Privacy-Preserving Contact Chaining

[Preliminary Report]

Aaron Segal[*]
Computer Science
Yale University
New Haven, CT, USA
aaron.segal@yale.edu

Joan Feigenbaum[†]
Computer Science
Yale University
New Haven, CT, USA
joan.feigenbaum@yale.edu

Bryan Ford[‡]
Swiss Federal Institute of Technology (EPFL)
Lausanne, Switzerland
bryan.ford@epfl.ch

ABSTRACT

How can government agencies acquire actionable, useful information about legitimate targets, while preserving the privacy of innocent parties and holding government agencies accountable? Towards understanding this crucial issue, we present the first privacy-preserving protocol for *contact chaining*, an operation that law-enforcement and intelligence agencies have used effectively. Our experiments suggest that a three-hop, privacy-preserving graph traversal producing 27,000 ciphertexts can be done in under two minutes.

Keywords

Privacy, Secure multiparty computation, Surveillance

1. INTRODUCTION

As networked devices become more available, more capable, and more ubiquitous in everyday life, tension mounts between users' desire to safeguard their personal information and government agencies' desire to use that personal information in their pursuit of criminals and terrorists. Many people assert that we are faced with an unpleasant, stark choice: Citizens can either have control over their personal information, or they can have law-enforcement and intelligence agencies with the tools that they need to keep the country safe. We regard this stark choice as a false dichotomy and assert that, by deploying appropriate technology in the context of sound policy and the rule of law, we can have both privacy and security.

In this paper, we continue the development of accountable, privacy-preserving surveillance that we began in [6]. We require that government surveillance be conducted according to *open processes*, *i.e.*, unclassified procedures laid out in public laws that all citizens have the right to read, to understand, and to challenge through the politicial process; procedures that do not have these properties are referred to as *secret processes*. We distinguish between *targeted users*, who are under suspicion and the targets of properly authorized warrants, and *untargeted users*; the latter are the vast majority of all users in any general-purpose, mass-communication system, and their private information must be protected from government scrutiny. In [6], we applied these principles to the problem of *set intersection*.

We now apply them to *contact chaining*. The goal is to use the topology of a *communication graph* (*e.g.*, a phone-call graph, email graph, or social network) to identify associates (or "contacts") of lawfully targeted users [1]. Agencies can then investigate those associates to determine whether they deserve further attention. It is useful to consider both direct contacts, *i.e.*, users who are neighbors in the communication graph, and extended contacts, *i.e.*, users who are at distance k in the communication graph, for an appropriate constant k. Without accountability and security mechanisms to limit an investigation's scope, contact chaining in a mass-communication network can sweep in a huge number of untargeted users. Section 2 presents an accountable contact-chaining protocol that bounds the scope of the search, uses encryption to protect untargeted users, and is computationally efficient, with time and communication complexity linear in the size of the output.

We posit that the *Openness Principle* put forth in [6] should govern all surveillance activity in a 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 processes.

This principle can be viewed as a requirement that an open "privacy firewall" be placed between government agencies and citizens' private information in a mass-communication network. Processes that move untargeted users' private information through the firewall must be open processes.

We briefly present our contact-chaining results in Section 2 and our open questions and future directions in Section 3. More detail on the Openness Principle, our results on contact chaining and set intersection, related work by others, and future directions can be found in [5, 6].

2. LAWFUL CONTACT CHAINING

The goal of contact chaining is to use information about social connections between identities, *e.g.*, records of phone calls between one number and another, to identify members of a criminal organization or terrorist group. Starting with one or more suspects whose identities are known, the gov-

[*] Supported by a Google Faculty Research Award.

[†] Supported in part by NSF grants CNS-1407454 and CNS-1409599, DHS grant FA8750-16-2-0034, and William and Flora Hewlett Foundation grant 2016-3834.

[‡] Supported by EPFL, the AXA Research Fund, and DHS grant FA8750-16-2-0034.

ernment aims to consider contacts of those suspects. These can be *direct contacts*, such as two people who spoke on the phone, or *extended contacts*, such as two people connected by a chain of two or more phone calls. If Alice calls Bob, and Bob calls Charlie, then Alice and Bob are direct contacts (as are Bob and Charlie). If neither Alice nor Charlie called the other during the period of investigation that defines the graph, then we say that they are extended contacts (or, more precisely, that they are at distance two in the graph).

