Catching Bandits and *Only* Bandits: Privacy-Preserving Intersection Warrants for Lawful Surveillance

Aaron Segal, Bryan Ford, and Joan Feigenbaum
Yale University

Abstract

Motivated in part by the Snowden revelations, we address the question of whether intelligence and law-enforcement agencies can gather actionable, relevant information about unknown electronic targets without conducting dragnet surveillance. We formulate principles that we believe effective, lawful surveillance protocols should adhere to in an era of big data and global communication networks. We then focus on intersection of cell-tower dumps, a specific surveillance operation that the FBI has used effectively. As a case study, we present a system that computes such intersections in a *privacy-preserving, accountable* fashion. Preliminary experiments indicate that such a system could be efficient and usable, suggesting that privacy and accountability need not be barriers to effective intelligence gathering.

1 Introduction

Much of the Snowden-triggered debate has revolved around the "balance" between national security and personal privacy. Both sides of these "balance" arguments presume that security and privacy represent a zero-sum tradeoff, a presumption that we believe is false – not just on policy grounds [16] but also for technical reasons. With the right *existing* technology deployed under the right policy framework, we can have both strong national security and strong privacy protections [8].

We observe and accept that certain demonstrably effective electronic-surveillance processes require "bulk" access to privacy-sensitive metadata. For example, by obtaining *cell-tower dumps* from related bank robbery sites – sets of about 150,000 total users whose cell phones had been in the area at particular times – the FBI *intersected* these sets to discover a phone used by the High Country Bandits [2]. Although the FBI's dragnet proved effective in catching these particular criminals, their incidental ingestion of many innocent users' phone numbers raises important privacy concerns.

Consistent with both US Constitutional and human-rights principles that allow government "search and seizure" in private spaces only via warrant processes grounded in *public* law, we propose that any electronic-surveillance activity searching or otherwise touching private user data or metadata must likewise be implemented via *open, public* processes that protect the privacy of innocent, untargeted users. These open processes can and should, for example, enable agencies like the FBI to catch criminals such as the High Country Bandits without ingesting 149,999 unrelated users' phone numbers into internal databases for potential use in arbitrary, secret surveillance activities now or in the future.

This paper takes preliminary steps toward enunciating basic principles for *open, privacy-preserving, accountable surveillance processes* and explains why this phrase is not an oxymoron. As a concrete case study, we present a prototype metadata-query system based on mature and practical *privacy-preserving set-intersection* methods [9, 13, 18]. This design supports warrant-based surveillance targeting not just known but *unknown* users, as in the FBI's targeting of the High Country Bandits and the NSA's targeting of the CO-TRAVELERs of terrorism suspects [17]. Our preliminary experimental results suggest that intersection of cell-tower dumps can indeed be implemented in a manner consistent with the principles of privacy-preserving, accountable surveillance.

Before proceeding, we wish to address the question of why "privacy-preserving, accountable surveillance" is an appropriate topic for a workshop on "free and open communications on the Internet." While it may be interesting and appealing to contemplate an Internet in which there is little or no surveillance, it would not be an effective way to increase the degree to which "Internet freedom" is a lived experience for ordinary people. Law-enforcement and intelligence agencies have been and currently are active in *every* national- or global-scale mass-communication system, and the Internet will be no exception. The Snowden revelations may have provided an opportunity to design protocols that allow government agencies to collect and use data that are demonstrably relevant to their missions *while respecting the privacy of ordinary citizens and being democratically accountable*. The FOCI community should seize that opportunity.

In Section 2, we present principles that could guide the development of privacy-preserving, accountable surveil-

lance protocols; we also explain why the intersection of cell-tower dumps is a natural domain in which to apply these principles. In Section 3, we flesh out our operational model of privacy-preserving, accountable set intersection and present the specific protocol that we used for our preliminary experiments. Section 4 contains the results of those experiments. Section 5 outlines related work, and Section 6 concludes with a (non-exhaustive) list of directions for further research in this area.

2 Privacy Principles for Surveillance

This section outlines several principles that we believe should govern electronic surveillance. We start with a basic principle stating that processes that use private data in bulk must be *open*, and we then outline several related properties that we expect such open processes to have. Finally, we summarize how these principles might be applied in the case of "set-intersection warrants."

