Defending Email Communication Against Profiling Attacks

Philippe Golle
Palo Alto Research Center
pgolle@parc.com

Ayman Farahat
Palo Alto Research Center
farahat@parc.com

ABSTRACT

We define message privacy against a *profiling* adversary, whose goal is to classify a population of users into categories according to the messages they exchange. This adversary models the most common privacy threat against email communication. We propose a protocol that protects senders and receivers of email messages from profiling attacks.

Categories and Subject Descriptors: C.2.0 [Computer-Communication Networks]: General — Security

General Terms: Security.

Keywords: Privacy, profiling, email, encryption.

1. INTRODUCTION

Unencrypted email communication offers no privacy. The contents of email messages are exposed to a number of intermediaries between senders and receivers, such as web-based email providers, Internet service providers, backbone Internet routers, as well as, potentially, intelligence agencies or governments. A recent proposal by a web-based email provider [4] to deliver contextual advertising alongside emails is a stark reminder of this threat to privacy.

Standard symmetric or public key encryption guarantees message privacy in a strong cryptographic sense, but it requires cumbersome mechanisms to exchange symmetric keys or distribute public keys. Only users with strong privacy needs are willing to bear the costs of encrypting their email messages. To date, the vast majority of email messages is still sent in plaintext.

In fact, encryption is an overkill that poorly addresses the most common privacy needs of email users. The communication of a typical email user is rarely of any value to an adversary, and is thus unlikely to be targeted in isolation. A whole population of users, on the other hand, may in aggregate attract the interest of an eavesdropping adversary. The real threat to privacy, therefore, comes from attacks which target a population of users rather than a specific individual. We call such attacks *profiling attacks.*

A profiling adversary may, for example, scan for keywords all emails sent and received by a population of users, and build profiles of users. Depending on the adversary, these profiles may be used for marketing purposes, or to generate leads in counter-terrorism or anti-criminal investigations. A profiling adversary typically has few resources to devote to any given message or user. It makes little sense indeed for a marketing or intelligence agency to expand vast amounts of resources to analyze a single message sent by a single person (except, of course, if that person is a known criminal or terrorist, but the privacy needs of known criminals or terrorists differ from those of the average population). To a profiling adversary, the real value lies in the mass classification of all messages sent by the whole user population.

This weaker but potentially more realistic model of the adversary allows us to propose a very efficient new solution to the problem of protecting the privacy of email communication. Our solution dispenses with the cumbersome requirement for key exchange that has hampered the deployment of email encryption, and requires little change to the existing email infrastructure.

2. PROFILING ATTACKS

The cryptographic literature typically defines message privacy with respect to a *targeted* attack. For example, the definition of semantic security [3] considers an adversary who must guess a random bit b, given the encryption $E(M_b)$ of one of two messages M_0 and M_1 of his choice. The attack is targeted in the sense that it involves only two messages M_0 and M_1 chosen by the adversary.

In contrast, we consider message privacy against an adversary $\mathcal{A}$ (called a profiling adversary), who observes passively *all* the messages exchanged between a large group of users, and attempts to identify all the users that satisfy a certain criterion based on the messages they send and receive. For example, $\mathcal{A}$ may be interested in messages that contain the keyword "Bomb". More sophisticated criteria may attempt to classify users into demographic or marketing categories.

We say that an encryption scheme offers privacy against profiling if $\mathcal{A}$, given all messages, is not significantly more successful at classifying users who encrypt their messages than it would be without seeing any message. Semantically secure encryption trivially guarantees privacy against profiling. More interestingly, a "weak" encryption scheme that is not secure in the cryptographic sense may still ensure privacy against profiling, if used by enough people that $\mathcal{A}$ has insufficient resources to break the encryption for more than a small fraction of users.

WPES'04, October 28, 2004, Washington, DC, USA.
Copyright 2004 ACM 1-58113-968-3/04/0010 ...$5.00.

In this paper, we propose an approach to privacy against profiling that is based on weak encryption (thus in particular we dispense with the need to exchange keys). We must ensure wide enough deployment of our weak encryption scheme to overwhelm the ability of the adversary to decrypt ciphertexts. Our approach is to design an encryption scheme that produces ciphertext that is indistinguishable from plaintext. This prevents the adversary from isolating users who encrypt their communication. From the view point of the adversary, it is thus as difficult to classify users who encrypt their messages as if every message were encrypted.

3. ENCRYPTING ENGLISH TEXT

In this section, we sketch the design of an encryption scheme for English text that uses techniques from natural language processing to produce ciphertext that is hard to distinguish from standard English text (to a machine observer), but hides the semantic of the plaintext. This encryption scheme is weak in the traditional cryptographic sense, yet it may very successfully thwart profiling attacks. For concreteness we present our encryption scheme for English text, but our techniques are general and could be adapted to other languages with little effort. (A complete description of this scheme, and experimental results, can be found in the full version of this paper [5].)

