Internet Engineering Task Force (IETF)
Request for Comments: 6235
Category: Experimental
ISSN: 2070-1721

E. Boschi
B. Trammell
ETH Zurich

May 2011

IP Flow Anonymization Support

Abstract

This document describes anonymization techniques for IP flow data and the export of anonymized data using the IP Flow Information Export (IPFIX) protocol. It categorizes common anonymization schemes and defines the parameters needed to describe them. It provides guidelines for the implementation of anonymized data export and storage over IPFIX, and describes an information model and Options-based method for anonymization metadata export within the IPFIX protocol or storage in IPFIX Files.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6235.

Table of Contents

1. Introduction

The standardization of an IP Flow Information Export (IPFIX) protocol [RFC5101] and associated representations removes a technical barrier to the sharing of IP flow data across organizational boundaries and with network operations, security, and research communities for a wide variety of purposes. However, with wider dissemination comes greater risks to the privacy of the users of networks under measurement, and to the security of those networks. While it is not a complete solution to the issues posed by distribution of IP flow information, anonymization (i.e., the deletion or transformation of information that is considered sensitive and that could be used to reveal the identity of subjects involved in a communication) is an important tool for the protection of privacy within network measurement infrastructures.

This document presents a mechanism for representing anonymized data within IPFIX and guidelines for using it. It is not intended as a general statement on the applicability of specific flow data anonymization techniques to specific situations or as a recommendation of any particular application of anonymization to flow data export. Exporters or publishers of anonymized data must take care that the applied anonymization technique is appropriate for the data source, the purpose, and the risk of deanonymization of a given application.

It begins with a categorization of anonymization techniques. It then describes the applicability of each technique to commonly anonymizable fields of IP flow data, organized by information element data type and semantics as in [RFC5102]; enumerates the parameters required by each of the applicable anonymization techniques; and provides guidelines for the use of each of these techniques in accordance with current best practices in data protection. Finally, it specifies a mechanism for exporting anonymized data and binding anonymization metadata to Templates and Options Templates using IPFIX Options.

1.1. IPFIX Protocol Overview

In the IPFIX protocol, { type, length, value } tuples are expressed in Templates containing { type, length } pairs, specifying which { value } fields are present in data records conforming to the Template, giving great flexibility as to what data is transmitted. Since Templates are sent very infrequently compared with Data Records, this results in significant bandwidth savings. Various different data formats may be transmitted simply by sending new Templates specifying the { type, length } pairs for the new data format. See [RFC5101] for more information.

The IPFIX information model [RFC5102] defines a large number of standard Information Elements (IEs) that provide the necessary { type } information for Templates. The use of standard elements enables interoperability among different vendors' implementations. Additionally, non-standard enterprise-specific elements may be defined for private use.

1.2. IPFIX Documents Overview

"Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information" [RFC5101] and its associated documents define the IPFIX protocol, which provides network engineers and administrators with access to IP traffic flow information.

"Architecture for IP Flow Information Export" [RFC5470] defines the architecture for the export of measured IP flow information out of an IPFIX Exporting Process to an IPFIX Collecting Process, and the basic terminology used to describe the elements of this architecture, per the requirements defined in "Requirements for IP Flow Information Export" [RFC3917]. The IPFIX Protocol document [RFC5101] then covers the details of the method for transporting IPFIX Data Records and Templates via a congestion-aware transport protocol from an IPFIX Exporting Process to an IPFIX Collecting Process.

"Information Model for IP Flow Information Export" [RFC5102] describes the Information Elements used by IPFIX, including details on Information Element naming, numbering, and data type encoding. Finally, "IP Flow Information Export (IPFIX) Applicability" [RFC5472] describes the various applications of the IPFIX protocol and their use of information exported via IPFIX and relates the IPFIX architecture to other measurement architectures and frameworks.

Additionally, "Specification of the IP Flow Information Export (IPFIX) File Format" [RFC5655] describes a file format based upon the IPFIX protocol for the storage of flow data.

This document references the Protocol and Architecture documents for terminology and extends the IPFIX Information Model to provide new Information Elements for anonymization metadata. The anonymization techniques described herein are equally applicable to the IPFIX protocol and data stored in IPFIX Files.

1.3. Anonymization within the IPFIX Architecture

According to [RFC5470], IPFIX Message anonymization is optionally performed as the final operation before handing the Message to the transport protocol for export. While no provision is made in the

architecture for anonymization metadata as in Section 6, this arrangement does allow for the rewriting necessary for comprehensive anonymization of IPFIX export as in Section 7. The development of the IPFIX Mediation [RFC6183] framework and the IPFIX File Format [RFC5655] expand upon this initial architectural allowance for anonymization by adding to the list of places that anonymization may be applied. The former specifies IPFIX Mediators, which rewrite existing IPFIX Messages, and the latter specifies a method for storage of IPFIX data in files.

More detail on the applicable architectural arrangements for anonymization can be found in Section 7.1

1.4. Supporting Experimentation with Anonymization

The status of this document is Experimental, reflecting the experimental nature of anonymization export support. Research on network trace anonymization techniques and attacks against them is ongoing. Indeed, there is increasing evidence that anonymization applied to network trace or flow data on its own is insufficient for many data protection applications as in [Bur10]. Therefore, this document explicitly does not recommend any particular technique or implementation thereof.