2.1 Open Processes for Law Enforcement

A basic tenet of democratic society is that law enforcement must follow *open processes*: procedures laid out in public law and subject to debate and revision through deliberation. Police need not disclose *whom* they may suspect of a particular crime or other details of an ongoing investigation, but their investigation must nevertheless follow rules and procedures established in *open* law books that everyone has a right to know and understand. And it is accepted that searching a person's home or personal records requires a narrowly targeted and properly authorized warrant based on probable cause.

We wish to formulate an openness principle for electronic surveillance that distinguishes between two classes of Internet users. A *targeted user* is one who is under suspicion and is the subject of a properly authorized warrant. All others are *untargeted users* – the vast majority of Internet users (and cell-phone users and users of any general-purpose, mass-communication system).

Just as search-warrant processes in free societies are grounded in open law, we believe that any "bulk" electronic-surveillance process that ingests, searches, or otherwise touches private[1] data of untargeted users must likewise be an open process. We refer to processes that are not open, public, and unclassified as *secret processes*, and we seek to limit their use (while admitting that there are circumstances in which they may be needed). Once law enforcement has legitimately employed an open process to identify, target, and obtain information about an Internet user suspected of a crime, however, it may potentially subject that targeted user's data to the full range of secret analysis tools and techniques in its arsenal.

One of the key reasons the NSA's mass-surveillance activities disclosed by Snowden are so troubling is that they tap into "bulk" data and metadata about untargeted users and ingest these private bulk data into secret processes that are codified only in secret FISA law and are subject only to secret oversight and accountability procedures (Figure 1a). In short, the public must simply "trust" the US government's evidence-free assertions that its mass ingestion and secret processing of privacy-sensitive data are (secretly) lawful and subject to adequate (secret) privacy protections and effective (secret) oversight. We cannot remotely envision the framers of the US Constitution being comfortable with such blind faith in secret mass-surveillance processes of this nature.

We therefore propose that a basic openness principle, comprising two main planks, should govern electronic-surveillance processes in a modern democracy:

I Any surveillance or law-enforcement process that obtains or uses private information about untargeted users shall be an open, public, unclassified process.

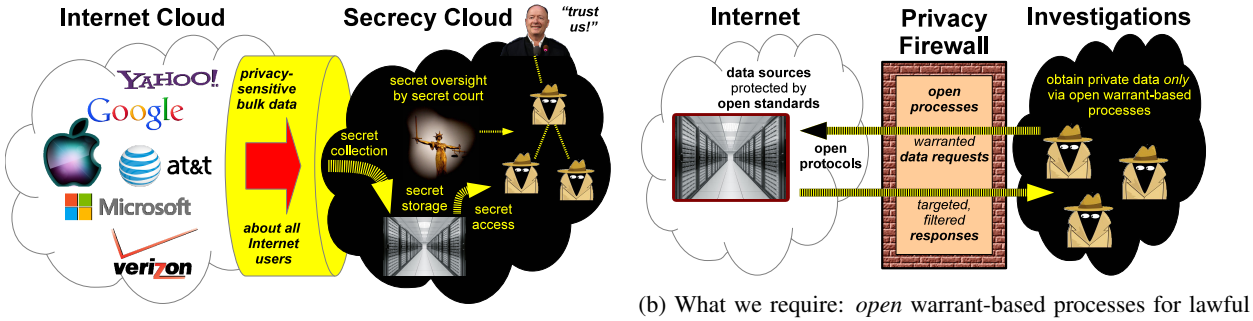
II Any secret surveillance or law-enforcement processes shall use only:
   (a) public information, and
   (b) private information about targeted users obtained under authorized warrants via open surveillance processes.

We view this openness principle as demanding that an open *privacy firewall* be placed in the path of private information flowing from the Internet to law enforcement (Figure 1b). Processes that search or ingest private data of untargeted users "through the firewall" must be open processes, but, once a user is targeted by a legitimate warrant and his data have been acquired via open processes, these targeted user data may potentially be subject to secret investigative processes.