In a nutshell, our encryption function replaces every word of plaintext with a word of ciphertext drawn from an English dictionary. The word of ciphertext is selected according to the output of a deterministic function φ_k, parameterized by the choice of an encryption key k. The function φ_k takes as input the plaintext word and possibly other variables (the index of the plaintext word in the text, etc.) and outputs a word of ciphertext. The function φ must achieve two conflicting goals: the ciphertext must hide the semantic of the plaintext, while preserving the appearance of English text.

We propose to draw φ_k from among functions that map a plaintext word to a ciphertext word that belongs to the same grammatical category and has approximately the same frequency in standard English. Such functions will hide the semantic contents of the message, except for what can be inferred from its grammatical structure. These functions will also produce ciphertext that passes the statistical tests that are most commonly used to determine whether a document consists of standard English text, such as tests based on Part-of-Speech tagging or Zipf's Law (see [6] for more detail on computational linguistics.) These statistical tests are by no means exhaustive, and we anticipate that the design of encryption schemes secure against a profiling adversary will take the flavor of an "arms' race", where every advance made by the adversary in distinguishing encrypted text from normal English text will be matched by a corresponding increase in the sophistication of the encryption functions φ_k.

The encryption functions we propose here bear some superficial resemblance to some lexical steganographic techniques such as Mimic functions [7] or [1], which are semantically secure but much less efficient (they produce ciphertext that is much larger than the plaintext).

4. ENHANCING EMAIL PRIVACY

We propose a protocol for email communication that is secure against profiling attacks. Our protocol ensures that encrypted emails reveal no semantic information about their content to a machine observer, and are statistically indistinguishable from normal emails.

Our protocol uses the encryption scheme of Section 3 to encrypt the email. Recall that the encryption function φ_k is parameterized by the choice of a key k. We propose to use a slow one-way function [2] to generate the key k, i.e. a publicly known function h that is moderately costly to evaluate (say, on the order of a few seconds of computation) and very hard to invert.

More precisely, the sender of an email proceeds as follows. Let us denote M the message body of the email to be sent, and H the header of that email (consisting of the address of the sender, the address of the recipient, the time at which the email is sent and potentially other fields). For simplicity, we assume that all the fields in H are known both to the sender and recipient of the email (if that is not the case, we may define H as a subset of the header that is known to both the sender and recipient).

The sender computes a key $k = h(H)$ by applying the slow one-way function to the header H, then encrypts the body of the email with φ_k. The message sent is $H||\varphi_k(M)$. Upon receiving this message, the recipient recovers the key k by computing $k = h(H)$, then uses k to decrypt the encrypted body $\varphi_k(M)$ of the email.

Note that our protocol dispenses with key exchange. Anyone can compute the decryption key from the header of the message. Whereas the intended recipient can decrypt a small number of messages at relatively low computational cost, a profiling adversary attempting to decrypt a large number of messages would incur a tremendous computational cost. Furthermore, since our encryption scheme produces ciphertext that is machine indistinguishable from plaintext, the adversary can not separate encrypted communication from the rest. Our protocol is thus secure against profiling attacks even if only a small fraction of users encrypt their communication (a human can distinguish ciphertext from plaintext but the cost of using a human for profiling large amounts of messages is prohibitive).

Conclusion. We propose a realistic model of the most common threat to email privacy, and a protocol that protects senders and recipients of email from that threat. We discuss our approach to email privacy in more detail and describe a prototype implementation of our techniques in the full version of this paper [5].

5. REFERENCES

[1] M. Chapman and G. Davida. Hiding the hidden: a software system for concealing ciphertext as innocuous text. In *Proc. of the ICICS '97,* pp. 333–345.

[2] C. Dwork and M. Naor. Pricing via processing or combatting junkmail. In *Proc. of CRYPTO '92.*

[3] S. Goldwasser and S. Micali. Probabilistic Encryption. In *J. Com. Sys. Sci.* 28 (1984), pp. 270–299.

[4] Gmail. `http://www.gmail.com`

[5] A. Korolova, A. Farahat and P. Golle. Defending email communication against profiling attacks. Full paper available at `http://crypto.stanford.edu/~pgolle/`

[6] Christopher D. Manning and Hinrich Schütze. *Review of Foundations of Statistical Natural Language Processing.* MIT press, 1999.

[7] P. Wayner. Mimic functions. In CRYPTOLOGIA, Volume 16, Number 3, pp. 193-214, July 1992.