# Evaluating the privacy properties of telephone metadata

Jonathan Mayer[a,b,1], Patrick Mutchler[a], and John C. Mitchell[a]

[a]Security Laboratory, Department of Computer Science, Stanford University, Stanford, CA 94305; and [b]Stanford Law School, Stanford University, Stanford, CA 94305

Edited by Cynthia Dwork, Microsoft Research Silicon Valley, Mountain View, CA, and approved March 1, 2016 (received for review April 27, 2015)

**Since 2013, a stream of disclosures has prompted reconsideration of surveillance law and policy. One of the most controversial principles, both in the United States and abroad, is that communications metadata receives substantially less protection than communications content. Several nations currently collect telephone metadata in bulk, including on their own citizens. In this paper, we attempt to shed light on the privacy properties of telephone metadata. Using a crowdsourcing methodology, we demonstrate that telephone metadata is densely interconnected, can trivially be reidentified, and can be used to draw sensitive inferences.**

surveillance | privacy | telephone | metadata | social network

Communications privacy law, in the United States and many other nations, draws a distinction between "content" and "metadata" (1). The former category reflects the substance of an electronic communication; the latter includes all other information about the communication, such as parties, time, and duration (2).[*]

When a government agency compels disclosure of content, the agency must usually comply with extensive substantive and procedural safeguards. Demands for metadata, by contrast, are often left to the near-total discretion of authorities. In the United States, for instance, a law enforcement officer can request telephone calling records with merely a subpoena—essentially a formal letter from the investigating agency (3). An intelligence program by the National Security Agency (NSA) has drawn particular criticism; under the business records provision of the USA PATRIOT Act (4), the agency acquired a substantial share of all domestic telephone metadata (5).[†]

In this paper, we empirically investigate factual assumptions that undergird policies of differential treatment for content and metadata. Using crowdsourced telephone logs and social networking information, we find that telephone metadata is densely interconnected, susceptible to reidentification, and enables highly sensitive inferences.[‡]

The balance of the paper is organized into three parts. First, we discuss our data collection methodology and properties of our participant population. We next present our results. Finally, we discuss implications for policy and future quantitative social science research. Additional methodological detail and figures are available in the *Supporting Information*.

## Methods

We collected the data in this study through an Android smartphone application (Fig. 1).[§] Potential participants could discover the project through academic websites, the Google Play store, and references in media coverage. The application automatically retrieved historical call and text message [Short Message Service (SMS)] metadata from device logs.[¶] In addition, the application retrieved information from a participant's Facebook account, to be used as ground truth for potential inferences.[#] Participants were provided an opportunity to view individualized features of their phone metadata, and then they were invited to uninstall the application. In total, 823 participants volunteered their metadata, which included 251,788 calls and 1,234,231 text messages. The *Supporting Information* provides additional detail on data sources and dataset properties (Figs. S1–S5 and *1. Dataset Methodology*, *1.1. Data Collection*, *1.2. Participants*, *1.3. Logs*, and *1.4. Sampling Bias*).

**Ethical Considerations.** Given the quantity and sensitivity of the data associated with this project, we instituted several informed consent mechanisms. Participants received extensive disclosure notices, both in the application and on the study website. In addition, the Facebook software library notified participants of the categories of social network information that the application was requesting. Each screen of the application, until information upload was complete, provided participants with an opportunity to withdraw. Furthermore, participants were furnished contact information for research staff such that they could request deletion of their information after using the application. The university institutional review board suggested helpful methodological refinements, and we began collecting data only after receiving the board's approval.

We also took a number of security precautions to safeguard participant information. Our application transmitted information to a cloud storage service only over an encrypted and authenticated connection [transport layer security (TLS)], and we retrieved information only over TLS. Credentials for accessing the data were restricted to the research team, and once the data were retrieved, the data were stored on encrypted devices at academic facilities.

**Dataset.** We provide a detailed treatment of our dataset in the *Supporting Information*. We note here, importantly, that our crowdsourced dataset is not a

## Significance

**Privacy protections against government surveillance are often scoped to communications content and exclude communications metadata. In the United States, the National Security Agency operated a particularly controversial program, collecting bulk telephone metadata nationwide. We investigate the privacy properties of telephone metadata to assess the impact of policies that distinguish between content and metadata. We find that telephone metadata is densely interconnected, can trivially be reidentified, enables automated location and relationship inferences, and can be used to determine highly sensitive traits.**

Author contributions: J.M., P.M., and J.C.M. designed research; J.M. and P.M. performed research; J.M. and P.M. contributed new analytic tools; J.M. and P.M. analyzed data; and J.M., P.M., and J.C.M. wrote the paper.

Conflict of interest statement: C.D. was jointly funded with J.C.M. on a 3-year Sloan Foundation grant in 2011, on a no-cost extension through 2015.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 5467.

[1] To whom correspondence should be addressed. Email: jmayer@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1508081113/-/DCSupplemental.

[*] The contours of the content–metadata distinction are well established for telephony and messaging, but are far more elusive for newer forms of communication.

