Proceedings on Privacy Enhancing Technologies ; 2019 (4):6–33

Kirill Nikitin*, Ludovic Barman*, Wouter Lueks, Matthew Underwood, Jean-Pierre Hubaux und Bryan Ford

Reducing Metadata Leakage from Encrypted Files and Communication with PURBs

Abstract: Most encrypted data formats leak metadata via their plaintext headers, such as format version, encryption schemes used, number of recipients who can decrypt the data, and even the recipients' identities. This leakage can pose security and privacy risks to users, *e.g.,* by revealing the full membership of a group of collaborators from a single encrypted e-mail, or by enabling an eavesdropper to fingerprint the precise encryption software version and configuration the sender used.

We propose that future encrypted data formats improve security and privacy hygiene by producing *Padded Uniform Random Blobs* or PURBs: ciphertexts indistinguishable from random bit strings to anyone without a decryption key. A PURB's content leaks *nothing at all*, even the application that created it, and is padded such that even its length leaks as little as possible.

Encoding and decoding ciphertexts with *no* cleartext markers presents efficiency challenges, however. We present cryptographically agile encodings enabling legitimate recipients to decrypt a PURB efficiently, even when encrypted for any number of recipients' public keys and/or passwords, and when these public keys are from different cryptographic suites. PURBs employ Padmé, a novel padding scheme that limits information leakage via ciphertexts of maximum length M to a practical optimum of $O(\log\log M)$ bits, comparable to padding to a power of two, but with lower overhead of at most 12% and decreasing with larger payloads.

Keywords: metadata, leakage, padding, traffic analysis

DOI 10.2478/popets-2019-0056

Received 2019-02-28; revised 2019-06-15; accepted 2019-06-16.

1 Introduction

Traditional encryption schemes and protocols aim to protect only their data payload, leaving related metadata exposed. Formats such as PGP [64] reveal in cleartext headers the public keys of the intended recipients, the algorithm used for encryption, and the actual length of the payload. Secure-communication protocols similarly leak information during key and algorithm agreement. The TLS handshake [45], for example, leaks in cleartext the protocol version, chosen cipher suite, and the public keys of the parties. This metadata exposure is traditionally assumed not to be security-sensitive, but important for the recipient's decryption efficiency.

Research has consistently shown, however, that attackers can exploit metadata to infer sensitive information about communication content. In particular, an attacker may be able to fingerprint users [40, 52] and the applications they use use [63]. Using traffic analysis [17], an attacker may be able to infer websites a user visited [17, 21, 39, 56, 57] or videos a user watched [43, 44, 50]. On VoIP, metadata can be used to infer the geo-location [35], the spoken language [61], or the voice activity of users [15]. Side-channel leaks from data compression [32] facilitate several attacks on SSL [5, 25, 48]. The lack of proper padding might enable an active attacker to learn the length of the user's password from TLS [53] or QUIC [1] traffic. In social networks, metadata can be used to draw conclusions about users' actions [26], whereas telephone metadata has been shown to be sufficient for user re-identification and for determining home locations [36]. Furthermore, by observing the format of packets, oppressive regimes can infer which technology is used and use this information for the purposes of incrimination or censorship. Most TCP packets that Tor sends, for example, are 586 bytes due to its standard cell size [27].

As a step towards countering these privacy threats, we propose that encrypted data formats should produce *Padded Uniform Random Blobs* or PURBs: ciphertexts designed to protect *all* encryption metadata. A PURB encrypts application content and metadata into a single blob that is indistinguishable from a random string,

***Corresponding Author: Kirill Nikitin:** EPFL, E-mail: kirill.nikitin@epfl.ch

***Corresponding Author: Ludovic Barman:** EPFL, E-mail: ludovic.barman@epfl.ch

Wouter Lueks: EPFL, E-mail: wouter.lueks@epfl.ch

Matthew Underwood: unaffiliated

Jean-Pierre Hubaux: EPFL, E-mail: jean-pierre.hubaux@epfl.ch

Bryan Ford: EPFL, E-mail: bryan.ford@epfl.ch