The intention of this document is to provide a common basis for interoperable exchange of anonymized data, furthering research in this area, both on anonymization techniques themselves as well as to the application of anonymized data to network measurement. To that end, the classification in Section 3 and anonymization export support in Section 6 can be used to describe and export information even about data anonymized using techniques that are unacceptably weak for general application to production datasets on their own.

While the specification herein is designed to be independent of the anonymization techniques applied and the implementation thereof, open research in this area may necessitate future updates to the specification. Assuming the future successful application of this specification to anonymized data publication and exchange, it may be brought back to the IPFIX working group for further development and publication on the Standards Track.

2. Terminology

Terms used in this document that are defined in the Terminology section of the IPFIX Protocol [RFC5101] document are to be interpreted as defined there. In addition, this document defines the following terms:

Anonymization Record: A record, defined by the Anonymization Options Template in Section 6.1, that defines the properties of the anonymization applied to a single Information Element within a single Template or Options Template.

Anonymized Data Record: A Data Record within a Data Set containing at least one Information Element with anonymized values. The Information Element(s) within the Template or Options Template describing this Data Record SHOULD have a corresponding Anonymization Record.

Intermediate Anonymization Process: An intermediate process that takes Data Records and transforms them into Anonymized Data Records.

Note that there is an explicit difference in this document between a "Data Set" (which is defined as in [RFC5101]) and a "data set". When in lower case, this term refers to any collection of data (usually, within the context of this document, flow or packet data) that may contain identifying information and is therefore subject to anonymization.

Note also that when the term Template is used in this document, unless otherwise noted, it applies both to Templates and Options Templates as defined in [RFC5101]. Specifically, Anonymization Records may apply to both Templates and Options Templates.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Categorization of Anonymization Techniques

Anonymization, as described by this document, is the modification of a dataset in order to protect the identity of the people or entities described by the dataset from disclosure. With respect to network traffic data, anonymization generally attempts to preserve some set of properties of the network traffic useful for a given application or applications, while ensuring the data cannot be traced back to the specific networks, hosts, or users generating the traffic.

Anonymization may be broadly classified according to two properties: recoverability and countability. All anonymization techniques map the real space of identifiers or values into a separate, anonymized space, according to some function. A technique is said to be recoverable when the function used is invertible or can otherwise be reversed and a real identifier can be recovered from a given replacement identifier. "Recoverability" as used within this

categorization does not refer to recoverability under attack; that is, techniques wherein the function used can only be reversed using additional information, such as an encryption key, or knowledge of injected traffic within the dataset, are not considered to be recoverable.

Countability compares the dimension of the anonymized space (N) to the dimension of the real space (M), and denotes how the count of unique values is preserved by the anonymization function. If the anonymized space is smaller than the real space, then the function is said to generalize the input, mapping more than one input point to each anonymous value (e.g., as with aggregation). By definition, generalization is not recoverable.

If the dimensions of the anonymized and real spaces are the same, such that the count of unique values is preserved, then the function is said to be a direct substitution function. If the dimension of the anonymized space is larger, such that each real value maps to a set of anonymized values, then the function is said to be a set substitution function. Note that with set substitution functions, the sets of anonymized values are not necessarily disjoint. Either direct or set substitution functions are said to be one-way if there exists no non-brute force method for recovering the real data point from an anonymized one in isolation (i.e., if the only way to recover the data point is to attack the anonymized data set as a whole, e.g., through fingerprinting or data injection).

This classification is summarized in the table below.

Recoverability / Countability	Recoverable	Non-recoverable
N < M	N.A.	Generalization
N = M	Direct Substitution	One-way Direct Substitution
N > M	Set Substitution	One-way Set Substitution

4. Anonymization of IP Flow Data

In anonymizing IP flow data as treated by this document, the goal is generally two-way address untraceability: to remove the ability to assert that endpoint X contacted endpoint Y at time T. Address untraceability is important as IP addresses are the most suitable field in IP flow records to identify real-world entities. Each IP address is associated with an interface on a network host and can

potentially be identified with a single user. Additionally, IP addresses are structured identifiers; that is, partial IP address prefixes may be used to identify networks just as full IP addresses identify hosts. This leads IP flow data anonymization to be concerned first and foremost with IP address anonymization.

Any form of aggregation that combines flows from multiple endpoints into a single record (e.g., aggregation by subnetwork, aggregation removing addressing completely) may also provide address untraceability; however, anonymization by aggregation is out of scope for this document. Additionally, of potential interest in this problem space but out of scope are anonymization techniques that are applied over multiple fields or multiple records in a way that introduces dependencies among anonymized fields or records. This document is concerned solely with anonymization techniques applied at the resolution of single fields within a flow record.

Even so, attacks against these anonymization techniques use entire flows and relationships between hosts and flows within a given dataset. Therefore, fields that may not necessarily be identifying by themselves may be anonymized in order to increase the anonymity of the dataset as a whole.

Due to the restricted semantics of IP flow data, there is a relatively limited set of specific anonymization techniques available on flow data, though each falls into the broad categories discussed in the previous section. Each type of field that may commonly appear in a flow record may have its own applicable specific techniques.