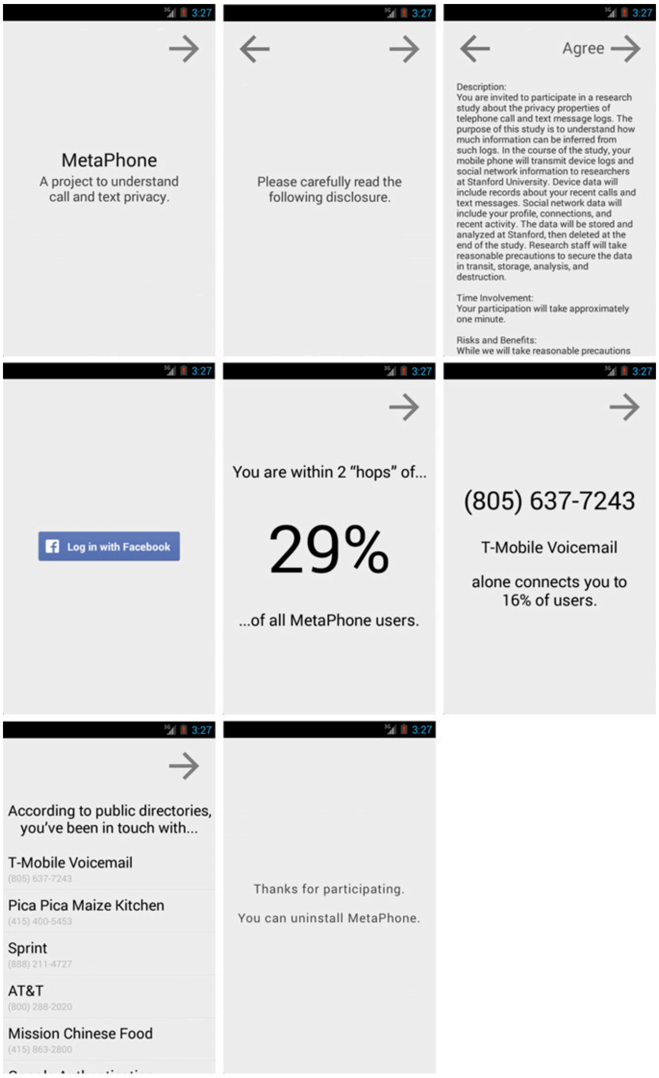
[†] While this article was in submission, Congress enacted the USA FREEDOM Act (6). Provisions codify the two-hop limit voluntarily imposed by the executive branch, as well as the proposed 18-month-duration limit. Data that are not associated with a query result will also remain with telecommunications services. These changes took effect on November 29, 2015.

[‡] In the interest of providing timely input on matters of public controversy, we presented our preliminary results in a series of online postings (webpolicy.org/2013/11/27/metaphone-seeing-someone/, webpolicy.org/2013/12/09/metaphone-the-nsa-three-hop/, webpolicy.org/2013/12/23/metaphone-the-nsas-got-your-number/, and webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata/).

[§] We initially approached several telecommunications providers about collaboration. All declined.

[¶] Metadata included the time of the call or SMS, whether the call or SMS was incoming or outgoing, the other phone number participating in the call or SMS, and the length (in seconds) of the call or the length (in characters) of the SMS.

[#] Facebook information included age, gender, relationship status, political leanings, religious affiliation, occupation, location, and interests.



MetaPhone
A project to understand call and text privacy.

Please carefully read the following disclosure.

Description:
You are invited to participate in a research study about the privacy properties of telephone call and text message logs. The purpose of this study is to understand how much information can be inferred from such logs. In the course of the study, your mobile phone will transmit device logs and social network information to researchers at Stanford University. Device data will include records about your recent calls and text messages. Social network data will include your profile, connections, and recent activity. The data will be stored and analyzed at Stanford, then deleted at the end of the study. Research staff will take reasonable precautions to secure the data in transit, storage, analysis, and destruction.

Time Involvement:
Your participation will take approximately one minute.

Risks and Benefits:
While we will take reasonable precautions

You are within 2 “hops” of...
29%
...of all MetaPhone users.

(805) 637-7243
T-Mobile Voicemail
alone connects you to 16% of users.

According to public directories, you’ve been in touch with...

| | |
|---|---|
| T-Mobile Voicemail | (805) 637-7243 |
| Pica Pica Maize Kitchen | (415) 400-5453 |
| Sprint | (888) 211-4727 |
| AT&T | (800) 288-2020 |
| Mission Chinese Food | (415) 863-2800 |

Thanks for participating.
You can uninstall MetaPhone.

**Fig. 1.** Example user experience flow in MetaPhone, the Android application that we developed to crowdsource a telephone metadata dataset.

random, representative sample of the US population. Participant requirements and recruiting introduced biases, reflected in skewed demographics. In addition, participant Facebook profiles did not include a uniform set of fields. Our results are, however, strongly suggestive of properties in the larger population. The privacy phenomena that we observe are not subtle, and their causes are generalizable.

## Results

In the following sections, we pose open questions about the privacy properties of telephone metadata. We then use our crowdsourced data to provide approximate empirical answers.

**Graph Structure.** Certain metadata surveillance programs impose a “hop” constraint, most notably the NSA’s domestic telephone program (7, 8).[‖] After accessing metadata on a suspected (“seed”) telephone number, an analyst can retrieve records for numbers one or more edges (“hops”) distant in a connectivity graph.[**] These restrictions are intended to constrain the volume of metadata that an agency can access. Although the NSA program initially allowed three hops, executive officials scaled it back to two hops following criticism (9).

