, there is both a real—and perceived—sense that the privacy invasiveness of the use of aggregate data is much less compared with data about individuals. Second there is the tremendous value of such aggregate information to governments; as Deville et al. observed, "In operational and governmental decisions, these data also may be valuable for supporting rapid responses to disruptive events or longer-term planning purposes." [53]. Finally it is not new that institutions use communications metadata—telephone companies have for decades—so while there is a radical transformation in the availability of data and the scale of its use, the idea that communications metadata might be used for future planning is not in itself new.

The Snowden disclosures led to public objections to the government's bulk collection of communications metadata, but this was bulk collection that could then be used to track individuals. Here the type of bulk usage is about tracking group behavior, not the actions of a single person or small group of people. Thus where the government actions are seen as beneficial, e.g., transportation planning, it is unlikely that there will be serious objections to the use of aggregated communications metadata. But tracking the movement of mobile phones is a way of determining anomalies in a population's activity. In situations of conflict, e.g., a government protest [57], while this is information that the government would really like to have, many in the public would object.

The penetration of mobile phones across the world, a phenomenon that occurred in the 2000s [167], has had another impact as well. This penetration provides governments unprecedented ability to gain insight into parts of the globe that were previously difficult to penetrate. Such spying is particularly important due to increased use of encryption. Knowing what is happening within your adversaries' borders through collecting their communications is now a particularly effective method to do so.

In some instances, there will be protections against the use of aggregated communications metadata. Some will be technical, e.g., the use of differential privacy to provide some anonymity for those within the dataset (see, e.g., [71]), while others will be policy and/or legislative, as in protections against state use of aggregated communications metadata to learn about the actions of a political opponent (this protection would not work under an authoritarian regime). But since the use of aggregated communications metadata to infer information about groups is simply too valuable, has been accepted publicly for sufficently long, and does not substantively impede the privacy of individuals to be stopped, I do not expect society to move against this use.

There may be some minor exceptions, e.g., determining how big a group must be before studying its characteristics is not seen as infringing on individuals' privacy is one issue. Another is whether there are protected traits that should not be explored through the use of metadata. The data is easy to collect, and abuses will be difficult to prevent.

The issue of inferring information about individuals—their devices, who they are, their use of the devices, and their personalities—leads to a different conclusion. The activities described under §5.4 Discerning Characteristics about Individuals are highly privacy invasive and therefore must be protected against. However, some of the information collected, such as is what is provided by Canvas, is important in order that the feature functions. So it is fair to say that the provider of the service has reason to collect the data. The question is how the provider uses the data.

Collection should not empower the provider to use the information in ways other than to service the feature. That may sound draconian, especially given the ad ecosystem that has developed over the last two decades. But this is simply applying the purpose limitation principle of the Fair Information Practices to the use of metadata. There is precedent in this type of restriction; the Federal Identity, Credential, and Access Management (FICAM) implementation guidance required that providers of authentication services used user data for the purpose for which it was intended *and not otherwise* [64].

Private-sector providers developed identity and authentication systems for citizens to log onto federal sites and services. Here's the privacy aspect: the FICAM implementation guidance prohibits identity providers from using authentication confirmations of a user's credentials for any purpose other than to manage authentication [64, p. 375]. In other words, there is no activity tracking of a user across federal sites, no sharing of the user's activity within the identity provider's company—*this is not allowed even for improving user experience*—and no sharing of the user's activities with third parties.

6.3 From whom should this information be protected?

This answer is both short—and long. The short answer is that users must be protected from government and the private sector's ability to infer private information. In a situation in which metadata can increasingly provide private information about an individual, legal and policy protections are needed. The long answer requires a deep legal analysis, and will, like the next question, wait for another day and another paper.

6.4 What are the right laws and policies to do so?

In [17], my coauthors and I examined the issue of how the differences between PSTN communications and networks and IP communications and networks led to a mismatch of U.S. wiretap laws. By surveying the privacy leakages in different types of uses of communications metadata, the current paper explores that mismatch from a somewhat different perspective. Presenting potential solutions—possible laws and policies to protect users' privacy against the uses of communications metadata—will be the subject of future work.

7 SUMMING UP

In this paper, I have made a number of novel observations about communications metadata. First, I have provided an in-depth, albeit brief, explanation for the change in the value of communications metadata, including the value arising from the "smart endpoints"

paradigm of the Internet. Second, I have developed a categorization of the uses of communications metadata based on their privacy impact. This provides a critical first step in developing legal and policy privacy protections for communications metadata. I have provided an outline of next steps in such privacy protections. This work provides both an intellectual framework for thinking about the privacy implications of the use of communications metadata and the beginnings of a roadmap for providing privacy protections for users of electronic communications.

I have focused here on privacy of communications metadata, but privacy and security are closely intertwined. In developing privacy protections for communications metadata, we will also be securing them in various ways. And that is a very good thing for all sorts of security, including public safety and national security.