As with IP addresses, Media Access Control (MAC) addresses uniquely identify devices on the network; while they are not often available in traffic data collected at Layer 3, and cannot be used to locate devices within the network, some traces may contain sub-IP data including MAC address data. Hardware addresses may be mappable to device serial numbers, and to the entities or individuals who purchased the devices, when combined with external databases. MAC addresses are also often used in constructing IPv6 addresses (see Section 2.5.1 of [RFC4291]) and as such may be used to reconstruct the low-order bits of anonymized IPv6 addresses in certain circumstances. Therefore, MAC address anonymization is also important.

Port numbers identify abstract entities (applications) as opposed to real-world entities, but they can be used to classify hosts and user behavior. Passive port fingerprinting, both of well-known and ephemeral ports, can be used to determine the operating system

running on a host. Relative data volumes by port can also be used to determine the host's function (workstation, web server, etc.); this information can be used to identify hosts and users.

While not identifiers in and of themselves, timestamps and counters can reveal the behavior of the hosts and users on a network. Any given network activity is recognizable by a pattern of relative time differences and data volumes in the associated sequence of flows, even without host address information. Therefore, they can be used to identify hosts and users. Timestamps and counters are also vulnerable to traffic injection attacks, where traffic with a known pattern is injected into a network under measurement, and this pattern is later identified in the anonymized dataset.

The simplest and most extreme form of anonymization, which can be applied to any field of a flow record, is black-marker anonymization, or complete deletion of a given field. Note that black-marker anonymization is equivalent to simply not exporting the field(s) in question.

While black-marker anonymization completely protects the data in the deleted fields from the risk of disclosure, it also reduces the utility of the anonymized dataset as a whole. Techniques that retain some information while reducing (though not eliminating) the disclosure risk will be extensively discussed in the following sections; note that the techniques specifically applicable to IP addresses, timestamps, ports, and counters will be discussed in separate sections.

4.1. IP Address Anonymization

Since IP addresses are the most common identifiers within flow data that can be used to directly identify a person, organization, or host, most of the work on flow and trace data anonymization has gone into IP address anonymization techniques. Indeed, the aim of most attacks against anonymization is to recover the map from anonymized IP addresses to original IP addresses thereby identifying the identified hosts. Therefore, there is a wide range of IP address anonymization schemes that fit into the following categories.

Scheme	Action
Truncation	Generalization
Reverse Truncation	Generalization
Permutation	Direct Substitution
Prefix-preserving Pseudonymization	Direct Substitution

4.1.1. Truncation

Truncation removes "n" of the least significant bits from an IP address, replacing them with zeroes. In effect, it replaces a host address with a network address for some fixed netblock; for IPv4 addresses, 8-bit truncation corresponds to replacement with a /24 network address. Truncation is a non-reversible generalization scheme. Note that while truncation is effective for making hosts non-identifiable, it preserves information that can be used to identify an organization, a geographic region, a country, or a continent.

Truncation to an address length of 0 is equivalent to black-marker anonymization. Complete removal of IP address information is only recommended for analysis tasks that have no need to separate flow data by host or network; e.g., as a first stage to per-application (port) or time-series total volume analyses.

4.1.2. Reverse Truncation

Reverse truncation removes "n" of the most significant bits from an IP address, replacing them with zeroes. Reverse truncation is a non-reversible generalization scheme. Reverse truncation is effective for making networks unidentifiable, partially or completely removing information that can be used to identify an organization, a geographic region, a country, or a continent (or Regional Internet Registry (RIR) region of responsibility). However, it may cause ambiguity when applied to data collected from more than one network, since it treats all the hosts with the same address on different networks as if they are the same host. It is not particularly useful when publishing data where the network of origin is known or can be easily guessed by virtue of the identity of the publisher.

Like truncation, reverse truncation to an address length of 0 is equivalent to black-marker anonymization.

4.1.3. Permutation

Permutation is a direct substitution technique, replacing each IP address with an address selected from the set of possible IP addresses, such that each anonymized address represents a unique original address. The selection function is often random, though it is not necessarily so. Permutation does not preserve any structural information about a network, but it does preserve the unique count of IP addresses. Any application that requires more structure than host-uniqueness will not be able to use permuted IP addresses.

There are many variations of permutation functions, each of which has trade-offs in performance, security, and guarantees of non-collision; evaluating these trade-offs is implementation independent. However, in general, permutation functions applied to anonymization SHOULD be difficult to reverse without knowing the parameters (e.g., a secret key for Hashed Message Authentication Code (HMAC). Given the relatively small space of IPv4 addresses in particular, hash functions applied without additional parameters could be reversed through brute force if the hash function is known, and SHOULD NOT be used as permutation functions. Permutation functions may guarantee non-collision (i.e., that each anonymized address represents a unique original address), but need not; however, the probability of collision SHOULD be low. Nevertheless, we treat even permutations with low but nonzero collision probability as a direct substitution. Beyond these guidelines, recommendations for specific permutation functions are out of scope for this document.