Durational limits are another form of constraint on metadata surveillance. In the NSA’s program, analysts can retrieve metadata for 5 years prior. A revision to the program, proposed by the White House, would shorten the accessible history to 18 months—the current retention period under federal communications regulation (10).

Our dataset enables us to quantify the impact of these surveillance limitations. We begin with a discussion of the structure of the telephone connectivity graph, then describe how we accounted for longitudinal considerations, and finally quantitatively assess the efficacy of these constraints.

Prior work on telephone graphs has emphasized a small-world network topology (11), largely treating the graphs as diffuse social networks. The literature emphasizes monotonic, heavy-tailed degree distributions, and especially power law distributions (12–19).

Our results are broadly consistent, with two refinements. We find that at the low end of node degree, among participants, probability density includes a peak and a one-sided heavy tail (Fig. 2, Fig. S6, and *2. Graph Structure and Analysis Methodology*, *2.1. Individual Participant Structure*). The intuitive explanation is that a small proportion of telephone subscribers makes essentially no telephone use, and another small proportion makes unusually heavy use. In future work, a nonmonotonic distribution—such as a variant of a log-normal distribution—would better approximate individual telephone use behavior (see ref. 18).

More importantly, we find that at the high end of node degree, there are hubs that connect meaningful proportions of the entire participant population (Fig. S7 and *2. Graph Structure and Analysis Methodology*, *2.2. Hub Structure*). These widely shared telephone numbers include customer service lines, Voice over Internet Protocol (VoIP) bridges, two-factor authentication services, and telemarketers (Fig. S8). Critically, for purposes of surveillance regulation, these high-degree nodes establish two-hop paths between large volumes of individual telephone subscribers (Fig. S9).

Because participants varied in the duration of telephone logs that they provided, and because some surveillance programs (including the NSA’s) extend beyond the time window of our dataset, we are compelled to extrapolate a longitudinal distribution of participant degree. We accomplished this by fitting curves to longitudinal degree data (Fig. S10 and *2. Graph Structure and Analysis Methodology*, *2.3. Estimating the Effects of Surveillance Regulation*).

With these preliminaries, we are able to quantitatively estimate the reach of a telephone metadata surveillance program under particular hop and duration limits. Fig. 3*A* depicts expected reach

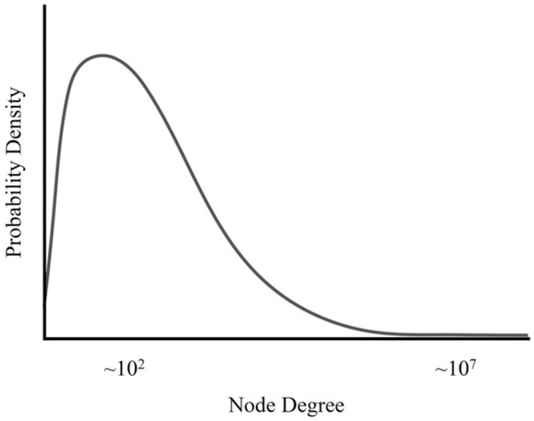


**Fig. 2.** Notional distribution of node degree in the telephone call and text message graphs, over approximately 1 year.

[‖] Our description emphasizes telephone metadata because that component of NSA bulk surveillance has been declassified. Officials have neither confirmed nor denied bulk surveillance of text messages.

[**] Our understanding of the NSA program is that, at each hop, an analyst can retrieve the subscriber’s communications records. Disclosures have not been entirely clear on this point.

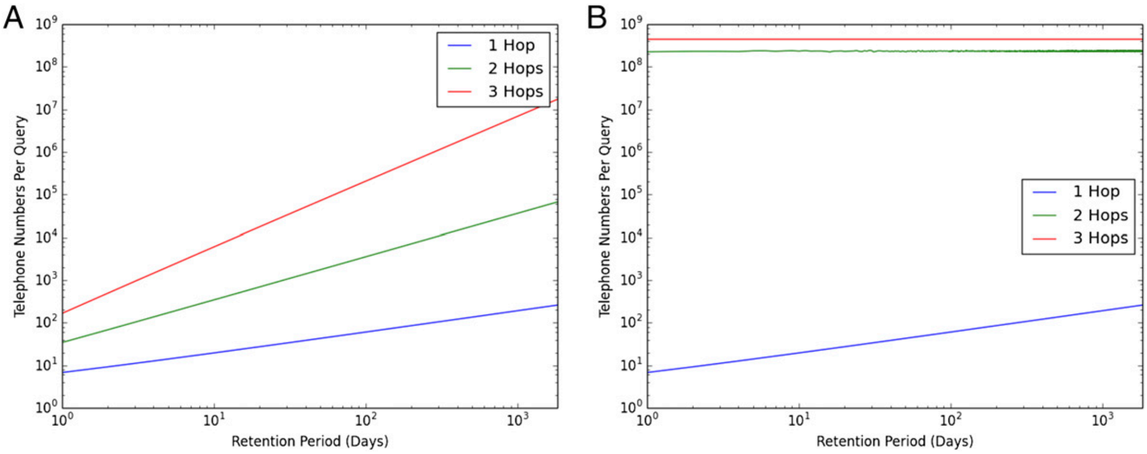


**Fig. 3.** Approximation of expected surveillance authority reach with one seed, in a combined US call and text message graph. (*A*) Naïve approach, assuming solely individual subscribers. (*B*) Bootstrapping approach, incorporating national hubs.

