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

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

March 12, 2012

Privacy Terminology and Concepts
draft-iab-privacy-terminology-01.txt

Abstract

Privacy is a concept that has been debated and argued throughout the last few millennia. Its most striking feature is the difficulty that disparate parties encounter when they attempt to precisely define it. In order to discuss privacy in a meaningful way, a tightly defined context is necessary. The specific context of privacy used within this document is that of personal data in Internet protocols. Personal data is any information relating to a data subject, where a data subject is an identified natural person or a natural person who can be identified, directly or indirectly.

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 consistent lexicon around privacy for IETF contributors to use when discussing privacy considerations within their work.

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

Status of This Memo

Copyright Notice

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 the difficulty that disparate parties encounter when they attempt to precisely define it. Each individual, group, and culture has its own views and preconceptions about privacy, some of which are mutually complimentary and some of which diverge. However, it is generally (but not unanimously) agreed that the protection of privacy is "A Good Thing." People often do not realize how they value privacy until they lose it.

In order to discuss privacy in a meaningful way, a tightly defined context is necessary. The specific context of privacy used within this document is that of "personal data" in Internet protocols. Personal data is any information relating to a data subject, where a data subject is an identified natural person or a natural person who can be identified, directly or indirectly.

A lot of work within the IETF involves defining protocols that can potentially transport personal data. Protocols are therefore capable of enabling both privacy protections and privacy breaches. Protocol architects often do not assume a specific relationship between the identifiers and data elements communicated in protocols and the humans using the software running the protocols. However, a protocol may facilitate the identification of a natural person depending on how protocol identifiers and other state are created and communicated.

One commonly held privacy objective is that of data minimization -- eliminating the potential for personal data to be collected. Often, however, the collection of personal data cannot not be prevented entirely, in which case the goal is to minimize the amount of personal data that can be collected for a given purpose and to offer ways to control the dissemination of personal data. This document focuses on introducing terms used to describe privacy properties that support data minimization.

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). These techniques are out of scope for this document.

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 discusses terms describing the most common ideas and concepts.

2. Basic Terms

Personal data: Any information relating to a data subject.

Data subject: An identified natural person or a natural person who can be identified, directly or indirectly.

Item of Interest (IOI): Any data item that an observer or attacker might be interested in. This includes attributes, identifiers, communication actions (such as sending data to or receiving data from certain communication partners), etc.

Initiator: The protocol entity that starts a communication interaction with a recipient. The term "initiator" is used rather than "sender" to highlight the fact that many protocols use bidirectional communication where both ends send and receive data

Recipient: A protocol entity that recieves communications from an initiator.

Attacker: An entity that intentionally works against some protection goal. It is assumed that an attacker uses all information available to infer information about its items of interest.

Observer: A protocol entity that is authorized to receive and handle data from an initiator and thereby is able to observe and collect information, potentially posing privacy threats depending on the context. These entities are not generally considered as "attackers" in the security sense, but they are still capable of privacy invasion.

3. Identifiability

Identity: Any subset of a data subject's attributes that identifies the data subject within a given context. Data subjects usually have multiple identities for use in different contexts.

Identifier: A data object that represents a specific identity of a protocol entity or data subject. See [RFC4949].

Identifiability: The extent to which a data subject is identifiable.

Identification: The linking of information to a particular data subject to infer the subject's identity.

The following sub-sections define terms related to different ways of reducing identifiability.

3.1. Anonymity

Anonymous: A property of a data subject in which an observer or attacker cannot identify the data subject within a set of other subjects (the anonymity set).

Anonymity: The state of being anonymous.

To enable anonymity of a data subject, there must exist a set of data subjects with potentially the same attributes, i.e., to the attacker or the observer these data subjects must appear indistinguishable from each other. The set of all such data subjects is known as the anonymity set and membership of this set may vary over time.

The composition of the anonymity set depends on the knowledge of the observer or attacker. Thus anonymity is relative with respect to the observer or attacker. An initiator may be anonymous only within a set of potential initiators -- its initiator anonymity set -- which itself may be a subset of all data subjects that may initiate communications. Conversely, a recipient may be anonymous only within a set of potential receipients -- its receipient 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 a data subject, such as a VoIP caller, to instruct an intermediary that he or she trusts not to populate the SIP From header field with the subject's authenticated and verified identity. The recipient of the call, as well as any other entity outside of the data subject'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 (the data subject). When PAI is used, the data subject becomes anonymous within the initiator anonymity set that is populated by every data subject making use of that specific intermediary.