4.1.4. Prefix-Preserving Pseudonymization

Prefix-preserving pseudonymization is a direct substitution technique, like permutation but further restricted such that the structure of subnets is preserved at each level while anonymizing IP addresses. If two real IP addresses match on a prefix of "n" bits, the two anonymized IP addresses will match on a prefix of "n" bits as well. This is useful when relationships among networks must be preserved for a given analysis task, but introduces structure into the anonymized data that can be exploited in attacks against the anonymization technique.

Scanning in Internet background traffic can cause particular problems with this technique: if a scanner uses a predictable and known sequence of addresses, this information can be used to reverse the substitution. The low-order portion of the address can be left unanonymized as a partial defense against this attack.

4.2. MAC Address Anonymization

Flow data containing sub-IP information can also contain identifying information in the form of the hardware (MAC) address. While MAC address information cannot be used to locate a node within a network, it can be used to directly and uniquely identify a specific device. Vendors or organizations within the supply chain may then have the information necessary to identify the entity or individual that purchased the device.

MAC address information is not as structured as IP address information. EUI-48 and EUI-64 MAC addresses contain an Organizational Unique Identifier (OUI) in the three most significant

bytes of the address; this OUI additionally contains bits noting whether the address is locally or globally administered. Beyond this, there is no standard relationship among the OUIs assigned to a given vendor.

Note that MAC address information also appears within IPv6 addresses as the EAP-64 address, or EAP-48 address encoded as an EAP-64 address, is used as the least significant 64 bits of the IPv6 address in the case of link-local addressing or stateless autoconfiguration; the considerations and techniques in this section may then apply to such IPv6 addresses as well.

Scheme	Action
Truncation	Generalization
Reverse Truncation	Generalization
Permutation	Direct Substitution
Structured Pseudonymization	Direct Substitution

4.2.1. Truncation

Truncation removes "n" of the least significant bits from a MAC address, replacing them with zeroes. In effect, it retains bits of OUI, which identifies the manufacturer, while removing the least significant bits identifying the particular device. Truncation of 24 bits of an EAP-48 or 40 bits of an EAP-64 address zeroes out the device identifier while retaining the OUI.

Truncation is effective for making device manufacturers partially or completely identifiable within a dataset while deleting unique host identifiers; this can be used to retain and aggregate MAC-layer behavior by vendor.

Truncation to an address length of 0 is equivalent to black-marker anonymization.

4.2.2. Reverse Truncation

Reverse truncation removes "n" of the most significant bits from a MAC address, replacing them with zeroes. Reverse truncation is a non-reversible generalization scheme. This has the effect of removing bits of the OUI, which identify manufacturers, before removing the least significant bits. Reverse truncation of 24 bits zeroes out the OUI.

Reverse truncation is effective for making device manufacturers partially or completely unidentifiable within a dataset. However, it may cause ambiguity by introducing the possibility of truncated MAC address collision. Also, note that the utility of removing manufacturer information is not particularly well covered by the literature.

Reverse truncation to an address length of 0 is equivalent to black-marker anonymization.

4.2.3. Permutation

Permutation is a direct substitution technique, replacing each MAC address with an address selected from the set of possible MAC addresses, such that each anonymized address represents a unique original address. The selection function is often random, though it is not necessarily so. Permutation does not preserve any structural information about a network, but it does preserve the unique count of devices on the network. Any application that requires more structure than host-uniqueness will not be able to use permuted MAC addresses.

There are many variations of permutation functions, each of which has trade-offs in performance, security, and guarantees of non-collision; evaluating these trade-offs is implementation independent. However, in general, permutation functions applied to anonymization SHOULD be difficult to reverse without knowing the parameters (e.g., a secret key for HMAC). While the EAP-48 space is larger than the IPv4 address space, hash functions applied without additional parameters could be reversed through brute force if the hash function is known, and SHOULD NOT be used as permutation functions. Permutation functions may guarantee non-collision (i.e., that each anonymized address represents a unique original address), but need not; however, the probability of collision SHOULD be low. Nevertheless, we treat even permutations with low but nonzero collision probability as a direct substitution. Beyond these guidelines, recommendations for specific permutation functions are out of scope for this document.

4.2.4. Structured Pseudonymization

Structured pseudonymization for MAC addresses is a direct substitution technique, like permutation, but restricted such that the OUI (the most significant three bytes) is permuted separately from the node identifier, the remainder. This is useful when the uniqueness of OUIs must be preserved for a given analysis task, but introduces structure into the anonymized data that can be exploited in attacks against the anonymization technique.

4.3. Timestamp Anonymization

The particular time at which a flow began or ended is not particularly identifiable information, but it can be used as part of attacks against other anonymization techniques or for user profiling, e.g., as in [Mur07]. Timestamps can be used in traffic injection attacks, which use known information about a set of traffic generated or otherwise known by an attacker to recover mappings of other anonymized fields, as well as to identify certain activity by response delay and size fingerprinting, which compares response sizes and inter-flow times in anonymized data to known values. Note that these attacks have been shown to be relatively robust against timestamp anonymization techniques (see [Bur10]), so the techniques presented in this section are relatively weak and should be used with care.

Scheme	Action
Precision Degradation	Generalization
Enumeration	Direct or Set Substitution
Random Shifts	Direct Substitution

4.3.1. Precision Degradation

Precision Degradation is a generalization technique that removes the most precise components of a timestamp, accounting for all events occurring in each given interval (e.g., one millisecond for millisecond level degradation) as simultaneous. This has the effect of potentially collapsing many timestamps into one. With this technique, time precision is reduced and sequencing may be lost, but the information regarding at which time the event occurred is preserved. The anonymized data may not be generally useful for applications that require strict sequencing of flows.