against time and hop count, assuming the call graph only includes individual subscribers. Fig. 3*B* uses a bootstrapping algorithm to incorporate the effects of high-degree hubs (Algorithm S1 and *2. Graph Structure and Analysis Methodology*, *2.3. Estimating the Effects of Surveillance Regulation*). An additional 3D surface visualization is provided in Fig. S11 and *2. Graph Structure and Analysis Methodology*, *2.3. Estimating the Effects of Surveillance Regulation*.

Applied to the NSA's program, our results strongly suggest that until 2013, analysts had legal authority to access telephone records for the majority of the entire US population. Under the more recent two-hop rule, the proposed 18-month-retention period, and an assumption that national and local hub numbers are removed from the call graph,[††] an analyst could in expectation access records for ∼25,000 subscribers with a single seed.

**Reidentification.** One of the chief defenses of metadata surveillance programs, including the NSA's, has been that the information is not identified. By relying on data that are not "personally identifiable information" (PII), the argument goes, metadata programs have a lesser privacy impact.[‡‡]

Prior work has demonstrated that the policy distinction between PII and non-PII is not based on sound science. Researchers have demonstrated "reidentification" risks in a number of applications, including health records (21, 22), location histories (23–25), web search queries (26), web browsing activity (27–29), movie reviews (30), and social network graphs (31, 32).

We contribute to this literature with an unsurprising result: telephone numbers are trivially reidentifiable. We conducted both automated and manual attempts at reidentification, and we found that both approaches were highly successful.

To quantify the feasibility of automated telephone number reidentification, we leveraged existing directory, search, and social network application programming interfaces (APIs). We randomly selected 30,000 numbers from our dataset and queried free, public interfaces hosted by Yelp, Google Places, and Facebook using these numbers. This approach matched identities for 9,576 (32%) of the numbers (Table 1). Matches included both businesses (from Yelp and Google Places) as well as individuals (from Facebook). These results are necessarily conservative; with access to commercial databases, a business or government agency would be able to achieve substantially higher match rates.

To assess the efficacy of manual reidentification, we randomly selected 250 phone numbers from our dataset and used two separate strategies for manual reidentification. First, we used a manual query interface for an inexpensive commercial database (Intelius). Second, we performed manual Google web searches and examined the results for identifying information. In total, we spent $19.95 for a month subscription to Intelius and 70-min running web searches. With these limited resources—far below those available to a large business or intelligence agency—we were still able to identify the overwhelming majority of the numbers (Table 2).

**Location Inferences.** The policy and law surrounding telephone metadata has conventionally distinguished call and text records from mobile location records. We used our dataset to investigate the extent to which location could be inferred from calls and text messages.

Prior work on mobile phone location has relied upon precise and dense Global Positioning System (GPS), wireless network, and cell tower measurements, using them to predict personal locations and movement patterns between those locations (33–35). In comparison, we show that home locations can often be predicted using imprecise and sparse telephone metadata. We accomplish this in two steps: (*i*) locating the businesses in a participant's phone logs using the reidentification techniques described above; and (*ii*) using those business locations to predict home locations.

Both Yelp and Google Places provide street addresses for reidentified businesses. We determined the latitude and longitude of these addresses using the Google Geocoding API. Following the intuition that most of the businesses an individual calls are clustered around their home, we used the DBSCAN algorithm (36) to find the largest cluster of calls based on business location information. We then predicted home location at the median latitude and longitude of the cluster.

**Table 1. Performance of telephone number reidentification (automated approaches)**

| Look-up source | Matched, % |
|---|---|
| Google Places | 16.6 |
| Yelp | 10.5 |
| Facebook | 13.7 |
| All Automated Sources | 31.9 |

[††] The Foreign Intelligence Surveillance Court has authorized the NSA to identify high-degree nodes (e.g., ref. 5). It is not apparent whether the NSA elects to eliminate these nodes when marking portions of the call graph as eligible for analysis, or whether the NSA merely eliminates these nodes when conducting subsequent analysis.

[‡‡] Definitions of PII vary. Some authorities do consider telephone numbers to be PII (e.g., ref. 20).