Without mechanisms to preserve privacy, a contact-chaining search can collect a surprisingly large group of users' information. For example, if the average cell-phone user contacts 30 individuals within the period of the investigation, a contact-chaining search out to distance three would capture 27,000 users on average – or many more if a heavy phone user is swept up by the search. In such a large group, the vast majority of contacts will not be collaborators of the targeted, primary suspect in the investigation. These untargeted users may nevertheless face government scrutiny, intrusive investigation, or a risk that their sensitive communications histories may be leaked accidentally.

Despite this risk, we recognize the potential law-enforcement value of information about the social contacts of targeted invidivuals. Therefore, we propose a *lawful contact-chaining protocol.* Such a protocol permits multiple government agencies working together to provide oversight and accountability [6]. Our protocol focuses on the case in which the government seeks information from multiple telecommunications providers about the communication graph formed by phone calls and text messages. Using this protocol, the agencies can retrieve a set of encrypted records from multiple telecoms, *each of which holds only part of a larger communication graph.* This set of encrypted data contains the identities of users within a certain distance of a target, but the identities cannot be decrypted unless the agencies cooperate. Under the lawful processes we propose, this cooperation would take the form of intersecton with other sets of encrypted data. These sets can come from privacy-preserving contact chaining, from cell-tower dumps, or from other sources of information about suspects. While any set may contain encrypted data about many untargeted users, few users will appear in *all* the sets, and those few will be suitable targets for further lawful investigation.

The same principles of oversight and accountability provided by multiple government agencies can apply to contact-chaining searches in other types of communication graphs, such as the social-network graph of Twitter or Facebook. These cases do not require our protocol, however: if one provider knows the entire communication graph, it can compute the output of the protocol without any interaction.

2.1 Privacy-Preserving Contact Chaining

2.1.1 Inputs and Parties to the Protocol

There are two types of parties in this protocol: Telecommunications companies (telecoms) and government agencies interested in performing lawful contact-chaining (agencies). The protocol computes a function of all parties' data.

The telecoms jointly hold an undirected communication graph $G = (V, E)$. Each telecom knows only a subset of the edges E. V contains vertices labeled with the phone numbers they represent, and E contains an edge between a and b if and only if phone number a has contacted phone number b or *vice versa* within the period of the investigation. Each phone number v is served by exactly one telecom. We assume telecoms know which telecom serves which phone number. Each telecom keeps records of all phone calls made by phones they serve, including calls made to phone numbers served by other telecoms. The subgraph known by telecom T is $G_T = (V, E_T)$ where E_T is the set of edges (a, b) such that a or b is a phone number served by T. Henceforth, for any phone number a, let $T(a)$ be the telecom that serves a.

The agencies must each hold a copy of a *warrant* in order to perform this protocol. A warrant is a triplet $(x,\ k,\ d)$. x is a target phone number. Because x belongs to a user targeted by the agencies, we assume that the agencies also know which telecom serves x. k is the (small) distance from x to which the agencies wish to "chain out." Choosing a small distance is important to limiting the scope of the investigation. However, many users' information might still be captured if some phone numbers have very many contacts. Suppose the target x calls the most popular pizza place in town. Now everyone else who has recently called that pizza place is at distance at most two from x.

Business phone numbers often have many more contacts than personal phone numbers do, and knowing that two individuals have contacted the same business does not usually indicate a relationship between those individuals. Therefore, the warrant also includes d, an upper bound on the degree of users that the agencies will "chain through." If a phone number has more than d contacts, then the agencies do not follow paths to other users through that phone number in their search (but do include that number in the output). The agencies disregard d for the initial target x, however, because they have already determined that contacts of x are of potential interest, even if x is a business.

This provides a reasonable limit to the scope of the investigation and hides what are very likely to be untargeted users from the government. In the uncommon scenario in which a business with many contacts also functions as a front for a criminal organization, the government could conduct further investigation, perhaps beginning a new contact-chaining search with that number as the initial target.

2.1.2 Security Assumptions

We assume some existing cryptographic infrastructure. All telecoms and agencies must have public encryption keys known to all other parties to the protocol and private decryption keys. For the purpose of interoperability with lawful intersection [6], agencies' keys must be for a commutative cryptosystem (*e.g.*, ElGamal). Each party must also have a private signing key and a public verification key.