Note that flow meters with low time precision (e.g., second precision, or millisecond precision on high-capacity networks) perform the equivalent of precision degradation anonymization by their design.

Also, note that degradation to a very low precision (e.g., on the order of minutes, hours, or days) is commonly used in analyses operating on time-series aggregated data, and may also be described as binning; though the time scales are longer and applicability more restricted, in principle, this is the same operation.

Precision degradation to infinitely low precision is equivalent to black-marker anonymization. Removal of timestamp information is only recommended for analysis tasks that have no need to separate flows in time, for example, for counting total volumes or unique occurrences of other flow keys in an entire dataset.

4.3.2. Enumeration

Enumeration is a substitution function that retains the chronological order in which events occurred while eliminating time information. Timestamps are substituted by equidistant timestamps (or numbers) starting from a randomly chosen start value. The resulting data is useful for applications requiring strict sequencing, but not for those requiring good timing information (e.g., delay- or jitter-measurement for quality-of-service (QoS) applications or service-level agreement (SLA) validation).

Note that enumeration is functionally equivalent to precision degradation in any environment into which traffic can be regularly injected to serve as a clock at the precision of the frequency of the injected flows.

4.3.3. Random Shifts

Random time shifts add a random offset to every timestamp within a dataset. Therefore, this reversible substitution technique retains duration and inter-event interval information as well as the chronological order of flows. Random time shifts are quite weak and relatively easy to reverse in the presence of external knowledge about traffic on the measured network.

4.4. Counter Anonymization

Counters (such as packet and octet volumes per flow) are subject to fingerprinting and injection attacks against anonymization or for user profiling as timestamps are. Data sets with anonymized counters are useful only for analysis tasks for which relative or imprecise magnitudes of activity are useful. Counter information can also be completely removed, but this is only recommended for analysis tasks that have no need to evaluate the removed counter, for example, for counting only unique occurrences of other flow keys.

Scheme	Action
Precision Degradation	Generalization
Binning	Generalization
Random noise addition	Direct or Set Substitution

4.4.1. Precision Degradation

As with precision degradation in timestamps, precision degradation of counters removes lower-order bits of the counters, treating all the counters in a given range as having the same value. Depending on the precision reduction, this loses information about the relationships between sizes of similarly sized flows, but keeps relative magnitude information. Precision degradation to an infinitely low precision is equivalent to black-marker anonymization.

4.4.2. Binning

Binning can be seen as a special case of precision degradation; the operation is identical, except for in precision degradation the counter ranges are uniform, and in binning, they need not be. For example, consider separating unopened TCP connections from potentially opened TCP connections. Here, packet counters per flow would be binned into two bins, one for 1-2 packet flows, and one for flows with 3 or more packets. Binning schemes are generally chosen to keep precisely the amount of information required in a counter for a given analysis task. Note that, also unlike precision degradation, the bin label need not be within the bin's range. Binning counters to a single bin is equivalent to black-marker anonymization.

4.4.3. Random Noise Addition

Random noise addition adds a random amount to a counter in each flow; this is used to keep relative magnitude information and minimize the disruption to size relationship information while avoiding fingerprinting attacks against anonymization. Note that there is no guarantee that random noise addition will maintain ranking order by a counter among members of a set. Random noise addition is particularly useful when the derived analysis data will not be presented in such a way as to require the lower-order bits of the counters.

4.5. Anonymization of Other Flow Fields

Other fields, particularly port numbers and protocol numbers, can be used to partially identify the applications that generated the traffic in a given flow trace. This information can be used in fingerprinting attacks, and may be of interest on its own (e.g., to reveal that a certain application with suspected vulnerabilities is running on a given network). These fields are generally anonymized using one of two techniques.

Scheme	Action
Binning	Generalization
Permutation	Direct Substitution

4.5.1. Binning

Binning is a generalization technique mapping a set of potentially non-uniform ranges into a set of arbitrarily labeled bins. Common bin arrangements depend on the field type and the analysis application. For example, an IP protocol bin arrangement may preserve 1, 6, and 17 for ICMP, UDP, and TCP traffic, and bin all other protocols into a single bin, to mitigate the use of uncommon protocols in fingerprinting attacks. Another example arrangement may bin source and destination ports into low (0-1023) and high (1024-65535) bins in order to tell service from ephemeral ports without identifying individual applications.

Binning other flow key fields to a single bin is equivalent to black-marker anonymization. Removal of other flow key information is only recommended for analysis tasks that have no need to differentiate flows on the removed keys, for example, for total traffic counts or unique counts of other flow keys.

4.5.2. Permutation

Permutation is a direct substitution technique, replacing each value with an value selected from the set of possible range, such that each anonymized value represents a unique original value. This is used to preserve the count of unique values without preserving information about, or the ordering of, the values themselves.

While permutation ideally guarantees that each anonymized value represents a unique original value, such may require significant state in the Intermediate Anonymization Process. Therefore, permutation may be implemented by hashing for performance reasons,

with hash functions that may have relatively small collision probabilities. Such techniques are still essentially direct substitution techniques, despite the nonzero error probability.