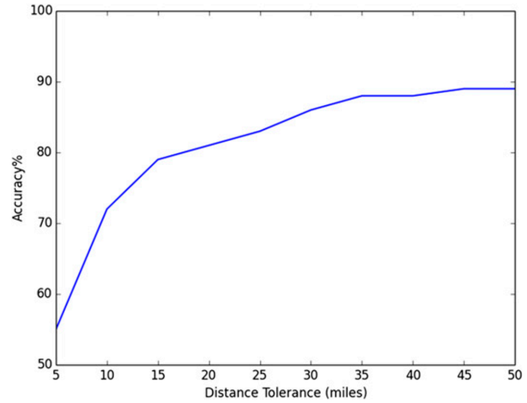


**Fig. 4.** Performance of automated home location prediction.

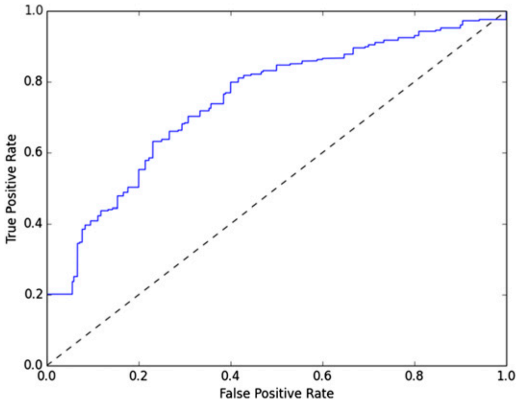


**Fig. 5.** Average performance of automated personal relationship prediction.

Among participants in our study, 418 listed a current city on Facebook.[§§] Of these participants, 241 (60%) had at least 10 calls to reidentified businesses. We were able to correctly predict the Facebook current city of 130 (57%) participants using the method described above. Fig. 4 presents the prediction accuracy at varying distance tolerances, measured from the center of a participant's current city.

**Relationship Inferences.** Another policy concern surrounding telephone metadata is that the metadata could be used to infer categories of interpersonal relationships. To understand the feasibility of drawing such inferences with an automated methodology, at scale, we focused on studying romantic relationships.

Prior work has applied supervised learning to a small sample of smartphone sensor and usage data and achieved good performance at predicting marital status (37, 53).[¶¶,##,∥∥] Related research has also demonstrated the feasibility of inferring relationship status from an online social network graph (38).

We built a classifier for whether a person was in a relationship, based on his or her call and text records. We began by selecting participants who were, according to their Facebook profile, single ($N = 148$) or in a relationship ($N = 309$). We then generated a range of features from telephone metadata and trained a support vector machine (*3. Relationship Inference Methodology*). Fig. 5 depicts the receiver operating characteristic for the resulting classifier.

Once a participant was labeled as in a relationship, we found that identifying the participant's partner was trivial. We tested several heuristics against the subset of participants with an identified relationship partner ($N = 211$) and achieved good performance (Table 3).

In sum, it appears feasible—with further refinement—to draw Facebook-quality relationship inferences from telephone metadata.

**Sensitive Trait Inferences.** Perhaps the greatest policy concern surrounding telephone metadata has been the possibility of drawing sensitive inferences. The issue is neatly encapsulated in a pair of December 2013 federal court opinions. One judge invalidated the NSA program, noting that "metadata from each person's phone 'reflects a wealth of detail about her familial, political, professional, religious, and sexual associations'" (39, 40). Less than 2 weeks later, another judge sustained the NSA program, dismissing sensitive inferences as merely a "parade of horribles" (41).

Data privacy researchers have not been so divided. In academic publications (e.g., ref. 42), court filings (e.g., ref. 40), and opinion pieces (e.g., ref. 43), scholars have persuasively argued that telephone metadata is highly sensitive. These claims have, however, been largely based on hypothetical scenarios and experiential intuition—not empirical results.

The closest related work has attempted inferences from a range of mobile phone features, including communications records, location estimates, and (in some papers) sensor and application logs. Prior results have suggested the feasibility of inferring age, gender, employment, and personality from these mobile phone data sources (refs. 37, 44–48 and ¶¶, ##, and ∥∥). Our study, motivated by the NSA's program and the state of American law, examines only call and text records.[***] We also attempt to draw particularly precise and particularly sensitive inferences about the participants in our study.[†††]

Using our dataset of reidentified phone numbers, we estimated the feasibility of drawing sensitive inferences from phone metadata. As with the reidentification task, we include results from both automated and manual approaches.

Automated inferences can be made directly from the results of Google Places and Yelp queries, which include business category information in their results. By labeling certain categories as sensitive, we identified the portion of participants that made a call or text to a potentially sensitive organization. Table 4 shows the portion of participants that made calls or texts to organizations matching sensitive categories.

Health Services was the most common category of sensitive organization. We further labeled medical specialist subsets of this category using more precise labels obtained from Google and Yelp queries. Table 5 shows the specialist categories that appear in at least 1% of participants' call logs.

Calls to religion-affiliated numbers provided an opportunity to validate the accuracy of automated sensitive inferences. A subset of participants both placed a call to a religious group and provided a religion on Facebook ($N = 18$). Among these, the most-called religious group overwhelmingly matched the Facebook religion ($N = 14$).

Our results suggest that, even without human review, a business or agency could draw sensitive inferences from a significant share of telephone records.

To simulate the inferences that might be drawn from manual telephone record analysis, we focused on participants who held a high proportion of their phone conversations with sensitive numbers. We then applied our automated and manual reidentification

[§§] This information is self-reported and may be out of date; future work would benefit from a more reliable source of location data.

[¶¶] Brdar S, Ćulibrk D, Crnojević V, Demographic attributes prediction on the real-world mobile data, Mobile Data Challenge 2012 (by Nokia) Workshop, June 18–19, 2012, Newcastle, UK.

[##] Mohrehkesh S, Ji S, Nadeem T, Weigle MC, Demographic prediction of mobile user from phone usage, Mobile Data Challenge 2012 (by Nokia) Workshop, June 18–19, 2012, Newcastle, UK.