Openness conceived in this manner may sound incompatible with the requirement that government agencies be able to keep secret the targets and details of active investigations, but it is not. Using appropriate security technology, a data-collection or surveillance *process* used in an investigation may be made fully public without revealing the *content* of any particular investigation.

Our focus here is on general electronic surveillance principles for law enforcement purposes, independent of any particular government or agency. The hot-button case of the NSA is complicated by the fact that the NSA was founded as a foreign-intelligence agency but has acquired *de facto* characteristics of law-enforcement agencies by: (a) increasingly serving to support and feed surveillance data to law-enforcement agencies such as the FBI and the DEA; (b) collecting and storing both

[1] Rigorous definitions of the term "private" are the subject of extensive study in computer security, law, philosophy, and many other fields; as such, they are beyond the scope of this paper.



(a) What we have: a cloud of secret mass surveillance processes

(b) What we require: *open* warrant-based processes for lawful electronic surveillance, creating a "privacy firewall."

Figure 1: Secret versus open electronic surveillance processes

US and non-US surveillance data alike, even if internal "searches" are allowed only on "non-US persons"; (c) being increasingly employed not just against wartime adversaries but against citizens of peaceful, allied, democratic states, who common sense dictates should have protection against "unreasonable search and seizure" regardless of the letter of US law [11]. To whatever extent the NSA or any government agency behaves like a domestic or international law-enforcement agency, we believe the above openness principle should apply.

2.2 Mass Surveillance

How should this openness principle be applied to *mass-surveillance* processes, *i.e.*, processes such as the cell-phone records-collection program that have the potential to collect or use *all* data in a particular category about *all* users? (As currently implemented, the cell-phone records-collection program realizes this potential [6, 10], but we do not think it should.) We refer to data sets collected and used in mass surveillance as *bulk data sets*.

We identify four particular "sub-principles" that we believe should apply to mass-surveillance processes:

Division of trust: No single agency or branch of government should have either the authority *or the technical means* to compromise the privacy of bulk data about untargeted users. Mass-surveillance processes must require the sign-off, oversight, and active participation of multiple independent authorities representing each branch of government.

Enforced scope limiting: Surveillance processes must incorporate scope-limiting mechanisms ensuring that no particular warranted-surveillance activity captures data from an overly broad group of users. For example, each warrant might have a specified limit on the number of users whose data may be touched by the warrant-authorized process.

Sealing time and notification: Surveillance processes that capture privacy-sensitive user data must impose a limit on the length of time that the users in question may be kept ignorant of the fact that their data were captured. After this time has expired, the process must ensure that the users are notified of the data access and given means to investigate the justification and/or obtain recompense for any unjust effects of the investigation. Higher levels of authority should be required to authorize longer sealing times. No level of authority should permit indefinite sealing times (even indirectly, on an "installment" basis).

Accountability: Surveillance processes must incorporate accounting mechanisms that enable all three branches of government, as well as civilian participants, to maintain and safely disclose relevant statistics on how frequently and extensively warranted-access mechanisms are used, *e.g.*, number of warrants per month of a given type, maximum number of individuals affected under any warrant, total number of individuals affected by all warrants in one month, or maximum secrecy period applied to any outstanding warrant in one month.

2.3 Case Study: Intersection Warrants Using Cell-Tower Dumps

Given a properly authorized warrant, we wish to enable law-enforcement agencies to target not just *known* users (those whose cell-phone numbers they already have and are covered by the warrant) but also *unknown* users (in our case, those whose cell-phone numbers they do not have but may be able to discover by *intersecting* several relevant cell-tower dumps). It may appear nonsensical to describe a user as both "unknown" and lawfully "targeted," but it is not. We may view such an *intersection warrant* as a type of "John Doe" warrant [3]: one in which the names or phone numbers of the person(s) of interest are unknown, but for which relevant times and locations are known, and for which there is sufficient evidence to convince a judge that there is probable cause to believe the given times and locations uniquely identify the unknown person(s) who committed a crime.