5. Parameters for the Description of Anonymization Techniques

This section details the abstract parameters used to describe the anonymization techniques examined in the previous section, on a per-parameter basis. These parameters and their export safety inform the design of the IPFIX anonymization metadata export specified in the following section.

5.1. Stability

A stable anonymization will always map a given value in the real space to a given value in the anonymized space, while an unstable anonymization will change this mapping over time; a completely unstable anonymization is essentially indistinguishable from black-marker anonymization. Any given anonymization technique may be applied with a varying range of stability. Stability is important for assessing the comparability of anonymized information in different datasets, or in the same dataset over different time periods. In practice, an anonymization may also be stable for every dataset published by a particular producer to a particular consumer, stable for a stated time period within a dataset or across datasets, or stable only for a single dataset.

If no information about stability is available, users of anonymized data MAY assume that the techniques used are stable across the entire dataset, but unstable across datasets. Note that stability presents a risk-utility trade-off, as completely stable anonymization can be used for longer-term trend analysis tasks but also presents more risk of attack given the stable mapping. Information about the stability of a mapping SHOULD be exported along with the anonymized data.

5.2. Truncation Length

Truncation and precision degradation are described by the truncation length or the amount of data still remaining in the anonymized field after anonymization.

Truncation length can generally be inferred from a given dataset, and need not be specially exported or protected. For bit-level truncation, the truncated bits are generally inferable by the least significant bit set for an instance of an Information Element described by a given Template (or the most significant bit set, in the case of reverse truncation). For precision degradation, the truncation is inferable from the maximum precision given. Note that

while this inference method is generally applicable, it is data dependent: there is no guarantee that it will recover the exact truncation length used to prepare the data.

In the special case of IP address export with variable (per-record) truncation, the truncation MAY be expressed by exporting the prefix length alongside the address.

5.3. Bin Map

Binning is described by the specification of a bin mapping function. This function can be generally expressed in terms of an associative array that maps each point in the original space to a bin, although from an implementation standpoint most bin functions are much simpler and more efficient.

Since the bin map for a bin mapping function is in essence the bin mapping key, and can be used to partially deanonymize binned data, depending on the degree of generalization, information about the bin mapping function SHOULD NOT be exported.

5.4. Permutation

Like binning, permutation is described by the specification of a permutation function. In the general case, this can be expressed in terms of an associative array that maps each point in the original space to a point in the anonymized space. Unlike binning, each point in the anonymized space corresponds to a single, unique point in the original space.

Since the parameters of the permutation function are in essence key-like (indeed, for cryptographic permutation functions, they are the keys themselves), information about the permutation function or its parameters SHOULD NOT be exported.

5.5. Shift Amount

Shifting requires an amount by which to shift each value. Since the shift amount is the only key to a shift function, and can be used to trivially deanonymize data protected by shifting, information about the shift amount SHOULD NOT be exported.

6. Anonymization Export Support in IPFIX

Anonymized data exported via IPFIX SHOULD be annotated with anonymization metadata, which details which fields described by which Templates are anonymized, and provides appropriate information on the anonymization techniques used. This metadata SHOULD be exported in

Data Records described by the recommended Options Templates described in this section; these Options Templates use the additional Information Elements described in the following subsection.

Note that fields anonymized using the black-marker (removal) technique do not require any special metadata support: black-marker anonymized fields SHOULD NOT be exported at all, by omitting the corresponding Information Elements from Template describing the Data Set. In the case where application requirements dictate that a black-marker anonymized field must remain in a Template, then an Exporting Process MAY export black-marker anonymized fields with their native length as all-zeros, but only in cases where enough contextual information exists within the record to differentiate a black-marker anonymized field exported in this way from a real zero value.

6.1. Anonymization Records and the Anonymization Options Template

The Anonymization Options Template describes Anonymization Records, which allow anonymization metadata to be exported inline over IPFIX or stored in an IPFIX File, by binding information about anonymization techniques to Information Elements within defined Templates or Options Templates. IPFIX Exporting Processes SHOULD export anonymization records for any Template describing exported anonymized Data Records; IPFIX Collecting Processes and processes downstream from them MAY use anonymization records to treat anonymized data differently depending on the applied technique.

Anonymization Records contain ancillary information bound to a Template, so many of the considerations for Templates apply to Anonymization Records as well. First, reliability is important: an Exporting Process SHOULD export Anonymization Records after the Templates they describe have been exported, and SHOULD export anonymization records reliably if supported by the underlying transport (i.e., without partial reliability when using Stream Control Transmission Protocol (SCTP)).

Anonymization Records MUST be handled by Collecting Processes as scoped to the Template to which they apply within the Transport Session in which they are sent. When a Template is withdrawn via a Template Withdrawal Message or expires during a UDP transport session, the accompanying Anonymization Records are withdrawn or expire as well and do not apply to subsequent Templates with the same Template ID within the Session unless re-exported.

The Stability Class within the anonymizationFlags IE can be used to declare that a given anonymization technique’s mapping will remain stable across multiple sessions, but this does not mean that

anonymization technique information given in the Anonymization Records themselves persist across Sessions. Each new Transport Session MUST contain new Anonymization Records for each Template describing anonymized Data Sets.

