Network Working Group
Internet-Draft
Intended status: Informational
Expires: July 13, 2012

M. Hansen
ULD Kiel
H. Tschofenig
Nokia Siemens Networks
R. Smith
JANET(UK)

January 10, 2012

Privacy Terminology
draft-iab-privacy-terminology-00.txt

Abstract

Privacy is a concept that has been debated and argued throughout the last few millennia by all manner of people. Its most striking feature is that nobody seems able to agree upon a precise definition of what it actually is. In order to discuss privacy in any meaningful way a tightly defined context needs to be elucidated. The specific context of privacy used within this document is that of "personal data", any information relating to a data subject; a data subject is an identified natural person or a natural person who can be identified, directly or indirectly. This context is highly relevant since a lot of work within the IETF involves defining protocols that can potentially transport (either explicitly or implicitly) personal data.

This document aims to establish a basic lexicon around privacy so that IETF contributors who wish to discuss privacy considerations within their work can do so using terminology consistent across the area.

Note: This document is discussed at https://www.ietf.org/mailman/listinfo/ietf-privacy

material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 13, 2012.

Table of Contents

1. Introduction

Privacy is a concept that has been debated and argued throughout the last few millennia by all manner of people, including philosophers, psychologists, lawyers, and more recently, computer scientists. Its most striking feature is that nobody seems able to agree upon a precise definition of what it actually is. Every individual, every group, and every culture have their own different views and preconceptions about the concept - some mutually complimentary, some distinctly different. However, it is generally (but not unanimously!) agreed that the protection of privacy is "A Good Thing" and often, people only realize what it was when they feel that they have lost it.

In order to discuss privacy in any meaningful way a tightly defined context needs to be elucidated. The specific context of privacy used within this document is that of "personal data", any information relating to a data subject; a (data) subject is an identified natural person or a natural person who can be identified, directly or indirectly. We A lot of work within the IETF involves defining protocols that can potentially transport personal data and can therefore either, by dint of design decisions when creating them, enable either privacy protection or result in privacy breaches. While identifiers and data elements communicated in protocols often do not assume a specific association with a human using the software. However, a protocol may help or simplify the re-identification of a natural person by the choice of their identifiers and other state that is established and communicated, particularly when information from various sources is combined and analyzed together.

Work in this area of privacy and privacy protection over the last few decades has centered on the idea of data minimization; it uses terminologies such as anonymity, unlinkability, unobservability, and pseudonymity. These terms are often used in discussions about the privacy properties of systems.

The core principal of data minimization is that the ability for others to collect personal data should be removed or at least minimized when this is either not desirable or when it cannot be entirely prevented.

Data minimization is the only generic strategy to enhance individual privacy in cases where valid personal information is used since all valid personal data inherently provides some linkability. Other techniques have been proposed and implemented that aim to enhance privacy by providing misinformation (inaccurate or erroneous information, provided usually without conscious effort to mislead or deceive) or disinformation (deliberately false or distorted

information provided in order to mislead or deceive). However, these techniques are out of scope for this document.

We use the term 'attacker' in this writeup to refer to an entity that violates the privacy expectations of a data subject. The attacker may not only be an entity that is external to the system but, in many cases, the attacker is actually one of the communication partners. When necessary we use the term initiator and responder to refer to the communication interaction of a protocol. This particular terminology is used to highlight that many protocols utilize bidirectional communication where both ends send and receive data. We assume that the attacker uses all information available to infer (probabilities of) his items of interest (IOIs). These IOIs may be attributes (and their values) of personal data, or may be actions such as who sent, or who received, which messages.

This document aims to establish a basic lexicon around privacy so that IETF contributors who wish to discuss privacy considerations within their work (see [I-D.iab-privacy-considerations]) can do so using terminology consistent across areas. Note that it does not attempt to define all aspects of privacy terminology, rather it just establishes terms to some of the most common ideas and concepts.

2. Anonymity

Definition: Anonymity of a subject means that the attacker cannot sufficiently identify the subject within a set of subjects, the anonymity set.

To enable anonymity of a subject, there always has to be an appropriate set of subjects with potentially the same attributes. The set of all possible subjects is known as the anonymity set, and membership of this set may vary over time.

The set of possible subjects depends on the knowledge of the attacker. Thus, anonymity is relative with respect to the attacker. Therefore, an initiator may be anonymous (initiator anonymity) only within a set of potential initiators - their initiator anonymity set - which itself may be a subset of all subjects who may send a message. Conversely a responder may be anonymous (responder anonymity) only within a set of potential responders - their responder anonymity set. Both anonymity sets may be disjoint, may overlap, or may be the same.