[∥∥] Ying JJC, Chang YJ, Huang CM, Tseng VS, Demographic prediction based on user's mobile behaviors, Mobile Data Challenge 2012 (by Nokia) Workshop, June 18–19, 2012, Newcastle, UK.

[***] The majority view in American courts is that the Fourth Amendment does not protect mobile phone location records. As a matter of statute, a court order is still required to obtain those records.

[†††] As a natural consequence of attempting such specific inferences, we examine fewer types of inference and draw inferences with lesser reliability than in prior work.

Table 2. Performance of telephone number reidentification (manual and combined approaches)

| Look-up source | Matched, % |
|---|---|
| Intelius | 65 |
| Google search | 58 |
| All automated sources | 26 |
| All sources | 82 |

Table 3. Performance of relationship partner identification heuristics

| Heuristic, maximum | Accuracy, % |
|---|---|
| Calls | 81 |
| Call duration | 45 |
| Days with a call | 77 |
| Texts | 76 |
| Text length | 68 |
| Days with a text | 76 |

Table 4. Participant interaction with sensitive organizations

| Category | Participants with ≥1 calls, % |
|---|---|
| Health services | 57 |
| Financial services | 40 |
| Pharmacies | 30 |
| Veterinary services | 18 |
| Legal services | 10 |
| Recruiting and job placement | 10 |
| Religious organizations | 8 |
| Firearms sales and repair | 7 |
| Political officeholders and campaigns | 4 |
| Adult establishments | 2 |
| Marijuana dispensaries | 0.4 |

Table 5. Participant interaction with health organizations

| Category | Participants with ≥1 calls, % |
|---|---|
| Dentistry and oral health | 18 |
| Mental health and family services | 8 |
| Ophthalmology and optometry | 6 |
| Sexual and reproductive health | 6 |
| Pediatrics | 5 |
| Orthopedics | 4 |
| Chiropractic care | 3 |
| Rehabilitation and physical therapy | 3 |
| Medical laboratories | 2 |
| Emergency or urgent care | 2 |
| Hospitals | 2 |
| Cardiology | 2 |
| Dermatology | 1 |
| Ear, nose, and throat | 1 |
| Neurology | 1 |
| Oncology | 1 |
| Substance abuse | 1 |
| Cosmetic surgery | 1 |

approaches, attempting to identify as many of each participant's contacts as possible.[‡‡‡]

The following vignettes are reflective of the types of inferences we were able to draw.

*i*) Participant A held conversations with a pharmacy specializing in chronic care, a patient service that coordinates management for serious conditions, several local neurology practices, and a pharmaceutical hotline for a prescription drug used solely to manage the symptoms and progression of relapsing-remitting multiple sclerosis.

*ii*) Participant B received a long phone call from the cardiology group at a regional medical center, talked briefly with a medical laboratory, answered several short calls from a local drugstore, and made brief calls to a self-reporting hotline for a cardiac arrhythmia monitoring device.

*iii*) Participant C placed frequent calls to a local firearm dealer that prominently advertises a specialty in the AR semiautomatic rifle platform. He also placed lengthy calls to the customer support hotline for a major firearm manufacturer; the manufacturer produces a popular AR line of rifles.

*iv*) Participant D placed calls to a hardware outlet, locksmiths, a hydroponics store, and a head shop in under 3 weeks.

*v*) Participant E made a lengthy phone call to her sister early one morning. Then, 2 days later, she called a nearby Planned Parenthood clinic several times. Two weeks later, she placed brief additional calls to Planned Parenthood, and she placed another short call 1 month after.

Using public sources, we were able to confirm that participant B had a cardiac arrhythmia and participant C owned an AR rifle. As for the remaining inferences, regardless of whether they were accurate, the mere appearance of possessing a highly sensitive trait assuredly constitutes a serious privacy impact.[§§§]

Our results lend strong support to the view that telephone metadata is extraordinarily sensitive, especially when paired with a broad array of readily available information. For a randomly selected telephone subscriber, over a short period, drawing these sorts of sensitive inferences may not be feasible. However, over a large sample of telephone subscribers, over a lengthy period, it is inevitable that some individuals will expose deeply sensitive information. It follows that large-scale metadata surveillance programs, like the NSA's, will necessarily expose highly confidential information about ordinary citizens.

## Discussion

The results of our study are unambiguous: there are significant privacy impacts associated with telephone metadata surveillance. Telephone metadata is densely interconnected, easily reidentifiable, and trivially gives rise to location, relationship, and sensitive inferences. In combination with independent reviews that have found bulk metadata surveillance to be an ineffective intelligence strategy (7, 8), our findings should give policymakers pause when authorizing such programs.

More broadly, this project emphasizes the need for scientifically rigorous surveillance regulation. Much of the law and policy that we explored in this research was informed by assumption and conventional wisdom, not quantitative analysis. To strike an appropriate balance between national security and civil liberties, future policymaking must be informed by input from the relevant sciences.

Our results also bear on commercial data practices. It is routine practice for telecommunications firms to collect, retain, and transfer subscriber telephone records, often dubbed "Customer Proprietary Network Information" (49, 50). Telecommunications regulation should also incorporate a scientifically rigorous understanding of the privacy properties of these data.