SCTP per-stream export [IPFIX-PERSTREAM] may be used to ease management of Anonymization Records if appropriate for the application.

The fields of the Anonymization Options Template are as follows:

IE	Description
templateId [scope]	The Template ID of the Template or Options Template containing the Information Element described by this anonymization record. This Information Element MUST be defined as a Scope Field.
informationElementId [scope]	The Information Element identifier of the Information Element described by this anonymization record. This Information Element MUST be defined as a Scope Field. Exporting Processes MUST clear then Enterprise bit of the informationElementId and Collecting Processes SHOULD ignore it; information about enterprise-specific Information Elements is exported via the privateEnterpriseNumber Information Element.
privateEnterpriseNumber [scope] [optional]	The Private Enterprise Number of the enterprise-specific Information Element described by this anonymization record. This Information Element MUST be defined as a Scope Field if present. A privateEnterpriseNumber of 0 signifies that the Information Element is IANA-registered.
informationElementIndex [scope] [optional]	The Information Element index of the instance of the Information Element described by this anonymization record identified by the informationElementId within the Template. Optional; need only be present when describing Templates that have multiple instances of the same Information Element. This

		Information Element MUST be defined as a Scope Field if present. This Information Element is defined in Section 6.2.
	anonymizationFlags	Flags describing the mapping stability and specialized modifications to the Anonymization Technique in use. SHOULD be present. This Information Element is defined in Section 6.2.3.
	anonymizationTechnique	The technique used to anonymize the data. MUST be present. This Information Element is defined in Section 6.2.2.

6.2. Recommended Information Elements for Anonymization Metadata

6.2.1. informationElementIndex

Description: A zero-based index of an Information Element referenced by informationElementId within a Template referenced by templateId; used to disambiguate scope for templates containing multiple identical Information Elements.

Abstract Data Type: unsigned16

Data Type Semantics: identifier

ElementId: 287

Status: Current

6.2.2. anonymizationTechnique

Description: A description of the anonymization technique applied to a referenced Information Element within a referenced Template. Each technique may be applicable only to certain Information Elements and recommended only for certain Information Elements; these restrictions are noted in the table below.

Value	Description	Applicable to	Recommended for
0	Undefined: the Exporting Process makes no representation as to whether or not the defined field is anonymized. While the Collecting Process MAY assume that the field is not anonymized, it is not guaranteed not to be. This is the default anonymization technique.	all	all
1	None: the values exported are real.	all	all
2	Precision Degradation/Truncation: the values exported are anonymized using simple precision degradation or truncation. The new precision or number of truncated bits is implicit in the exported data and can be deduced by the Collecting Process.	all	all
3	Binning: the values exported are anonymized into bins.	all	all
4	Enumeration: the values exported are anonymized by enumeration.	all	timestamps
5	Permutation: the values exported are anonymized by permutation.	all	identifiers
6	Structured Permutation: the values exported are anonymized by permutation, preserving bit-level structure as appropriate; this represents prefix-preserving IP address anonymization or	addresses	

	structured MAC address anonymization.			
7	Reverse Truncation: the values exported are anonymized using reverse truncation. The number of truncated bits is implicit in the exported data, and can be deduced by the Collecting Process.	addresses		
8	Noise: the values exported are anonymized by adding random noise to each value.	non-identifiers	counters	
9	Offset: the values exported are anonymized by adding a single offset to all values.	all	timestamps	

Abstract Data Type: unsigned16

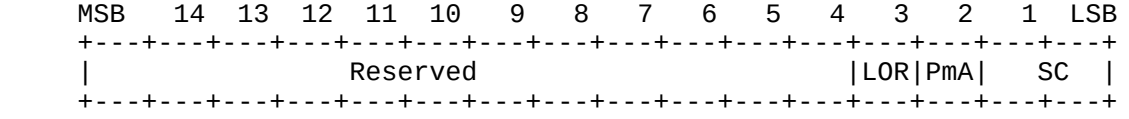
Data Type Semantics: identifier

ElementId: 286

Status: Current

6.2.3. anonymizationFlags

Description: A flag word describing specialized modifications to the anonymization policy in effect for the anonymization technique applied to a referenced Information Element within a referenced Template. When flags are clear (0), the normal policy (as described by anonymizationTechnique) applies without modification.



anonymizationFlags IE

bit(s) (LSB = 0)	name	description
0-1	SC	Stability Class: see the Stability Class table below, and Section 5.1.
2	PmA	Perimeter Anonymization: when set (1), source Information Elements as described in [RFC5103] are interpreted as external addresses, and destination Information Elements as described in [RFC5103] are interpreted as internal addresses, for the purposes of associating anonymizationTechnique to Information Elements only; see Section 7.2.2 for details. This bit MUST NOT be set when associated with a non-endpoint (i.e., source or destination) Information Element. SHOULD be consistent within a record (i.e., if a source Information Element has this flag set, the corresponding destination element SHOULD have this flag set, and vice versa.)
3	LOR	Low-Order Unchanged: when set (1), the low-order bits of the anonymized Information Element contain real data. This modification is intended for the anonymization of network-level addresses while leaving host-level addresses intact in order to preserve host level-structure, which could otherwise be used to reverse anonymization. MUST NOT be set when associated with a truncation-based anonymizationTechnique.
4-15	Reserved	Reserved for future use: SHOULD be cleared (0) by the Exporting Process and MUST be ignored by the Collecting Process.