For example, the FBI caught the High Country Ban-

dits [2] by intersecting three cell-tower dumps, representing the sets of cell-phone numbers that had been used near three different bank-robbery sites at the times of the robberies. In total, these dumps contained 150,000 cell-phone numbers, but their intersection contained only one: that of a High Country Bandit. Similarly, the NSA's CO-TRAVELER program [17] searches for unknown associates of known surveillance targets by first intersecting cell-tower dumps from times and locations at which a particular known target appeared and then interpreting the intersection as the set of cell-phone numbers of people who may be "traveling with" the known target.

In Sections 3 and 4, we present and evaluate a protocol that computes the intersection of cell-tower dumps and obeys the principles articulated above. This is a natural test case for us for at least two reasons. First, intersections of cell-tower dumps have proven useful in catching criminals; this distinguishes them from many of the other surveillance activities featured in the Snowden revelations, the practical utility of which is at best unclear. Second, privacy-preserving set intersection is a well-studied, mature, and practical technology [9,13,18].

Note that our protocol is not specific to cell-tower dumps and could also be used to query other "time-and-place" metadata collections in an open, lawful manner.

3 Lawful Intersection Attacks

This section first outlines the assumptions and principals involved in our lawful intersection-warrant protocol, then describes the operation of the protocol, and finally summarizes its key security properties.

3.1 Principals

Our model for lawful intersection attack involves the following three types of principals. For simplicity, we assume here that these principals will participate in an intersection-attack mechanism in an honest-but-curious way. That is, they will not attempt to violate the rules of the mechanism, but they may use their own views of all data they see to acquire additional information.[2]

Sources: entities that produce metadata records embodying information of the form, "user X was observed to be near location Y at time Z." The obvious examples are phone companies whose cell towers produce logs of the users who appeared in the vicinity of a given cell tower at a given time, but our model extends to other producers of metadata of this general form.

[2] This assumption could be relaxed significantly by requiring all principals to produce zero-knowledge correctness proofs of their intermediate results, using standard and well-known techniques. We leave these details to future work, however, and we would still need to assume the correctness of the original inputs – *e.g.*, logged phone numbers.

Repository: any entity tasked with storing metadata for surveillance or law-enforcement purposes. This may be the phone companies that produced the records (*i.e.*, the same as the metadata sources), a government agency, or some specialized independent agency. While "who stores the data" is an important question in general, it is orthogonal to our goals, and our model is agnostic with respect to its answer.

Agencies: a set of *multiple* independent but cooperating government agencies across whom our model divides surveillance authority. While our model is formally agnostic with respect to the number or specific natures of the authorities across whom trust is divided, we will use the US's 3-branch constitutional model as a concrete example, in which it might be appropriate to divide surveillance authority across three agencies:

- The **Executive Agency** represents the executive branch and is responsible for *requesting* surveillance warrants – *e.g.*, an agency like the NSA or FBI.
- The **Judicial Agency** represents the judicial branch and is responsible for *authorizing* requested warrants, after verifying independently that they are legally justified and suitably scoped.
- The **Legislative Agency** reports to the legislative branch and is responsible for ensuring that accurate and sufficiently detailed data are gathered and regularly reported to Congress on how and to what extent these surveillance capabilities are employed.

3.2 Lawful Set-Intersection Protocol

The lawful set intersection protocol we present is similar in structure to the protocol of Vaidya and Clifton [18].

Our protocol is built on two commutative encryption schemes: ElGamal and Pohlig-Hellman. A commutative encryption scheme has the property that a message encrypted sequentially under multiple encryption keys can be decrypted by applying the corresponding decryption keys *in any order*. The ElGamal and Pohlig-Hellman encryption schemes are not only commutative but mutually commutative - that is, a message encrypted under a combination of encryption keys from the two cryptosystems can still be decrypted by the corresponding decryption keys, again regardless of order.

We use the randomized, public-key ElGamal encryption scheme for long-term encryption of stored data. Each agency, or "participant," in the protocol needs an ElGamal key pair, the public key for which is known to the sources of private information. The Pohlig-Hellman encryption scheme is symmetric-key and deterministic, and the participants in the protocol use it to blind the data prior to intersection. Because Pohlig-Hellman is deterministic and commutative, there is a one-to-one correspondence between data items and their encryptions un-

der any fixed set of Pohlig-Hellman keys, regardless of the order in which those keys are applied. Short-term Pohlig-Hellman keys are generated by each participant during the protocol execution and discarded at the execution's end.