As an example consider RFC 3325 (P-Asserted-Identity, PAI) [RFC3325], an extension for the Session Initiation Protocol (SIP), that allows subjects, such as a VoIP caller, to instruct an intermediary he or she trusts not to populate the SIP From header field with its authenticated and verified identity. The recipient of the call, as well as any other entity outside the data subjects's trust domain, would therefore only learn that the SIP message (typically a SIP INVITE) was sent with a header field 'From: "Anonymous" <sip:anonymous@anonymous.invalid>' rather than the subject's address-of-record, which is typically thought of as the "public address" of the user, i.e., the data subject. When PAI is used the subject becomes anonymous within the initiator anonymity set that is populated by every subject making use of that specific intermediary.

Note that this example assumes that other personal data cannot be inferred from the other SIP protocol payloads, which is a useful assumption to be made in the analysis of one specific protocol extension but not for analysis of an entire architecture.

3. Unlinkability

Definition: Unlinkability of two or more Items Of Interest (e.g., subjects, messages, actions, ...) means that within a particular set of information, the attacker cannot distinguish whether these IOIs are related or not (with a high enough degree of probability to be useful).

Unlinkability of two (or more) messages may of course depend on whether their content is protected against the attacker. In the cases where this is not true, messages may only be unlinkable if we assume that the attacker is not able to infer information about the initiator or responder from the message content itself. It is worth noting that even if the content itself does not betray linkable information explicitly, deep semantical analysis of a message sequence can often detect certain characteristics which link them together, e.g., similarities in structure, style, use of some words or phrases, consistent appearance of some grammatical errors, etc.

The unlinkability property can be considered as a more "fine-grained" version of anonymity since there are many more relations where unlinkability might be an issue than just the relation of "anonymity" between subjects and IOIs. As such, it may sometimes be necessary to explicitly state to which attributes anonymity refers to (beyond the subject to IOI relationship). An attacker might get to know information on linkability of various messages while not necessarily reducing anonymity of the particular subject. As an example an attacker, in spite of being able to link all encrypted messages in a set of transactions, does not learn the identify of the subject who is the source of the transactions.

There are several items of terminology heavily related to unlinkability:

Definition: We use the term "profiling" to mean learning information about a particular subject while that subject remains anonymous to the attacker. For example, if an attacker concludes that a subject plays a specific computer game, reads specific news article on a website, and uploads certain videos, then the subjects activities have been profiled, even if the attacker is unable to identify that specific subject.

Definition: "Relationship anonymity" of a pair of subjects means that sender and recipient (or each recipient in case of multicast) are unlinkable. The classical MIX-net [Chau81] without dummy traffic is one implementation with just this property: The attacker sees who sends messages when, and who receives messages when, but cannot figure out who is sending messages to whom.

Definition: The term "unlinkable session" refers the ability of the system to render a set of actions by a subject unlinkable from one another over a sequence of protocol runs (sessions). This term is useful for cases where a sequence of interactions between an initiator and a responder is necessary for the application logic rather than a single-shot message. We refer to this as a session. When doing an analysis with respect to unlinkability we compare this session to a sequence of sessions to determine linkability.

Definition: We use the term "fingerprinting" to refer to any parameter (or set of parameters) that an attacker can observe for the purpose of re-identification. Fingerprinting is a form of tracking by associating activities of a communication software at different times, potentially with different communication partners, but without explicitly sharing state information (as it would be the case with cookies [RFC6265]). For example, the Panopticlick project by the Electronic Frontier Foundation uses parameters an HTTP-based Web browser shares with sites it visits to determine the uniqueness of the browser [panopticlick].

4. Undetectability

Definition: Undetectability of an item of interest (IOI) means that the attacker cannot sufficiently distinguish whether it exists or not.

In contrast to anonymity and unlinkability, where the IOI is protected indirectly through protection of the IOI's relationship to a subject or other IOI, undetectability is the direct protection of an IOI. For example, undetectability can be regarded as a possible and desirable property of steganographic systems.

If we consider messages as IOIs, then undetectability means that messages are not sufficiently discernible from, e.g., "random noise".

5. Pseudonymity

Definition: A pseudonym is an identifier of a subject other than one of the subject's real names. An identifier, as defined in [id], is "a lexical token that names entities".

In the context of IETF protocols almost all identifiers are pseudonyms since there is typically no requirement to use real names in protocols. However, in certain scenario it is reasonable to assume that real names will be used and it will be worthwhile to point out this circumstance.