In the protocol below, we refer to "the agencies' sending" messages to one or more telecoms. Exactly which agency transmits messages to the telecoms is not important to our protocol, but a telecom will disregard any message not accompanied by signatures from all agencies. One simple topology is for a single agency to act as a relay, forwarding reponses from the telecoms to the other agencies and signatures on agency messages to the telecoms.

Our protocol preserves the privacy of untargeted users as long as all parties execute the protocol in an honest-but-curious manner, at least one of the government agencies does not collude with the others, and no telecom colludes with government agencies. A colluding group containing all agencies would be equivalent to the current situation, in which the government does not provide meaningful accountability of its own surveillance activities; what we propose is a replacement for this situation, but it does require the government to follow its own laws, once set. A telecom's colluding with a government agency would amount to sending that agency free information about its users or submitting incorrect information to the protocol. But telecoms have no business purpose to deviate from the protocol and risk legal action. In practice, existing legal tools allow law enforcement agencies to gather information about the phone history of a suspect with a valid warrant, but such information cannot generally be used for further contact chaining.

2.1.3 Desired Outputs and Privacy Properties

The goal of the protocol is for the agencies to obtain a set of ciphertexts, each of which is the encryption of a phone number v such that the distance in the communication graph from v to the targeted phone number x is at most k. The set should not contain encryptions of numbers v such that each path from x to v of length at most k contains an intermediate vertex of degree greater than d; the "intermediate" vertices in a path are all vertices except the endpoints x and v.

Every phone number in this set must be encrypted with each of the agencies' public ElGamal keys. The agencies should all have the same output. The telecoms should not learn the agency's output. Instead, each telecom's output should contain only a list of which of the phone numbers it serves were sent to the government agencies. This allows the telecoms to play an additional accountability role.

The agencies can act as appropriate to further investigate these encrypted phone numbers. The set of encrypted phone numbers can be intersected with, say, the encrypted numbers of people on a terrorist watch list [6].

In the basic protocol presented here, the agencies and telecoms learn some additional information. Specifically, the agencies learn which provider serves each encrypted phone number in the output set and the distance from x of each encrypted phone number. Each telecom learns which of the phone numbers it serves appear in the agencies' output, as well as the distance of each of those phone numbers from the target phone number x. In [5, Section 4.1.5], we present a modified version of the protocol that uses a DC-net-based anonymity subprotocol to prevent the agencies from learning which telecom serves which encrypted phone numbers.

As long as our security assumptions hold, the agencies collectively learn *no* information about the edge set E except what is implied by the output. Furthermore, the agencies cannot learn any of the phone numbers that appear in encrypted form in the output (unless implied by the size of the encrypted output and the leaked service information), nor can agencies cause a phone number not within distance k of x to appear in the output, even in encrypted form.

2.1.4 Lawful Contact-Chaining Protocol

The protocol below amounts to a distributed breadth-first search of the communication graph run by the agencies making queries of the telecoms. However, all messages the agencies receive from the telecoms will be encrypted. Let $\text{Enc}_T(m)$ be the encryption of message m under telecom T's public key. Call such an encryption a *telecom ciphertext*. Let $\text{Enc}_\mathcal{A}(m)$ be the encryption of m under the public keys of all agencies, and call such an encryption an *agency ciphertext*.

To manage the breadth-first search, the agencies (or at least the investigating agency) will maintain a queue $\mathbf{Q}$, containing vertices yet to explore. $\mathbf{Q}$ contains tuples for unexplored vertices a of the form $(\text{Enc}_{T(a)}(a), T(a), j)$. These tuples contain the telecom ciphertext for a, a record of which telecom owns a, and an integer j indicating the remaining distance from a still to be covered by the search.