Each participant's input is its ElGamal private key and a set of data that has been encrypted under the ElGamal public keys of all agencies. The agencies do not generate these sets – rather, they are distributed to the agencies by the repositories. If there are k agencies, they are given numbers 1 through k so that, when the j^{th} agency is done acting on a set of data, it can pass the set on to the $(j+1)^{st}$, and it can receive new sets from the $(j-1)^{st}$.

Assuming all agencies execute the protocol with honest-but-curious behavior, the protocol's output for each participant will be the intersection of all sets. Optionally, each agency may also supply a threshold limiting the size of the intersection it is willing to reveal. If the size of the intersection of all sets would be above any agency's threshold, no agency will learn anything except the cardinality of the intersection, which agency's threshold was violated, and some intermediate values discussed in Section 3.3. The protocol runs as follows:

Initialize. Each agency first generates a temporary Pohlig-Hellman key to be used only during this execution of the protocol and then discarded. The first agency then obtains the ElGamal-encrypted sets to be intersected from the Repository.

Phase 1. Each agency in turn uses its ElGamal private key to remove a layer of ElGamal encryption from each item in each set to be intersected; it then adds a layer of Pohlig-Hellman encryption to each item, using the temporary key it generated in the Initialize step. The agency then randomly shuffles each encrypted set independently, while keeping the sets separate, and forwards the sets to the next agency. The phase is complete when agency k decrypts the final layer of ElGamal encryption from all items in all sets, leaving all sets encrypted under every agency's Pohlig-Hellman keys only.

Phase 2. Agency k broadcasts the resulting Pohlig-Hellman-encrypted data sets to all other participants. Each participant then computes the desired intersection: *i.e.*, the encrypted elements that appear in all sets – or, more generally, the elements that appear in some threshold number of the sets as defined by the intersection warrant. Because Pohlig-Hellman is a commutative and deterministic encryption scheme, two identical data items will have identical encryptions at this stage, making computation of the intersection trivial despite the encryption.

Phase 3. If any agency sees that the number of distinct items (*e.g.*, phone numbers) appearing in the resulting set intersection is above a warrant-specified limit on the number of individuals the warrant is permitted to target, the agency deletes its Pohlig-Hellman key, sends a message to all other agencies, and refuses to continue with the protocol. (The agency requesting the warrant might then be required to produce a new, more narrowly targeted warrant and try again.)

Phase 4. If the intersection's cardinality meets the requirements of the warrant, then the agencies collectively decrypt the items in the intersection. As in Phase 1, each agency in turn uses its Pohlig-Hellman key to decrypt each element of the intersection set, shuffles the intersection set, then forwards it to the next agency. The phase completes when when agency k decrypts the last layer of Pohlig-Hellman encryption and forwards the plaintext result to the other agencies.

For simplicity, we describe Phases 1 and 4 above as strict "cascades," each agency processing the full data set before passing it to the next. A simple performance optimization our prototype implements is for different input sets to start at different agencies – i.e., to start and end at different "points" around a circle – thus spreading computational load and increasing parallelism. This is only one of many potential optimizations, however.

3.3 Protocol Properties

We now analyze our cell-tower-dump intersection mechanism with respect to our openness principle for mass-surveillance processes. We accomplish division of trust by having all data be encrypted in advance with the public keys of the agencies that request, authorize, or oversee the surveillance. Without participation of all of these agencies, the data cannot be decrypted – even if the Repository is compromised, for example – and no unauthorized surveillance can be performed unilaterally.

The protocol also provides the means to enforce a limited scope of investigation. The sizes of all sets and intersections are visible to all participants in Phase 3; so the warrant can specify a limit on the number of users whose data may be revealed. Any participant can stop the protocol, *before* any metadata records are revealed in cleartext, if the size of the intersection is above this limit.

If the protocol completes, it gives the same output to each participant. This makes it easy to notify users whose data were viewed after some sealing time and to maintain statistics for the purpose of accountability. These processes are beyond the scope of the intersection protocol, but one of the participants in the protocol can be responsible for maintaining them.