ACKNOWLEDGMENTS

This work was partially supported by NSF grant: CNS 1923528. This work has greatly benefited from comments provided by David Balenson, Jono Spring, Tom Walcott, and other participants in the New Security Paradigms Workshop. I am very grateful for their help.

REFERENCES

[1] Acar, Gunes, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, Claudia Diaz, "The Web Never Forgets: Persistent Tracking in the Wild," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security,* November 2014, pp. 674–689.

[2] Amar, Yousef, Hamed Haddadi, Richard Mortier, Anthony Brown, James Colley, and Andy Crabtree, "'An analysis of home iot network traffic and behaviour," https://arxiv.org/abs/1803.05368.

[3] Amerini, Irene, Rudy Becarelli, Roberto Caldelli, Alessio Melani, and Moreno Niccolai, "Combining Features of On-Board Sensors," *IEEE Transactions of Information Forensics and Security,* Vol. 12, No. 10 (October 20, 2017).

[4] Anderson, Nate, "How 'cell tower dumps' caught the High Country Bandits—and why it matters," *ArsTechnica,* August 29, 2013 [last viewed May 21, 2020].

[5] Android, Android 9 Release Notes, https://source.android.com/setup/start/p-release-notes [last viewed May 17, 2020].

[6] Apple and Google, "Exposure Notification FAQ 1.1," https://blog.google/documents/73/Exposure_Notification_-_FAQ_v1.1.pdf [last viewed May 21 2020].

[7] Apthorpe, Noah, Dillon Reisman, Nick Feamster, "A Smart Home is No Castle: Privacy Vulnerabilities of Encrypted IoT Traffic," May 18, 2017.

[8] Apthorpe, Noah, Danny Yuxing Huang, Dillon Reisman, Arvind Narayanan, and Nick Feamster, "Keeping the Smart Home Private with Smart(er) IoT Traffic Shaping, *Proceedings on Privacy Enhancing Technologies,* 2019, pp. 128–148.

[9] Alaca, Furkan and Paul van Oorschot, "Device Fingerprinting for Augmenting Web Authentication:Classification and Analysis of Methods," *Annual Computer Security Applications Conference (ASAC'32),* 2016.

[10] Arackaparambil, Chrisil, Sergey Bratus, Anna Shubina, and David Kotz, "On the Reliability of Wireless Fingerprinting using Clock Skews," *Proceedings ACM WiSec,* 2010.

[11] Backes, Michael, Goran Doychev, Markus Durmuth, and Boris Kopf, "Speaker Recognition in Encrypted Voice Streams," *European Symposium on Research in Computer Security,* 2010, pp. 508-523.

[12] Bakeless, John, *Spies of the Confederacy,* J.P. Lippincott, 1970.

[13] Bajardi, Paolo, Matteo Delfino, Andre Panisson, Giovanni Petri, and Michele Tizzoni, "Unveiling patterns of international communities in a global city using mobile phone data,"*EPJ Data Science,* Vol. 4, Article 3 (2015).

[14] BBC News, "Yo app warns Israeli citizens of missile strikes," July 14, 2014.

[15] Becker. R., R. Cáceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. "A Tale of One City: Using Cellular Network Data for Urban Planning," *IEEE Pervasive Computing,* Vol. 10, No. 4 (October-December 2011).

[16] Beddit Sleep Monitor, http://www.beddit.com/

[17] Bellovin, Steven, Matt Blaze, Susan Landau, and Stephanie Pell, "It's Too Complicated: How the Internet Upends *Katz, Smith,* and Electronic Surveillance Law," *Harvard Journal of Law and Technology,* Vol. 30, No. 1 (2017).

[18] Bengtsson Linus, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux, "Using Mobile Phone Data to Predict the Spatial Spread of Cholera," *Scientific Reports,* Vol. 5, Article 8923 (2015).

[19] Bengtsson, Linus, Xin Lu, Anna Thorson, Richard Garfield, and Johan von Schreeb, "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti," PLOS Medicine, August 30, 2011.

[20] Bensinger, Greg, "Talking Less, Paying More for Voice," *Wall Street Journal,* June 6, 2012.

[21] Beresford Alaister, Stajano Frank, "Location privacy in pervasive computing," *Pervasive Computing,* January-March 2003.

[22] Bergman, Ronen, "The Hezbollah Connection," *New York Times Magazine,* February 10, 2015.

[23] Blondel, Vincent and Gautier Krings, *NetMob 2011: Book of Abstracts,*, 2011.

[24] Blondel, Vincent, Adeline Decuyper, Pierre Deville, Yves-Alexandre De Montjoye, Jameson Toole, Vincent Traag, and Dashun Wang, eds., *Mobile Phone Data for Development: Analysis of Mobile Phone Datasets for the Development of the Ivory Coast; Selected Contributions to the D4D Challenge Sponsored by Orange,* May 1-3, 2013, https://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf.

[25] Blondel, Vincent, Adeline Decuyper, and Guatier Krings, "A survey on results of mobile phone analysis," *EPJ Data Science,* Vol. 4, No. 10 (2015).