There remains much future work to be done in this space. To conduct this study, we were compelled to rely on a small and

[‡‡‡] Although several of these participants consented to being identified in this publication, out of recognition for the associated privacy risks, we use only pseudonyms.

[§§§] More generally, a probabilistic sensitive inference—even with less than even likelihood—could constitute a significant privacy risk.

unrepresentative dataset. Future efforts would benefit from population-scale data; the challenges are in sourcing the data, not computing on them. Future work could also pair telephone records with more comprehensive ground truth than the Facebook data we accessed. Subscriber records and cell site location information, for instance, would better enable testing for inferences. Another potential direction is testing more advanced approaches to automated inferences; the machine-learning techniques we applied in this study were effective, although relatively rudimentary.

**ACKNOWLEDGMENTS.** We thank the participants in this study, who altruistically gave up their privacy, such that the public might better understand government surveillance. We also thank Dan Boneh and Edward Felten for invaluable suggestions. This project was supported by the National Science Foundation Team for Research in Ubiquitous Secure Technology Research Center.

1. Electronic Communications Privacy Act, 18 U.S. Code Sect. 2510(8) (2012).
2. United States v. Forrester, 512 F.3d 500, 9th Cir. (July 6, 2007).
3. Stored Communications Act, 18 U.S. Code Sect. 2703(c)(2) (2012).
4. USA PATRIOT Act, 50 U.S. Code Sect. 1861 (2012).
5. In re FBI Application, No. BR 13-109, FISA Ct. (August 29, 2013).
6. USA FREEDOM Act, Pub. L. No. 114-23, 129 Stat. 268 (June 2, 2015).
7. President's Review Group (December 12, 2013) *Liberty and Security in a Changing World* (President's Review Group on Intelligence and Communications Technologies, Washington, DC). Available at www.whitehouse.gov/sites/default/files/docs/2013-12-12_rg_final_report.pdf. Accessed April 2016.
8. Privacy and Civil Liberties Oversight Board (2014) *Report on the Telephone Records Program Conducted Under Section 215 of the USA PATRIOT Act and on the Operations of the FISC* (Privacy and Civil Liberties Oversight Board, Washington, DC).
9. Privacy and Civil Liberties Oversight Board (2015) *Recommendations Assessment Report* (Privacy and Civil Liberties Oversight Board, Washington, DC).
10. Retention of Telephone Toll Records, 47 C.F.R. Sect. 42.6 (2015).
11. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442.
12. Abello J, Pardalos PM, Resende MGC (1998) On maximum clique problems in very large graphs. *External Memory Algorithms. DIMACS Workshop: External Algorithms and Visualization, May 20–22 1998*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, eds Abello JM, Vitter JS (American Mathematical Society, Providence, RI), pp 119–130.
13. Abello J, Resende MG, Sudarsky S (2002) Massive quasi-clique detection. *LATIN 2002: Theoretical Informatics, 5th Latin American Symposium, Cancun, Mexico April 2002 Proceedings*, Lecture Notes in Computer Science, ed Rajsbaum S (Springer, Berlin), Vol 2286, pp 598–612.
14. Aiello W, Chung F, Lu L (2001) A random graph model for power law graphs. *Expo Math* 10(1):53–66.
15. Nanavati AA, et al. (2006) On the structural properties of massive telecom call graphs: findings and implications, Conference on Information and Knowledge Management, CIKM '06, November 5–11, 2006, Arlington, VA (Association for Computing Machinery, New York), pp 435–444.
16. Onnela JP, et al. (2007) Structure and tie strengths in mobile communication networks. *Proc Natl Acad Sci USA* 104(18):7332–7336.
17. Pandit V, et al. (2008) Extracting dense communities from telecom call graphs. *Third International Conference on Communication Systems Software and Middleware and Workshops, COMSWARE 2008* (IEEE, New York), pp 82–89.
18. Seshadri M, et al. (2008) Mobile call graphs: beyond power-law and lognormal distributions, The 14th ACM SIGKDD International Conference on Knowledge, KDD '08, August 24–27, 2008, Las Vegas (Association for Computing Machinery, New York), pp 596–604.
19. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. *Science* 324(5930):1071–1076.
20. California Online Privacy Protection Act, Cal. Bus. and Prof. Code Sect. 22577 (2012).
21. Sweeney L (2002) k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst* 10(5):557–570.
22. Sweeney L, Abu A, Winn J (2013) *Identifying Participants in the Personal Genome Project by Name, Technical Report 1021-1* (Harvard Univ Data Privacy Lab, Cambridge, MA).
23. Golle P, Partridge K (2009) On the anonymity of home/work location pairs. *Proceedings of the 7th International Conference on Pervasive Computing* (Springer, Berlin), pp 390–397.
24. Zang H, Bolot J (2011) Anonymization of location data does not work: A large-scale measurement study. *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, MobiCom '11* (Association for Computing Machinery, New York), pp 145–156.
25. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the Crowd: The privacy bounds of human mobility. *Sci Rep* 3:1376.
26. Ohm P (2010) Broken promises of privacy. *UCLA Law Rev* 57:1701–1777.