Finally, this process protects the privacy of untargeted users. The only information leaked apart from the output are the *sizes* of intersections of any two or more sets involved in the protocol. (This property is proven in [18] for a protocol using the same structure as ours; the proof generalizes straightforwardly to our case.)

This small information leakage reveals how many

users appear in multiple sets but does not reveal any specific user identities or metadata other than those in the requested intersection. Because only aggregate properties of the sets are leaked, we feel this leak should not represent a major privacy issue – and when a query fails because of an empty or too-large intersection, the leaked statistics may help the agency that requested the warrant formulate a revised request for a better scoped warrant.

4 Implementation and Evaluation

This section presents the results of our preliminary experiments with the lawful set-intersection protocol of Section 3. Recall that each network node that participates in this protocol acts on behalf of an "agency," in the terminology of Section 3.1. It is this set of agencies across whom trust, *e.g.*, the power to authorize an intersection attack, is divided in our model. Because the public keys of *all* agencies are used to encrypt the data stored in the Repository, each agency effectively uses its private key to "authorize" the selection and decryption of results responsive to a particular intersection warrant.

4.1 Prototype Implementation

Our implementation of the lawful set-intersection protocol is written in Java and available on GitHub.[3] The prototype does not use any external libraries for cryptography beyond Java's standard BigInteger class.

The program is run on multiple servers. The servers connect to each other over TCP sockets, and, for simplicity, we use a directed cycle as the connection graph. Each server sends data only to the next server in the cycle and receives data only from the previous server in the cycle. Each participant takes one set of encrypted data and a 1024-bit ElGamal private key as input. The data must have been sequentially encrypted under all participants' public keys before it can be used in the protocol; because this encryption is done offline and in advance, the protocol itself does not require access to public keys.

To test the protocol, we ran it many times on data sets of various sizes. We used PlanetLab [5] on a network of three computers located across the United States in order to take into account the potential effects of latency on end-to-end running time. The computers we used are located at Yale in Connecticut, University of Texas in Dallas, and University of California in Riverside.

The tests were all run using only these three nodes, each node with one data set. We expect the running time would increase considerably as a function of the number of participants, but three is a natural number of nodes in this context, representing a distribution of authority across three branches of government (Section 3.1).

[3] https://github.com/DeDiS/Surveillance

Items	Data sent per node (KB)	CPU time per node (s)	End-to-End runtime (s)
10	21	0.6	4.1
25	46	1.3	6.0
50	86	2.6	9.6
75	127	3.8	12.6
100	167	5.0	15.5
250	410	12.4	38.2
500	815	24.7	69.1
750	1220	36.9	103.0
1000	1625	49.3	137.2
2500	4055	123.0	369.9
5000	8106	245.6	724.9
7500	12156	369.4	1034.9
10000	16206	493.8	1402.3
50000	81009	2560.5	6971.2

Table 1: Experimental Results

4.2 Query Efficiency

Prior to execution, we randomly generated data sets for each trial for each node. We ran the protocol 10 times each with different-sized data sets, ranging from 10 items per set to 50,000 items per set. We measured three variables: the bandwidth, or amount of data each node transmitted during the protocol; the CPU time each node used in performing calculations; and the total end-to-end time, from the start of the protocol's execution to the production of output. After running each test 10 times, we averaged the results; these averages are presented in Table 1.

In the High Country Bandits case, the FBI processed information from about 150,000 users total [2]. Our largest test, with 50,000 data items per set, tested our protocol's efficiency with an equally large amount of data. The average amount of time needed to run the protocol in this experiment was 6971.2 seconds, just under two hours. Considering the amount of time it would take a law-enforcement agency to set up its own set-intersection program, two hours seems quite reasonable. Further, because all the key computations in this protocol are "embarrassingly parallel," delay could probably be reduced by orders of magnitude with a moderate and readily feasible investment in processing power at each agency.

These tests were run with an intersection size of three. We also tested these benchmarks with an intersection size of 10 and found that the average times did not change by more than one second in any case and that the data sent per node always increased by 3 KB.

Our results indicate that the amount of data sent over the network, CPU time, and end-to-end time all increase linearly with the size of the data sets, which is what we

would expect from this protocol.