[26] Blumenstock, Joshua, Gabriel Cadamuro, and Robert On, "Predicting poverty and wealth from mobile phone metadata," *Science,* Vol. 350, Issue 6264 (November 27, 2015), pp. 1073-1076.

[27] Bojinov, Hristo, Yan Michalevsky, Gabi Nakibly, and Dan Boneh, "Mobile Device Identification via Sensor Fingerprinting," 2014.

[28] Borgman, Christine, Jillian Wallis, and Matthew Mayernick, "Who's Got the Data? Interdependencies in Science and Technology Collaborations," Computer Supported Cooperative Work, Vol. 21, Issue 6 (2012), pp. 485-523.

[29] Borrmann, Donald A., William T. Kvetkas, Charles V. Brown, Michael J. Flatley, and Robert Hunt, *The History of Traffic Analysis: World War I—Vietnam,* Center for Cryptologic History, National Security Agency, 2013.

[30] Borgman, Christine, *Big data, little data, no data: Scholarship in the networked world,* MIT Press, 2015.

[31] Brandeis, Louis. Dissenting opinion in *Olmstead v. United States,* 277 U.S. 438, 1928.

[32] "The Essence of the Fundamental Rights to Privacy and Data Protection: Finding the Way Through the Maze ofthe CJEU's Constitutional Reasoning," *German Law Journal,* Vol. 20, pp. 864-883, 2019.

[33] Brunson, Jason Cory and Richard C. Laubenbacher, "Applications of Network Analysis to Routinely Collected Health Care Data: A Systematic Review," JAMIA, Vol. 25, Issue 2 (2018), pp. 210-221.

[34] Calabrese, Francesco, Laura Ferrari, and Vincent D. Blondel, "Urban Sensing Using Mobile Phone Network Data: A Survey of Research," *ACM Computing Surveys,* Vol. 47, No. 2, Article 25 (November 2014).

[35] Calabrese, Francesco, Piero Lovisolo, Colonna Massimo, Dario Parata, and Carlo Ratti, "Real-Time Urban Monitoring Using CellPhones: A Case Study in Rome," *Transactions on IntelligentTransportation Systems,* Vol. 12, No. 1 (March 2011), pp. 141-151.

[36] Calabrese, Francesco, Estaban Moro, Vincent Blondel, and Alex 'Sandy' Pentland, eds., *NetMob: Book of Abstracts,* 2017, https://netmob.org/www17/assets/img/bookofabstract_oralt_2017.pdf.

[37] Carpita, Maurizio and Anna Simonetto, "Big Data to Monitor Big Social Events: Analysing the mobile phone signals inthe Brescia Smart City," *Electronic Journal of Applied Statistical Analysis,* Vol. 05, Issue 01, December 2014, pp. 31-41.

[38] Chairunnunda, Prima, Nam Pham, and Urs Hengartner, "Privacy: Gone with the Typing! Identifying Web Users by their Typing Patterns," *PETS* 2011.

[39] Chen, Ben, "Systems and Methods for Utilizing Wireless Communications to Suggest Connections for a User," United States Patent Application 20160014677, July 10, 2014, http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.html&r=1&p=1&f=G&l=50&d=PG01&S1=%2820160114.PD.+AND+%28Facebook.AS.+OR+Facebook.AANM.%29%29&OS=PD/1/14/2016+and+%28AN/Facebook+or+AANM/Facebook%29&RS=%28PD/

[40] Chen, M. Keith and Ryne Rohla, "The effect of partisanship and political advertising on close family ties," *Science,* Vol. 360, Issue 6392, pp. 1020-1024.

[41] Chen, You, Nancy Lorenzi, Steve Nyemba, Jonathan Schildcrout, and Bradley Malin, "We Work with Them? Healthcare Workers Interpretation of Organizational Relations Mined from Electronic Health Records," International Journal of Medical Information, Vol. 83, No. 7 (2014).

[42] Chen, You, Nancy Lorenzi, Warren Sandberg, Kelly Walgast, and Bradley Malin, "Identifying Collaborative Care Teams through Electronic Medical Utilization Records," *Journal of the American Medical Informatics Association,* Volume 24, Issue e1, April 2017, Pages e111–e120,

[43] Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia, "Who is Tweeting on Twitter: Human, Bot, or Cyborg?," ACSAC, 2010, https://www.eecis.udel.edu/ hnw/paper/acsac10.pdf.

[44] Cisco, *NetFlow Configuration Guide, Cisco IOS Release 15M& T,* https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/netflow/configuration/15-mt/nf-15-mt-book/ios-netflow-ov.html#GUID-0C91B715-F791-4F90-BF13-4654A1D7AFBB [last viewed November 10, 2020].

[45] Clark, David, "The Design Philosophy of the DARPA Internet Protocols," *Proceedings SIGCOMM 88, Computer Communication Review,* Vol. 18, No. 4, August 1988, pp. 106-114.