Note: This example ignores the fact that other personal data may be inferred from the other SIP protocol payloads. This caveat makes the analysis of the specific protocol extension easier but cannot be assumed when conducting analysis of an entire architecture.

3.2. Pseudonymity

Pseudonym: An identifier of a subject other than one of the subject's real names.

Real name: The opposite of a pseudonym. For example, a natural person may possess the names that appear on his or her birth certificate or on other official identity documents issued by the state. A natural person's real name typically comprises his or her given names and a family name. A data subject may have multiple real names over a lifetime, including legal names. Note that from a technological perspective it cannot always be determined whether an identifier of a data subject is a pseudonym or a real name.

Pseudonymous: A property of a data subject in which the subject is identified by a pseudonym.

Pseudonymity: The state of being pseudonymous.

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 scenarios it is reasonable to assume that real names will be used (with vCard [RFC6350], for example).

Pseudonymity is strengthened when less personal data can be linked to the pseudonym; when the same pseudonym is used less often and across fewer contexts; and when independently chosen pseudonyms are more frequently used for new actions (making them, from an observer's or attacker's perspective, unlinkable).

For Internet protocols it is important whether protocols allow pseudonyms to be changed without human interaction, the default length of pseudonym lifetimes, to whom pseudonyms are exposed, how data subjects are able to control disclosure, how often pseudonyms

can be changed, and the consequences of changing them. These aspects are described in [I-D.iab-privacy-considerations].

3.3. Identity Confidentiality

Identity confidentiality: A property of a data subject wherein any party other than the recipient cannot sufficiently identify the data subject within the anonymity set. In comparison to anonymity and pseudonymity, identity confidentiality is concerned with eavesdroppers and intermediaries.

As an example, consider the network access authentication procedures utilizing the Extensible Authentication Protocol (EAP) [RFC3748]. EAP includes an identity exchange where the Identity Response is primarily used for routing purposes and selecting which EAP method to use. Since EAP Identity Requests and Responses are sent in cleartext, eavesdroppers and intermediaries along the communication path between the EAP peer and the EAP server can snoop on the identity. To address this treat, as discussed in RFC 4282 [RFC4282], the user's identity can be hidden against these observers with the cryptography support by EAP methods. Identity confidentiality has become a recommended design criteria for EAP (see [RFC4017]). EAP-AKA [RFC4187], for example, protects the EAP peer's identity against passive adversaries by utilizing temporal identities. EAP-IKEv2 [RFC5106] is an example of an EAP method that offers protection against active observers with regard to the data subject's identity.

3.4. Identity Management

Identity Provider (IdP): An entity (usually an organization) that has a relationship with a data subject and is responsible for providing authentication and authorization information to relying parties (see below). To facilitate the provision of authentication and authorization, an IdP will usually go through a process of verifying the data subject's identity and issuing the subject a set of credentials. Each function that the IdP performs -- identity verification, credential issuing, providing authentication assertions, providing authorization assertions, and so forth -- may be performed by separate entities, but for the purposes of this document, it is assumed that a single entity is performing all of them.

Relying Party (RP): An entity that relies on authentication and authorization of a data subject provided by an identity provider, typically to process a transaction or grant access to information or a system.

4. Unlinkability

Unlinkability: Within a particular set of information, a state in which an observer or attacker cannot distinguish whether two items of interest are related or not (with a high enough degree of probability to be useful to the observer or attacker).

Unlinkability of two or more messages may depend on whether their content is protected against the observer or attacker. In the cases where this is not true, messages may only be unlinkable if it is assumed that the observer or attacker is not able to infer information about the initiator or receipient from the message content itself. It is worth noting that even if the content itself does not betray linkable information explicitly, deep semantic analysis of a message sequence can often detect certain characteristics that link them together, including similarities in structure, style, use of particular words or phrases, consistent appearance of certain grammatical errors, and so forth.

There are several items of terminology highly related to unlinkability:

Correlation: The combination of various pieces of information about a data subject. For example, if an observer or attacker concludes that a data subject plays a specific computer game, reads a specific news article on a website, and uploads specific videos, then the data subject's activities have been correlated, even if the observer or attacker is unable to identify the specific data subject.