Further tests showed that total data sent and total CPU usage across participants were not affected if the data were concentrated in one or two sets, as opposed to being spread equally over all three sets. However, we found that the end-to-end delay can increase by up to a factor of two if the data are spread out. This result is unsurprising, because unbalanced sets render less effective the optimization mentioned at the end of Section 3.2, wasting time while the small-set-input participants idle, waiting for data to be sent by large-set-input participants.

5 Related Work

Private set intersection [9, 13, 18] is but a sliver of a large body of work on privacy-preserving algorithms [1]. We are not the first to propose employing such algorithms for targeted lawful surveillance. Kamara recently explored the use of Private Information Retrieval (PIR) for metadata queries [12]. Kroll, Felten, and Boneh explore mechanisms to distribute trust and improve privacy and accountability in queries [14]. These methods focus on queries for *known* targets, whereas we wish to demonstrate that proper use of cryptography can support powerful privacy-preserving surveillance of *unknown* targets. Non-cryptographic techniques have also been explored to protect privacy in video surveillance [7].

6 Conclusions and Open Problems

From the experimental results in Section 4, we conclude that privacy-preserving, accountable set intersection may indeed be achievable at scale. This in turn leads us to be optimistic about the feasibility of the broader goal articulated in Section 1: maintaining constitutional rights in powerful, evolving, digital-communication systems while simultaneously equipping law-enforcement and intelligence agencies to use these systems to combat and prevent crime and terrorism. There is a great deal of further work to be done along these lines, and we briefly describe a portion of it here.

The principles given in Section 2 are a first stab at an appropriate foundation for privacy-preserving, accountable surveillance. We hope that they will stimulate discussion and be refined and revised by the relevant research communities.

6.1 Enhancements and Generalizations

The protocol that we have implemented leaks the sizes of pairwise intersections (but not the contents of those intersections) in the case of three participants; more generally, it leaks the sizes of the j-wise intersections, where $1 < j < k$, and k is the number of participants. As explained in Sections 3 and 4, this is not a show stopper on privacy or efficiency grounds, but it leaves open the question of whether there is a similarly efficient protocol with the same accountability properties that reveals no information except the k-wise intersection.

In principle, one could achieve this ideal level of privacy by starting with a general secure, multiparty computation (SMPC) protocol and augmenting it with the appropriate accountability features. How well such an approach would scale is an open question.

Starting with the Fairplay platform for secure, two-party computation [15], there has been much work on general-purpose SMPC platforms. The goal of this research is to provide languages, compilers, run-time environments, and other platform elements that enable programmers who are not experts in cryptography or SMPC to write ordinary code and transform it into executable, multiparty protocols with the desired security properties. There are now many such platforms whose performance and usability are improving (see, *e.g.,* [19]). One could, in principle, achieve the goals put forth in Section 3 simply by writing a set-intersection program and using, say, Sharemind [4] to translate it into a privacy-preserving, distributed set-intersection protocol (rather than implementing privacy-preserving set intersection "from scratch" as in Section 4). Open questions include whether the resulting protocol would be efficient enough to use at scale and how to make it accountable.

Of course, intersection of cell-tower dumps is but one of many computations that could be of use to law-enforcement and intelligence agencies. It would be interesting to identify other such computations and to apply to them the principles and computational approaches that we have explored in this paper.

6.2 Openness in Lawful Surveillance

One remaining high-level issue is the tension between the openness principle proposed in Section 2.1 – requiring that processes handling "bulk" electronic surveillance data be open – and the traditional desire of intelligence agencies to protect "sources *and methods*," especially from the knowledge of criminals or terrorists being investigated. We emphasize that satisfying our openness principle by no means demands exposing *all* intelligence methods – only those few involved in implementing the "privacy firewall" in Figure 1b. All the details of any particular investigation – who is being investigated, when, the details of a particular warrant such as which metadata sets are to be intersected, how those sets were chosen, and how the decrypted results are processed *after* being lawfully queried through the "privacy firewall" – could still rely on closely guarded intelligence methods.