[46] Cole, Matthew, "OPSEC Failure of Spies," Black Hat USA 2013, December 3, 2013, https://www.youtube.com/watch?v=BwGsr3SzCZc.

[47] Cole, David, "We Kill People Based on Metadata," *New York Review of Books,* May 10, 2014.

[48] Conover, Michael, Clayton Davis, Emilio Ferrara, Karissa McKelvery, Filippo Menczer, and Alessandro Flammini, "The Geospatial Characteristics of a Social Movement Communication Network," PLOS ONE, March 6, 2013.

[49] Committee on Responding to Section 5(d) of Presidential Policy Directive 28: The Feasibility of Software to Provide Alternatives to Bulk Signals Intelligence Collections, National Research Council, *Bulk Collection of Signals Intelligence: Technical Options,* National Academies Press, 2015.

[50] Cortes, Corrina, Daryl Pregibon, and Chris Volinsky, "Communities of Interest," IDA '01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis, September 2001, pp. 105-114.

[51] Das, Anupam, Nikita Borisov, and Edward Chou, "Every Move You Make: Exploring Practical Issues in Smartphone Motion Sensor Fingerprinting and Countermeasures," Proceedings in Privacy Enhancing Technologies, 2018 (1), pp. 88-108.

[52] Daubert, Jorg, Alexander Wiesmaier, and Panayotis Kikiras, "A View on Privacy & Trust in IoT," *IEEE ICC—Workshop on Security and Privacy for Internet of Things and Cyber-Physical Systems,* 2015.

[53] Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Science,* 111(45) (2014), pp. 15888-15893.

[54] Dey, Sanorita, Nirupam Roy, Wenyuan Xu, Romit Roy Choudhury, and Srihari Nelakuditi, "AccelPrint: Imperfections of Accelerometers MakeSmartphones Trackable," NDSS 2014.

[55] Diesner, J. and K.M. Carley, "Exploration of Communication Networks from the Enron Email Corpus," Workshop on Link Analysis, Counterterrorism, and Security, April 23, 2005.

[56] Diesner, Jana, Terrill L. Frantz, and Kathleen M. Carley, "Communication Networks from the Enron Email Corpus: 'It's Always About the People. Enron is no Different'," *Computational and Mathematical Organization Theory,* Vol. 11 (2005), pp. 201-228.

[57] Dobra, Adrian, Nathelie E. Williams, and Nathan Eagle, "Spatiotemporal Detection of Unusual Human Population Behavior Using Mobile Phone Data," *PLOS One,* Vol. 10, NO.3 (2015).

[58] Douglass, Rex W., David A. Meyer, Megha Ram, David Rideout, and Dongjin Song, "High resolution population estimates from telecommunications data," *EPJ Data Science,* Vol. 4, Article 4 (2015).

[59] Dublin Core Metadata Initiative, "DCMI Type Vocabulary," http://dublincore.org/documents/dcmi-terms/#section-7 [last viewed December 22, 2017].

[60] Eagle, Nathan, Michael Macy, and Rob Claxton, "Network Diversity and Economic Development," *Science,* Vol. 328 (May 21, 2010), pp. 1029-1031.

[61] Eckersley, Peter, "How Unique is Your Web Browser," *Privacy Enhancing Technologies Symposium,* 2010.

[62] European Commission, "Proposal for a Regulation of the European Parliament and of the Council concerning the respect for privacy life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), 2017/0003 (COD), Brussels, October 1, 2017.

[63] European Commission, *Directive 2009/136/EC of the European Parliament and the Council,* November 25, 2009.

[64] Federal Chief Information Officers Council and Federal Enterprise Architecture, *Federal Identity, Credential, and Access Management (FICAM) Roadmap and Implementation Guidance,* Version 2.0, December 2, 2011.

[65] Filippova, Katja and Keith Hall, "Improved video categorization from text metadata and user comments," SIGIR 2011.

[66] Fitzgerald, Patrick, "The Evolving Role of Technology in the Work of Leading Investigators and Prosecutors," Palantir, June 12, 2013, accessed December 12, 2016, 11:39-12:44.

[67] Gambs, Sebastien, Marc-Olivier Killijian and Miguel Nunez del Prado Cortez, "De-anonymization attack on geolocated data," *Journal of Computer and System Sciences,* Vol. 80, Issue 8 (December 2014), pp. 1597-1614.

[68] Ginsburg, Douglas *U.S. v. Maynard,* 615 F. 3d 544 (D.C. Cir 2010).

[69] Gilliland, Anne, "Setting the Stage," in *Introduction to Metadata,* second edition, Getty Research Institute, 2008.

[70] Golle, Philippe and Kurt Partridge, "On the Anonymity of Home/Work Location Pairs," *International Conference on Pervasive Computing,* 2009, pp. 390-397.

[71] Google, "COVID-19 Community Mobility Report," https://www.gstatic.com/covid19/mobility/2020-04-05_US_Mobility_Report_en.pdf

[72] Gorman, Siobhan, "NSA's Domestic Spying Grows as Agency Sweeps Up Data," *Wall Street Journal,* March 10, 2008.