27. Krishnamurthy B, Naryshkin K, Wills C (2011) Privacy leakage vs. Protection measures: The growing disconnect. *IEEE Secur Priv* 11(3):14–20.
28. Mayer JR, Mitchell JC (2012) Third-party web tracking: Policy and technology. *2012 IEEE Symposium on Security and Privacy (SP), May, 20–23, 2012, San Francisco* (IEEE, New York), pp 413–427.
29. Englehardt S, et al. (2015) Cookies that give you away: The surveillance implications of Web tracking. *Proceedings of the 24th International Conference on World Wide Web, WWW '15* (Association for Computing Machinery, New York), pp 289–299.
30. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. *Proceedings of the 2008 IEEE Symposium on Security and Privacy, SP '08* (IEEE Computer Society, Washington, DC), pp 111–125.
31. Backstrom L, Dwork C, Kleinberg J (2007) Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. *Proceedings of the 16th International Conference on World Wide Web, WWW '07* (Association for Computing Machinery, New York), pp 181–190.
32. Narayanan A, Shmatikov V (2009) De-anonymizing social networks. *30th IEEE Symposium on Security and Privacy* (IEEE, New York), 173–187.
33. Zheng Y, Li Q, Chen Y, Xie X, Ma WY (2008) Understanding mobility based on GPS data. *Proceedings of the 10th International Conference on Ubiquitous Computing, UbiComp '08* (Association for Computing Machinery, New York), pp 312–321.
34. Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from GPS trajectories. *Proceedings of the 18th International Conference on World Wide Web, WWW '09* (Association for Computing Machinery, New York), pp 791–800.
35. Anagnostopoulos T, Anagnostopoulos C, Hadjiefthymiades S (2012) Efficient location prediction in mobile cellular networks. *Int J Wirel Inf Networks* 19(2): 97–111.
36. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Data Mining and Knowledge Discovery* 2(2):169–194.
37. Zhong E, Tan B, Mo K, Yang Q (2013) User demographics prediction based on mobile data. *Pervasive and Mobile Computing* 9(6):823–837.
38. Backstrom L, Kleinberg J (2014) Romantic partnerships and the dispersion of social ties. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '14* (Association for Computing Machinery, New York), pp 831–841.
39. Klayman v. Obama, 957 F. Supp. 2d 1, Dist. Ct. DC (December 16, 2013).
40. Declaration of Professor Edward W. Felten, ACLU v. Clapper, No. 13-cv-03994, Southern Dist. Ct. NY (August 26, 2013).
41. ACLU v. Clapper, 959 F. Supp. 2d 724, Southern Dist. Ct. NY (December 27, 2013).
42. Landau S (2013) Making sense from Snowden. *IEEE Secur Priv* 11(4):54–63.
43. Blaze M (June 19, 2013) Phew, NSA is just collecting metadata. *Wired*. Available at www.wired.com/2013/06/phew-it-was-just-metadata-not-think-again.
44. Chittaranjan G, Blom J, Gatica-Perez D (2013) Mining large-scale smartphone data for personality studies. *Pers Ubiquitous Comput* 17(3):433–450.
45. de Montjoye YA, Quoidbach J, Robic F, Pentland A (2013) Predicting Personality Using Novel Mobile Phone-Based Metrics. *Social Computing, Behavioral-Cultural Modeling and Prediction*, Lecture Notes in Computer Science (Springer, Berlin) Vol 7812, pp 48–55.
46. Arai A, et al. (2014) Understanding user attributes from calling behavior. *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia, MoMM'14* (Association for Computing Machinery, New York), pp 95–104.
47. Jahani E, et al. (2015) Predicting gender from mobile phone metadata, NetMob 2015, April 7–10, 2015, Cambridge, MA, eds Moro E, de Montjoye Y-A, Blondel V, Pentland A, pp 110–113.
48. Toole JL, et al. (2015) Tracking employment shocks using mobile phone data. *J R Soc Interface* 12(107):pii: 20150185.
49. The Telecommunications Act, 47 U.S. Code Sect. 222 (2012).
50. Customer Proprietary Network Information, 47 C.F.R. Sects. 64.2001–64.2011 (2015).
51. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA* 106(36):15274–15278.
52. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland A (2014) Daily stress recognition from mobile phone data, weather conditions and individual traits. *Proceedings of the 22nd ACM International Conference on Multimedia, MM'14* (Association for Computing Machinery, New York), pp 477–486.
53. Laurila JK, et al. (2013) From big smartphone data to worldwide research: The mobile data challenge. *Pervasive and Mobile Computing* 9(6):752–771.
54. Apple (2015) *ResearchKit Technical Overview* (Apple, Cupertino, CA). Available at researchkit.org/docs/docs/Overview/GuideOverview.html. Accessed April 2015.
55. The White House Office of the Press Secretary (2014) *Presidential Policy Directive/PP-28* (The White House, Washington, DC).
56. Dwork C (2006) Differential privacy. *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, Lecture Notes in Computer Science (Springer, Berlin), Vol 4052, pp 1–12.
57. Mayer J, Narayanan A (2013) Privacy substitutes. *Stanford Law Rev Online* 66:89–96.
58. Pew Research Center (2015) *The Smartphone Difference* (Pew Research Center, Washington, DC), Available at www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015. Accessed April 2015.
59. Federal Communications Commission (2014) *Local Telephone Competition: Status as of December 31, 2013* (Federal Communications Commission, Washington, DC).
60. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.