The agencies represent their output as a list $\mathbf{C}$ of agency ciphertexts. Each telecom T represents its output as a list $\mathbf{L}_T$ of plaintext users served by that telecom whose information the agencies requested. The protocol is as follows:

1. The agencies start by agreeing upon a warrant (x, k, d). They encrypt x under the public key of $T(x)$.
2. The agencies initialize $\mathbf{Q}$ to contain $(\text{Enc}_{T(x)}(x), T(x), k)$.
3. The agencies initialize the output list $\mathbf{C}$ to be empty.
4. Each telecom T initializes its output list $\mathbf{L}_T$ to be empty.
5. While $\mathbf{Q}$ is not empty, do the following:
   - (a) The agencies dequeue $(\text{Enc}_{T(a)}(a), T(a), j)$ from $\mathbf{Q}$. They send the pair $(\text{Enc}_{T(a)}(a), j)$ to $T(a)$.
   - (b) a's provider, $T(a)$, decrypts a from its telecom ciphertext. It adds a to $\mathbf{L}_T$.
   - (c) $T(a)$ encrypts a under the agencies' public keys, and sends $\text{Enc}_\mathcal{A}(a)$ to the agencies.
   - (d) If $j = 0$, $T(a)$ is done. Go to step 5g.
   - (e) Otherwise, $T(a)$ encrypts each neighbor b of a under $T(b)$'s public key, creating a telecom ciphertext for b.
   - (f) $T(a)$ sends the number of ciphertexts generated this way, $\deg(a)$, as well as all telecom ciphertexts generated in the previous step, to the agencies. $T(a)$ sends the ciphertexts in the form of pairs $(\text{Enc}_{T(b)}(b), T(b))$.
   - (g) The agencies add $\text{Enc}_\mathcal{A}(a)$ to $\mathbf{C}$.
   - (h) If $\deg(a) > d$ and $j \neq k$ (i.e. $a \neq x$), the agencies discard all telecom ciphertexts received for a's neighbors (they refuse to sign these ciphertexts in future protocol steps, and do not send them to telecoms).
   - (i) Otherwise, for each telecom ciphertext received, the agencies add $(\text{Enc}_{T(b)}(b), T(b), j - 1)$ to $\mathbf{Q}$.
6. The agencies' final output is the list $\mathbf{C}$. Each telecom T's final output is $\mathbf{L}_T$.

The inner loop can be executed many times in parallel, up to the point of completely emptying $\mathbf{Q}$ at the beginning of the loop. Many messages to the same telecom can also be batched and sent together, thereby reducing the number of signing and verifying operations so that they depend only on k and not on the size of the input or output.

Correctness and privacy of the basic protocol are argued in the longer technical report [5, Section 4.2].

2.2 Contact-Chaining Protocol Performance

We implemented the basic protocol in Java and tested its running time, CPU time, and network utilization. Our implementation uses the variant of our protocol in which the agencies completely exhaust the search queue $\mathbf{Q}$ each round, sending all queries at any given distance from x to the telecoms at once in batches, for greater parallelism. All telecoms receive their batch of queries at the same time and operate on those queries using eight parallel threads.

We use 2048-bit DSA signatures, 2048-bit RSA encryption for the telecoms, and ElGamal encryption for the agencies' output. Our Java program supports any number of agencies and telecoms, but we chose to run tests with three government agencies and four telecoms.

For our contact graph, we used an anonymized data set of 1.6 million users from Pokec, a Slovakian social network [3].