[73] Gratzer, Vanessa and David Naccache, "Cryptog- raphy, Law Enforcement, and Mobile Communications," *IEEE Security and Privacy,* Vol. 4, No. 6 (November/December 2006).

[74] Greenwald, Glenn, "NSA collecting phone records of millions of Americans daily," *Guardian,* Jun 6, 2013.

[75] Greschbach, Benjamin, Gunnar Kreitz and Sonja Buchegger, "The Devil is in the Metadata—New Privacy Challenges in Decentralised Online Social Networks," IEEE International Conference on Pervasive Computing and Communications Workshops, 2012.

[76] Griffiths, Rudyard, ed., *Does State Spying Make Us Safer: the Munk Debate on Mass Surveillance* 2014.

[77] Gundlegard, David, Clas Rydergren, Nils Bryeer, "Travel demand estimation and network assignment based on cellular network data," *Computer Communications* (2016), pp. 29-42.

[78] Gungdogdu, Didem, Ozlem D. Incel, Albert A. Saleh, and Bruno Lepri, "Countrywide arrhythmia: emergency event detection using mobile phone data," *EPJ Data Science,* Vol. 5, Article number 25 (2016).

[79] Han, Shin-Kap, "The Other Ride of Paul Revere: The Brokerage Role in the Making of the American Revolution," *Mobilization: An International Quarterly,* Vol. 14, Issue 2 (2009), pp. 143-162.

[80] Harris, Shane, "How the NSA Became a Killing Machine," The Daily Beast, November 9, 2014 (update April 14, 2017), https://www.thedailybeast.com/how-the-nsa-became-a-killing-machine [last viewed March 25, 2020].

[81] Hayden, Michael in "Johns Hopkins Foreign Affairs Symposium Presents: The Price of Privacy: Re-Evaluating the NSA," April 7, 2014, at 17.59.

[82] Hern, Alex, "Fitness tracking app Strava gives away location of secret US army bases," *Guardian,* January 28, 2018.

[83] Hersh, Seymour, "The Intelligence Gap," *New Yorker,* December 6, 1999.

[84] Hu, Yunhua, Hang Li, Yunbo Cao, Li Teng, Dmitriy Meyerzon, and Qinghua Zheng, "Automatic extraction of titles from general documents using machine learning,"Information Processing and Management, 42 (2006) 1276–1293.

[85] Hupperich, Thomas, Henry Hosseini, and Thorsten Holz, "Leveraging Sensor Fingerprinting for Mobile Device Identification," *Detection of Intrusions and Malware & Vulnerability Assessment—DIMVA 2016,* pp. 377-396.

[86] Hupperich, Thomas, Davide Maiorca, Marc Kuhrer, Thorsten Holz, and Giorgio Giacinto, "On the Robustness of Mobile Device Fingerprinting: Can Mobile Users Escape Modern Web-Tracking Mechanisms?," *Annual Computer Security Applications Conference,* December 2015.

[87] Isaacman, Sibren, Vanessa Frias-Martinez, Lingzi Hong, Enrique Frias-Martinez, "Climate Change Induced Migrations from a Cell Phone Perspective," *NetMob,* p. 46, 2017.

[88] Jahani, Eaman, Pål Roe Sundsø, Johannes Bjelland, Asif Iqbal, Alex Pentland, and Yves-Alexandre, de Montjoye, "Predicting Gender from Mobile Phone Metadata in *NetMob 2015.*

[89] Jourdan, Theo, Antoine Boutet, and Carole Frindel, "Toward Privacy in IoT Devices for Activity Recognition," *15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (Nov 2018).

[90] Kahn, David, *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet,* Scribner, 1996.

[91] Karagiannis, Thomas, Andre Broido, Michalis Faloutsos, and kc claffy, "Transport Layer Identification of P2P Traffic," IMC 2004.

[92] Karagiannis, Thomas, Konstantina Papagiannaka, and Michalis Faloutsos, *BLINC: Multilevel Traffic Classification in the Dark,* SIGCOMM'05, August 22-26, 2005.

[93] M. Karnan, M. Akila, and N. Krishnaraj, "Biometric personal authentication using keystroke dynamics: A review," Applied Soft Computing, Vol. 11, Issue 2, pp: 1565-1573.

[94] Kastrenakes, Jacob, "FCC fines Verizon $ 1.35 miillion over 'supercookie' tracking," TheVerge, March 7, 2016, https://www.theverge.com/2016/3/7/11173010/verizon-supercookie-fine-1-3-million-fcc.

[95] Kerr, Orin, "Websurfing and the Wiretap Act," *Washington Post,* June 4, 2015.

[96] Kerr, Orin, "Websurfing and the Wiretap Act, part 2: the Third Circuit's ruling," *Washington Post,* November 19, 2015.

[97] Kim, Hyunchul, kc claffy, Maria Fomenkov, Dhiman Barman, Michalis Faloutsos, KiYoung Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices," *ACM CoNEXT 2008,* December 10-12, 2008.