A generic open process such as the set-intersection primitive tends to be usable in many different, specific ways – as in the contrasting High Country Bandits [2] and NSA CO-TRAVELER [17] examples. Had CO-TRAVELER not been disclosed by Snowden, for example, then this specific method of using set intersection to find unknown associates of known targets as they travel might well remain a closely guarded secret, even if the basic intersection-warrant mechanism were well-known, openly debated, and instituted in public policy.

Finally, a basic tenet of democratic society and the rule of law is that it is better to risk a few criminals' going uncaught, because they know and understand public law-enforcement processes "too well," than to risk that secret law-enforcement processes, however well intentioned at the outset, might become unaccountable and evolve into "star-chamber" tools of political repression and authoritarianism. This democratic principle of openness must be carried into the electronic world; with the right tools, the principle need not tie the hands of legitimate, accountable law-enforcement processes.

7 Acknowledgements

We thank the FOCI reviewers for their valuable and insightful feedback. Our prototype intersection-warrant implementation is derived from private set-intersection code by Ennan Zhai. This work was supported in part by the National Science Foundation under grant 1016875, the Office of Naval Research under grant N00014-12-1-0478, the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory under contract FA8750-13-2-0058, and DARPA and SPAWAR Systems Center Pacific under contract N66001-11-C-4018. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing the office policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

[1] Charu C. Aggarwal and Philip S. Yu. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-Preserving Data Mining: Models and Algorithms*, chapter 2, pages 11–52. Springer, 2008.

[2] Nate Anderson. How "cell tower dumps" caught the High Country Bandits—and why it matters. *arstechnica*, August 29, 2013.

[3] Meredith A. Bieber. Meeting the statute or beating it: Using John Doe indictments based on DNA to meet the statute of limitations. *University of Pennsylvania Law Review*, 150(3):1079–1098, 2002.

[4] Dan Bogdanov and Aivo Kalu. Pushing back the rain – how to create trustworthy services in the cloud. *ISACA Journal*, 3:49–51, 2013.

[5] Brent Chun et al. PlanetLab: An overlay testbed for broad-coverage services. In *ACM Computer Communications Review*, July 2003.

[6] Ryan Devereaux, Glenn Greenwald, and Laura Poitras. Data Pirates of the Caribbean: The NSA Is Recording Every Cell Phone Call in the Bahamas. *The Intercept*, May 20, 2014.

[7] Frédéric Dufaux and Touradj Ebrahimi. Scrambling for privacy protection in video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1168–1174, 2008.

[8] Joan Feigenbaum and Bryan Ford. Is Data Hoarding Necessary for Lawful Surveillance? *The Huffington Post*, April 19, 2014.

[9] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas. Efficient private matching and set intersection. In *EUROCRYPT (EUROCRYPT)*. Springer, May 2004.

[10] Glenn Greenwald. NSA collecting phone records of millions of Verizon customers daily. *The Guardian*, June 6, 2013.

[11] Human Rights Council. The right to privacy in the digital age: Report of the Office of the United Nations High Commissioner for Human Rights, June 2014.

[12] Seny Kamara. Restructuring the NSA Metadata Program. In *Workshop on Applied Homomorphic Cryptography (WAHC)*, March 2014.

[13] Lea Kissner and Dawn Xiaodong Song. Privacy-preserving set operations. In *Annual International Cryptology Conference (CRYPTO)*. Springer, August 2005.

[14] Joshua A. Kroll, Edward W. Felten, and Dan Boneh. Secure protocols for accountable warrant execution, April 2014.

[15] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella. Fairplay – a secure two-party computation system. In *USENIX Security Symposium*, August 2004.

[16] Daniel J. Solove. *Nothing to Hide: The False Tradeoff Between Privacy and Security*. Yale University Press, 2011.

[17] Ashkan Soltani and Barton Gellman. New documents show how the NSA infers relationships based on mobile location data. *The Washington Post*, December 10, 2013.

[18] Jaideep Vaidya and Chris Clifton. Secure set intersection cardinality with application to association rule mining. *Journal of Computer Security*, 13(4):593–622, 2005.

[19] Yihua Zhang, Aaron Steele, and Marina Blanton. Picco: A general-purpose compiler for private distributed computation. In *ACM Conference on Computer and Communications Security*, November 2013.