For Internet protocols it is important whether protocols allow identifiers to be recycled dynamically, what the lifetime of the pseudonyms are, to whom they get exposed, how subjects are able to control disclosure, and how often they can be changed over time (and what the consequences are when they are regularly changed). These aspects are described in [I-D.iab-privacy-considerations].

Achieving anonymity, unlinkability, and maybe undetectability may enable the ideal of data minimization. Unfortunately, it would also prevent a certain class of useful two-way communication scenarios. Therefore, for many applications, we need to accept a certain amount of linkability and detectability while attempting to retain unlinkability between the subject and their transactions. This is achieved through appropriate kinds of pseudonymous identifiers. These identifiers are then often used to refer to established state or are used for access control purposes, see [I-D.iab-identifier-comparison].

The term 'real name' is the antonym to "pseudonym". There may be multiple real names over a lifetime -- in particular legal names. For example, a human being may possess the names which appear on their birth certificate or on other official identity documents issued by the State; for a legal person the name under which it operates and which is registered in official registers (e.g., commercial register or register of associations). A human being's real name typically comprises their given name and a family name. Note that from a mere technological perspective it cannot always be determined whether an identifier of a subject is a pseudonym or a real name.

Additional useful terms are:

Definition: The "holder" of the pseudonym is the subject to whom the pseudonym refers.

Definition: A subject is "pseudonymous" if a pseudonym is used as identifier instead of one of its real names.

Definition: Pseudonymity is the state of remaining pseudonymous through the use of pseudonyms as identifiers.

Sender pseudonymity is defined as the sender being pseudonymous, recipient pseudonymity is defined as the recipient being pseudonymous.

Anonymity through the use of pseudonyms is stronger where ...

- the less personal data of the pseudonym holder can be linked to the pseudonym;
- the less often and the less context-spanning pseudonyms are used and therefore the less data about the holder can be linked;
- the more often independently chosen pseudonyms are used for new actions (i.e., making them, from an observer's perspective, unlinkable)

6. Acknowledgments

Parts of this document utilizes content from [anon_terminology], which had a long history starting in 2000 and whose quality was improved due to the feedback from a number of people. The authors would like to thank Andreas Pfitzmann for his work on an earlier draft version of this document.

Within the IETF a number of persons had provided their feedback to this document. We would like to thank Scott Brim, Marc Linsner, Bryan McLaughlin, Nick Mathewson, Eric Rescorla, Alissa Cooper, Scott Bradner, Nat Sakimura, Bjoern Hoehrmann, David Singer, Dean Willis, Christine Runnegar, Lucy Lynch, Trend Adams, Mark Lizar, Martin Thomson, Josh Howlett, and Mischa Tuffield.

7. Security Considerations

This document introduces terminology for talking about privacy within IETF specifications. Since privacy protection often relies on security mechanisms then this document is also related to security in its broader context.

8. IANA Considerations

This document does not require actions by IANA.

9. References

9.1. Normative References

[I-D.iab-privacy-considerations] Cooper, A., Tschofenig, H., Aboba, B., Peterson, J., and J. Morris, "Privacy Considerations for Internet Protocols", draft-iab-privacy-considerations-01 (work in progress), October 2011.

[id] "Identifier - Wikipeadia", Wikipedia , URL: http://en.wikipedia.org/wiki/Identifier, 2011.

9.2. Informative References

[Chau81] Chaum, D., "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms", Communications of the ACM , 24/2, 84-88, 1981.

[I-D.iab-identifier-comparison] Thaler, D., "Issues in Identifier Comparison for Security Purposes", draft-iab-identifier-comparison-00 (work in progress), July 2011.

[RFC3325] Jennings, C., Peterson, J., and M. Watson, "Private Extensions to the Session Initiation Protocol (SIP) for Asserted Identity within Trusted Networks", RFC 3325, November 2002.

[RFC6265] Barth, A., "HTTP State Management Mechanism", RFC 6265, April 2011.

[anon_terminology] Pfitzmann, A. and A. Pfitzmann, "A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management", URL: http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf , version 034, 2010.

[panopticlick] Eckersley, P., "How Unique Is Your Web Browser?", Electronig Frontier Foundation , URL: https://panopticlick.eff.org/browser-uniqueness.pdf, 2009.

Authors' Addresses

Marit Hansen
ULD Kiel

EMail: marit.hansen@datenschutzzentrum.de

Hannes Tschofenig
Nokia Siemens Networks
Linnoitustie 6
Espoo 02600
Finland

Phone: +358 (50) 4871445
EMail: Hannes.Tschofenig@gmx.net
URI: http://www.tschofenig.priv.at

Rhys Smith
JANET(UK)

EMail: rhys.smith@ja.net