[98] Klausen, Jyette, Christopher Marks, and Tauhid Zamen, "Finding Online Extremists in Social Networks," *Operations Research,* Vol. 66, Issue 4 (August 2018), pp. 957-976.

[99] Kohno, Tadayoshi, Andre Broido, and K.C. Claffy, "Remote physical device fingerprinting," IEEE Transactions on Dependable and Secure Computing, Vol. 2, No. 2 (2005).

[100] Krystosek, Paul, Nancy Ott, Geoffrey Sanders, and Timothy Shimeall, *Network Traffic Analysis with SiLK: Analyst's Handbook for SiLK Version 3.15.0 and Later,* August 2020.

[101] Krikorian, Raffi, "Map of a Twitter Status Object," April 18, 2010.

[102] Kung, Kevin S., Kael Greco, Stanislav Sobelevsky, and Carlo Ratti, "Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data," *PLOS One,* Vol. 9, No. 6 (2014).

[103] Laperdrix, Pierre, Walter Rudametkin, and Benoit Baudry, "Beauty and the Beast: Diverting modern webbrowsers to build unique browser fingerprints," *37th IEEE Symposium on Security and Privacy,* May 2016.

[104] Renaud Lambiotte, Estaban Moro, Vincent Blondel, and Alex "Sandy" Pentland, *NetMob: Book of Abstracts,* 2019.

[105] Landau, Susan, Hubert Le Van Gong, and Robin Wilton, "Achieving Privacy in a Federated Identity Management System," *Financial Cryptography and Data Security,* 2009.

[106] Landau, Susan, "Under the Radar: NSA's Efforts to Secure Private-Sector Telecommunications Infrastructure," *Journal of National Security Law and Policy,* Vol. 7, No.3 (2014), pp. 411-442.

[107] Landau, Susan, "Transactional information is remarkably revelatory," *Proceedings of the National Academy of Sciences*, Vol. 113, No. 20 (May 17, 2016), pp. 5467-5469.

[108] Landau, Susan *Listening In: Cybersecurity in an Insecure Age,* Yale University Press, 2017.

[109] Landau, Susan and Asaf Lubin, "Examining the Anomalies, Explaining the Value: Should the USA FREEDOM Act's Metadata Program be Extended?," to appear, *Harvard National Security Journal.*

[110] Layton, Edward, *And I was There: Pearl Harbor and Midway,* William Morrow and Co., 1985.

[111] Lee, Bartholomew, "Wireless–Its Evolution from Mysterious Wonder to Weapon of War, 1902 to 1912," https://www.californiahistoricalradio.com/wp-content/uploads/2013/01/BartWirelessWar190205Lee.pdf [last viewed May 5, 2020].

[112] Leith, Douglas, "Web Browser Privacy: What Do Browsers SayWhen They Phone Home?," SCSS Technical Report, https://www.scss.tcd.ie/Doug.Leith/pubs/browser_privacy.pdf [last viewed May 17, 2020].

[113] Li, Huaxin, Zheyu Xu, Haojin Zhu, Di Ma, Shuan Li, and Kai Xing, "Demographics Inference Through Wi-Fi Network Traffic Analysis," *IEEE INFOCOM,* 2016.

[114] Lichtblau, Eric, "Police are Using Phone Tracking as a Routine Tool," *New York Times,* March 31, 2012.

[115] Lisovich, Michael, Deidre K. Mulligan, and Stephen Wicker, "Inferring Personal Information from Demand-Response Systems," *IEEE Security and Privacy,* Vol. 8, No. 1 (January 2010), 11-20.

[116] Lu, Xin David J. Wrathall, Pal Roe Sundsoye, Md. Nadiruzzaman, Erik Wetter, Asif Iqbal, Taimur Qureshi, Andrew Tatem, Geoffrey Canright, Kenth Engø-Monsen, Linus Bengtsson, "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh," *Global Environment Change,* Vol. 38 (May 2016), pp. 1-7.

[117] Manousakas, Dionysis, Cecilia Mascolo, Alastair R. Beresford, Dennis Chan, and Nikhil Sharma, "Quantifying Privacy Loss of Human Mobility Graph Topology," Privacy Enhancing Technologies Symposium, 2018.

[118] Ma, Huina, Xin Shuai. Yong-Yuel Ahn, and Yohan Bohlen, "Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Cote I'voire," EPJ Data Science, Vol. 4, Article 15 (2015).

[119] Malmi, Eric and Igmar Weber, "You Are What Apps You Use: Demographic Prediction Based on User's Apps," *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 2016.

[120] Masood, Rahat, Benjamin, Zi Hao Zhao, Hassan Jameel Asghar, and Mohamed Ali Kaafar, "Touch and You're Trapp(ck)ed: Quantifying the Uniqueness of Touch Gestures for Tracking," PETS 2018.

[121] Jonathan Mayer, Patrick Mutchler, and John C. Mitchell, "Evaluating the privacy properties of telephone metadata," *Proc Natl Acad Sci.,* Vol. 113, No. 20 (May 17, 2016), pp. 5536-5541.