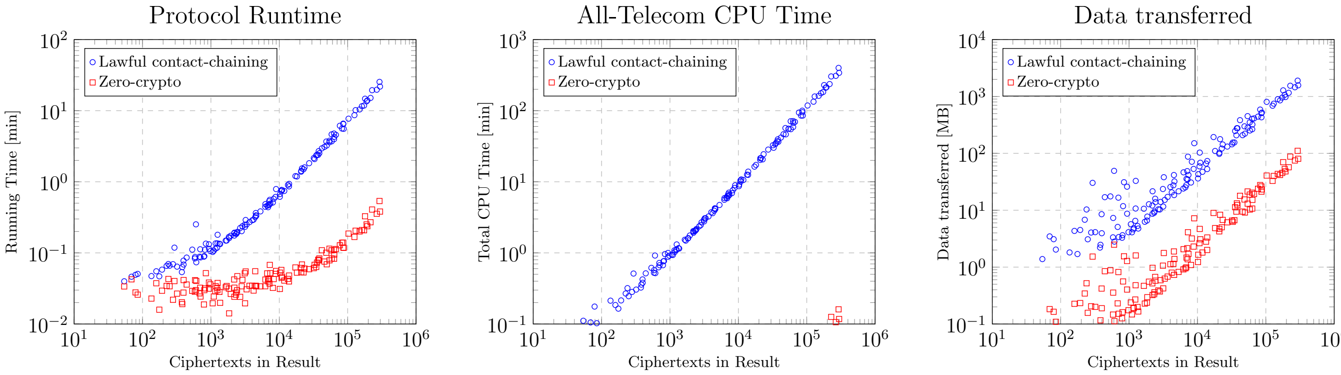

Figure 1: Performance Evaluation of the Lawful Contact-Chaining Protocol

To simulate multiple providers, we assigned each user to one of four telecom servers. Each telecom server was given a different number of the users, in proportion to the subscriber bases of the world's largest four telecoms [4].

These results help in evaluating how practical our lawful contact-chaining protocol would be in practice. However, our data set is small compared to the databases held by real telecommunications companies, each company handles that data using different technologies, and absolute costs might vary. Therefore, we also produced a implementation of the contact-chaining protocol that omits all cryptographic operations, and does not preserve the privacy of users. By comparing the performance of our lawful contact-chaining protocol with the zero-cryptography protocol, we can get a sense of the "cost of privacy and accountability."

Additional detail about our Java implementation and experimental setup can be found in [5, Section 4.3].

2.2.1 Results

Our implementation of lawful contact-chaining performed well, showing a linear relationship between the number of ciphertexts in the output and the running time, CPU time, and data usage of the protocol. We display graphs of our recorded data in Figure 1.

We found that our protocol was able to process about 197 ciphertexts per second on average. Returning to our example of a network with an average of 30 contacts per user, a search with $k = 2$ would have 900 users in the output, and a search with $k = 3$ would have 27,000 users in the output. In our experiments, we found that a search that returned 937 ciphertexts took 6.86 seconds to run, and a search that returned 27,338 ciphertexts took 109.55 seconds to run.

The zero-cryptography version of our program ran, predictably, more quickly than the lawful privacy-preserving version. The total CPU time across all telecoms needed for our zero-crypto implementation never rose above ten seconds, even in the largest cases. We conclude that, even given the scale of database operations that real telecoms perform, the cost of adding privacy-preservation to the contact-chaining protocol is reasonable.

More detailed analysis is presented in [5, Section 4.3].

3. OPEN PROBLEMS AND FUTURE WORK

Thus far, we have explored accountable, privacy-preserving protocols for set intersection and contact chaining. Another operation of potential interest is the retrieval of targeted users' postings on Facebook and other social networks, including those that are shared only with a small subset of the target's "friends." Accountable surveillance of social-network postings may present novel protocol-design challenges, because it deals with one-to-many communication, while previous work dealt with pairwise communication.

It may be possible to speed up our contact-chaining protocols by using elliptic-curve cryptography instead of RSA. Our assumption that all parties are honest-but-curious might be weakened, *e.g.*, by using zero-knowledge techniques to obtain versions of our protocols that are secure against a rogue agent's maliciously modifying telecom-supplied data in order to falsely incriminate a victim. One may wish to generalize the differential-privacy approach of Kearns *et al.* [2] to handle to indirect contacts as well as direct contacts.

Finally, the Openness Principle of [6] is but one step toward a full understanding of how democratic processes and the rule of law can be carried into the digital world. Further investigation, much of it interdisciplinary, is needed.

4. REFERENCES

[1] Tim Cushing. NSA Appears To Be Chaining Calls Using Phone Numbers One Hop Out As New Originating Selectors. *Techdirt*, July 3, 2014.

[2] Michael Kearns, Aaron Roth, Zhiwei Steven Wu, and Grigory Yaroslavtsev. Private algorithms for the protected in social network search. *Proceedings of the National Academy of Sciences*, 113(4):913–918, 2016.

[3] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[4] mobiThinking. Global mobile statistics 2014 Part A: Mobile subscribers; handset market share; mobile operators. *mobiForge*, May 16, 2014.

[5] Aaron Segal, Joan Feigenbaum, and Bryan Ford. Open, privacy-preserving protocols for lawful surveillance. http://arxiv.org/abs/1607.03659, July 2016.

[6] Aaron Segal, Bryan Ford, and Joan Feigenbaum. Catching bandits and only bandits: Privacy-preserving intersection warrants for lawful surveillance. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*, San Diego, CA, August 2014. USENIX Association.