Relationship anonymity: When an initiator and receipient (or each recipient in the case of multicast) are unlinkable. The classical MIX-net [Chau81] without dummy traffic is one implementation with this property: the observer sees who sends and receives messages and when they are sent and received, but it cannot figure out who is sending messages to whom.

Unlinkable protocol interaction: When one protocol interaction is not linkable to another protocol interaction of the same protocol.

An example of a protocol that does not provide this property is Transport Layer Security (TLS) session resumption [RFC5246] or the TLS session resumption without server side state [RFC5077]. In RFC 5246 [RFC5246] a server provides the client with a session_id in the ServerHello message and caches the master_secret for later exchanges. When the client initiates a new connection with the server it re-uses the previously obtained session_id in its ClientHello message. The server agrees to resume the session by

using the same session_id and the previously stored master_secret for the generation of the TLS Record Layer security association. RFC 5077 [RFC5077] borrows from the session resumption design idea but the server encapsulates all state information into a ticket instead of caching it. An attacker who is able to observe the protocol exchanges between the TLS client and the TLS server is able to link the initial exchange to subsequently resumed TLS sessions when the session_id and the ticket is exchanged in clear (which is the case with data exchange in the initial handshake messages).

Fingerprinting: The process of an observer or attacker partially or fully identifying a device, application, or initiator based on multiple information elements communicated to the observer or attacker. 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].

5. Undetectability

Undetectability: The state in which an observer or attacker cannot sufficiently distinguish whether an item of interest 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 means the IOI is directly protected. For example, undetectability is as a desirable property of steganographic systems.

If we consider the case where an IOI is a message, then undetectability means that the message is not sufficiently discernible from other messages (from, e.g., random noise).

Achieving anonymity, unlinkability, and undetectability may enable extreme data minimization. Unfortunately, this would also prevent a certain class of useful two-way communication scenarios. Therefore, for many applications, a certain amount of linkability and detectability is usually accepted while attempting to retain unlinkability between the data subject and his or her transactions. This is achieved through the use of 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].

6. Example

[To be provided in a future version once the guidance is settled.]

7. 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, Scott Bradner, Nat Sakimura, Bjoern Hoehrmann, David Singer, Dean Willis, Christine Runnegar, Lucy Lynch, Trend Adams, Mark Lizar, Martin Thomson, Josh Howlett, Mischa Tuffield, S. Moonesamy, Ted Hardie, Zhou Sujing, Claudia Diaz, Leif Johansson, and Klaas Wierenga.

8. 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.

9. IANA Considerations

This document does not require actions by IANA.

10. References

10.1. 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 - Wikipedia", Wikipedia , URL: http://en.wikipedia.org/wiki/Identifier, Dec 2011.

10.2. 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.

[RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.

[RFC4017] Stanley, D., Walker, J., and B. Aboba, "Extensible Authentication Protocol (EAP) Method Requirements for Wireless LANs", RFC 4017, March 2005.

[RFC4187] Arkko, J. and H. Haverinen, "Extensible Authentication Protocol

Method for 3rd Generation Authentication and Key Agreement (EAP-AKA)", RFC 4187, January 2006.

[RFC4282] Aboba, B., Beadles, M., Arkko, J., and P. Eronen, "The Network Access Identifier", RFC 4282, December 2005.

[RFC4949] Shirey, R., "Internet Security Glossary, Version 2", RFC 4949, August 2007.

[RFC5077] Salowey, J., Zhou, H., Eronen, P., and H. Tschofenig, "Transport Layer Security (TLS) Session Resumption without Server-Side State", RFC 5077, January 2008.

[RFC5106] Tschofenig, H., Kroeselberg, D., Pashalidis, A., Ohba, Y., and F. Bersani, "The Extensible Authentication Protocol-Internet Key Exchange Protocol version 2 (EAP-IKEv2) Method", RFC 5106, February 2008.

[RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.

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

[RFC6350] Perreault, S., "vCard Format Specification", RFC 6350, August 2011.

[anon_terminology] Pfitzmann, A. and M. Hansen, "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

Alissa Cooper
CDT

EMail: acooper@cdt.org