[122] Matthew Mayernick and Amelia Acker, "Tracing the Traces: The Critical Role of Metadata within Networked Communications," Journal of the Association for Information Science and Technology, September 19, 2017.

[123] Microsoft, "Windows 10, version 1709 basic level Windows diagnostic events and fields," December 12, 2018 https://docs.microsoft.com/en-us/windows/privacy/basic-level-windows-diagnostic-events-and-fields-1709 [last viewed December 23, 2018].

[124] de Montjoye, Yves-Alexandre, Jordi Quoidbach, Florent Robic, and Alex (Sandy) Pentland, "Predicting Personality Using Novel Mobile Phone-Based Metrics," *Social Computing, Behavioral-Cultural Modeling and Prediction* 2013, pp. 48–55.

[125] de Montjoye, Yves-Alexandre, Cesar A. Hidalgo, Michel Verleysen, and Vincent D Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports,* Vol. 3 2013.

[126] de Montjoye1, Yves-Alexandre, Erez Shmueli, Samuel S. Wang, Alex Sandy Pentland, "openPDS: Protecting the Privacy of Metadata through SafeAnswers," *PLOS One,* Vol. 9, Issue 7 (2015).

[127] de Montjoye, Yves-Alexandre, L. Radaelli, V. K. Singh, and A. Pentland, "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," *Science,* Vol. 347, no. 6221, pp. 536-539.

[128] Moore, Tyler and Richard Clayton, "Discovering Phishing Mailboxes Using Email Metadata," Seventh APWG eCrime Researchers Summit (eCrime), Las Croabas, PR, October 2012.

[129] Motahari, Sara, Ole Mengshoel, Phyllis Reuther, Sandeep Appala, Luca Zoia, and Jay Shah, "The Impact of Social Affinity on Phone CallingPatterns: Categorizing Social Ties from Call Data Records," *Proceedings of the 6th Workshopon Social Network Mining and Analysis,* 2012.

[130] Muriello, Daniel, Stephen Heise, and Jie Chen, "Associating Users and Cameras in a Social Networking System," United States Patent 9,485,923, November 1, 2016, http://patft.uspto.gov/netacgi/nph-Parser?Sect2=PTO1&Sect2=HITOFF&p=1&u=/netahtml/PTO/search-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN/9485423 [last viewed December 22, 2018].

[131] National Research Council, *Bulk Collection of Signals Intelligence: Technical Options,* National Academies Press, 2015.

[132] National Security Agency, "Tor Stinks," https://edwardsnowden.com/docs/doc/tor-stinks-presentation.pdf [last viewed May 21, 2020].

[133] President Barack Obama, Remarks, https://edition.cnn.com/2013/06/07/politics/nsa-data-mining, June 10, 2013.

[134] Onnela, Jukka-Pekka, Samuel Arbesman, Marta C. Gonzalez, Albert-Laszlo Barabas, Nicholas A. Christakis, "Geographic Constraints on Social Network Groups," *PLOS One* (2011).

[135] Pai, Sameer, Marci Meingast, Tanya Roosta, Sergio Bermudez, Stephen B. Wicker, Deirdre K. Mulligan, Shankar Sastrym "Transactional Confidentiality in Sensor Networks," *IEEE Security and Privacy,* Vol. 6, No. 4, pp. 28-35, Jul/Aug, 2008.

[136] Private communication with the author.

[137] Court of Justice of the European Union, Judgement of the Court, *Patrick Breyer v. Germany,* October 19, 2016.

[138] Patel, Vishal, Rama Chellappa, Deepak Chama, Brandon Barbello, "Continuous User Authentication on Mobile Devices: Recent progress and remaining challenges," *IEEE Signal Processing Magazine,* Vol. 33, Issue 4, pp. 49-61, July 1, 2016.

[139] Pattara-Atikom, Wasan and Ratchata Peachavanish, "Estimating road traffic congestion from cell dwell time using neural network," *Proceedings from telecommunications, 7th international conference on ITS* (pp. 1–6).

[140] Peng, Wei and Tong Sun, "Method and system for identifying a key influencer in social media utilizing topic modeling and social diffusion analysis," US8312056B1, granted November 11, 2013.

[141] Pomerantz, Jeffrey, *Metadata,* MIT Press, 2015.

[142] Rowe, Ryan, German Creamer, Shlomo Heshkop, and Salvatore Stolfo, "Automated Social Hierarchy Detection through Email Network Analysis," *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining Social Network Analysis,* pp. 109-117, 2007.

[143] Sapiezynski, Piotr , Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann, "Tracking Human Mobility Using WiFi Signals," *PLOS One,* July 1, 2015.

[144] Science and Technology Directorate, Department of Homeland Security, *Study on Mobile Device Security,* April 2017.

[145] Sen, Subhabrata, Oliver Spatscheck, and Dongmei Wang, "Accurate, Scalable In Network Idnentification of P2P Traffic Using Application Signatures," *WWW2004,* May 17-22, 2004.

[146] Seneviratne, Suranga, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti, "Your Installed Apps Reveal Your Gender and More!," *Mobile Computing and Communications Review,* Vol. 18, No. 3 (July 2014).

[147] Sense Sleep Monitor, https://hello.is.

[148] Soto, Victor, Vanessa Frias-Martinez, Jesus Virseda and Enrique Frias-Martinez, "Prediction of Socioeconomic Levels Using Cell Phone Records," *International Conference on User Modeling, Adaptation, and Personalization,* 2011.

[149] Statista, "Smartphone penetration rate as share of the population in the United States from 2010 to 2021," https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/ [last viewed May 19, 2020].

[150] Steenbruggen, John, Maria Teresa Borzacchiello, Peter Nijkamp, and Henk Scholten, "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities," *GeoJournal,* Vol. 78 (2013), pp. 223-243.

[151] Sterbenz, James, "Intelligence in Future Broad Networks: Challenges and Opportunities in High-Speed Networks," 2002 International Zurich Seminar on Broadband Communications, 2002.

[152] Staiano, Jacopo, Fabio Pianesi, Bruno Lepri, Nicu Sebe, Nadav Aharony, and Alex Pentland, "Friends Don't Lie," *Proceedings ofthe 2012 ACM Conference on Ubiquitous Computing—UbiComp,* 2012.

[153] Subbian, Karthik, "Offline Trajectories," https://patents.justia.com/patent/10149111, December 4, 2018 [last viewed August 3, 2020].

[154] Tatem, Andrew, Youliang Qui, David L. Smith, Oliver Sabot, Abdullah S. Ali, and Bruno Moonen, "The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents," *Malaria Journal*, Vol. 8, Article number: 287 (2009).

[155] Twitter, "About public and protected tweets," https://help.twitter.com/en/safety-and-security/public-and-protected-tweets [last viewed April 28, 2020].

[156] U.S. Patent Application Publication No. 2012/0304206 (November 29, 2012)

[157] Vaccari, A., F. Dal Fiore, E. Beinat, A. Biderman, and C. Ratti, "Current Amsterdam: studying social dynamics through mobile phones network Data," *Imagining Amsterdam,*Amsterdam, Netherlands, 19–21 November 2009.

[158] In re Application of the FBI for an Order Requiring the Production of Tangible Things from Verizon Business Network Services, Inc. on Behalf of MCI Communication Services, Inc., No. BR 13- 80 (FISC Apr. 25, 2013)).

[159] Viana, Aline Carneiro, Adriano Di Luzio, Katia Jaffrès-Runser, Alessandro Mei, Julinda Stefa, "Accurately Inferring Personality Traits from the Use of Mobile Technology," 2018.

[160] Weslowski, Amy, Nathen Eagle, Andew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, Carolyn O. Buckee, "Quantifying the impact of human mobility on malaria," *Science,* Vol. 338 (2012) 267–270.

[161] Wesolowski, Amy, Nathan Eagle, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility," *Journal of the Royal Society* (2013).

[162] Wesolowski, Amy, Taimur Qureshi, Maciej F. Boni, Pal Roe Sundsøy, Michael A. Johansson, Syed Basit Rasheed, Kenth Engo-Monsen, and Caroline O. Buckee, "Impact of human mobility on the emergence of dengue epidemics in Pakistan," *Proceedings of the National Academies of Science,* September 22, 2015.

[163] Weyuker, Elaine, Thomas Ostrand, and Robert Bell, "D too many cooks spoil the broth? Using the number of developers to enhance detect prediction models?," Journal of Empirical Software Engineering, Vol. 13, Issue 5, 2008.

[164] Wilkinson, Gerard, Tom Bartindale, Tom Nappey, Michael Evans, Peter Wright, and Patrick Olivier, "Media of Things: Supporting the Production of Metadata Rich Media Through IoT Sensing," CHI, 2018.

[165] Williams, Nathelie E., Timothy A. Thomas, Matthew Dunbar, Nathan Eagle, Adrian Dobra, "Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data," *PLOS One,* Vol. 10, Issue 7 (2015).

[166] Wolff, Josephine, *You'll See this Message when it is Too Late: The Legal and Economic Aftermath of Cybersecurity Breaches,* MIT Press, 2019.

[167] Mobile Cellular Subscriptions (per 100 people), WORLD BANK, https://data.worldbank.org/indicator/IT.CEL.SETS.P2?

[168] Charles V. Wright, Lucas Ballard, Fabian Monrose, and Gerald M. Masso, "Language Identification of Encrypted VoIP Traffic: Alejandra y Roberto or Alice and Bob?," Sixteenth USENIX Security Symposium, August 2007.

[169] Ziegeldorf, Jan Henrik, Oscar Garcia Morchon, and Klaus Wehrle, "Privacy in the Internet of Things: Threats and Challenges," Security and Communication Networks, June 10, 2013.