The Stability Class portion of this flags word describes the stability class of the anonymization technique applied to a referenced Information Element within a referenced Template. Stability classes refer to the stability of the parameters of the anonymization technique, and therefore the comparability of the mapping between the real and anonymized values over time. This determines which anonymized datasets may be compared with each other. Values are as follows:

Bit 1	Bit 0	Description
0	0	Undefined: the Exporting Process makes no representation as to how stable the mapping is, or over what time period values of this field will remain comparable; while the Collecting Process MAY assume Session level stability, Session level stability is not guaranteed. Processes SHOULD assume this is the case in the absence of stability class information; this is the default stability class.
0	1	Session: the Exporting Process will ensure that the parameters of the anonymization technique are stable during the Transport Session. All the values of the described Information Element for each Record described by the referenced Template within the Transport Session are comparable. The Exporting Process SHOULD endeavor to ensure at least this stability class.
1	0	Exporter-Collector Pair: the Exporting Process will ensure that the parameters of the anonymization technique are stable across Transport Sessions over time with the given Collecting Process, but may use different parameters for different Collecting Processes. Data exported to different Collecting Processes are not comparable.
1	1	Stable: the Exporting Process will ensure that the parameters of the anonymization technique are stable across Transport Sessions over time, regardless of the Collecting Process to which it is sent.

Abstract Data Type: unsigned16

Data Type Semantics: flags

ElementId: 285

Status: Current

7. Applying Anonymization Techniques to IPFIX Export and Storage

When exporting or storing anonymized flow data using IPFIX, certain interactions between the IPFIX protocol and the anonymization techniques in use must be considered; these are treated in the subsections below.

7.1. Arrangement of Processes in IPFIX Anonymization

Anonymization may be applied to IPFIX data at three stages within the collection infrastructure: on initial export, at a mediator, or after collection, as shown in Figure 1. Each of these locations has specific considerations and applicability.

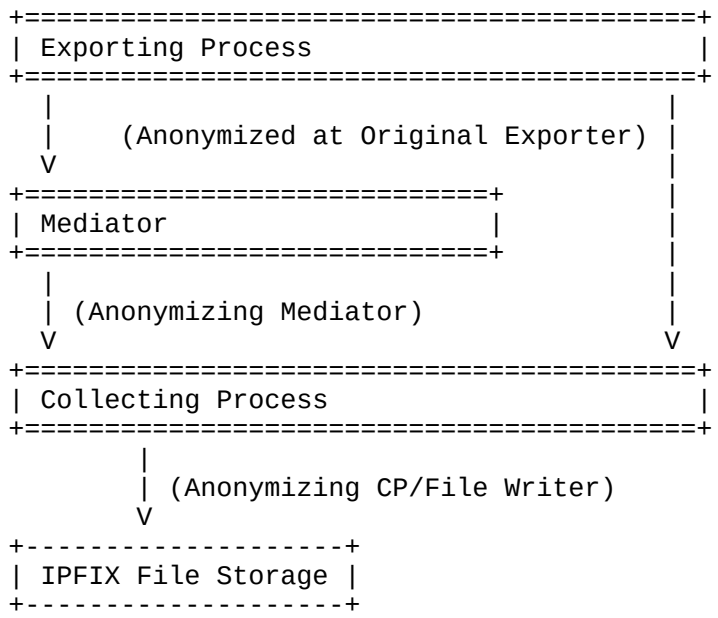


Figure 1: Potential Anonymization Locations

Anonymization is generally performed before the wider dissemination or repurposing of a dataset, e.g., adapting operational measurement data for research. Therefore, direct anonymization of flow data on initial export is only applicable in certain restricted circumstances: when the Exporting Process (EP) is "publishing" data to a Collecting Process (CP) directly, and the Exporting Process and Collecting Process are operated by different entities. Note that certain guidelines in Section 7.2.3 with respect to timestamp anonymization may not apply in this case, as the Collecting Process may be able to deduce certain timing information from the time at which each Message is received.

A much more flexible arrangement is to anonymize data within a Mediator [RFC6183]. Here, original data is sent to a Mediator, which performs the anonymization function and re-exports the anonymized data. Such a Mediator could be located at the administrative domain boundary of the initial Exporting Process operator, exporting

anonymized data to other consumers outside the organization. In this case, the original Exporter SHOULD use TLS [RFC5246] as specified in [RFC5101] to secure the channel to the Mediator, and the Mediator should follow the guidelines in Section 7.2, to mitigate the risk of original data disclosure.

When data is to be published as an anonymized dataset in an IPFIX File [RFC5655], the anonymization may be done at the final Collecting Process before storage and dissemination, as well. In this case, the Collector should follow the guidelines in Section 7.2, especially as regards File-specific Options in Section 7.2.4

In each of these data flows, the anonymization of records is undertaken by an Intermediate Anonymization Process (IAP); the data flows into and out of this IAP are shown in Figure 2 below.

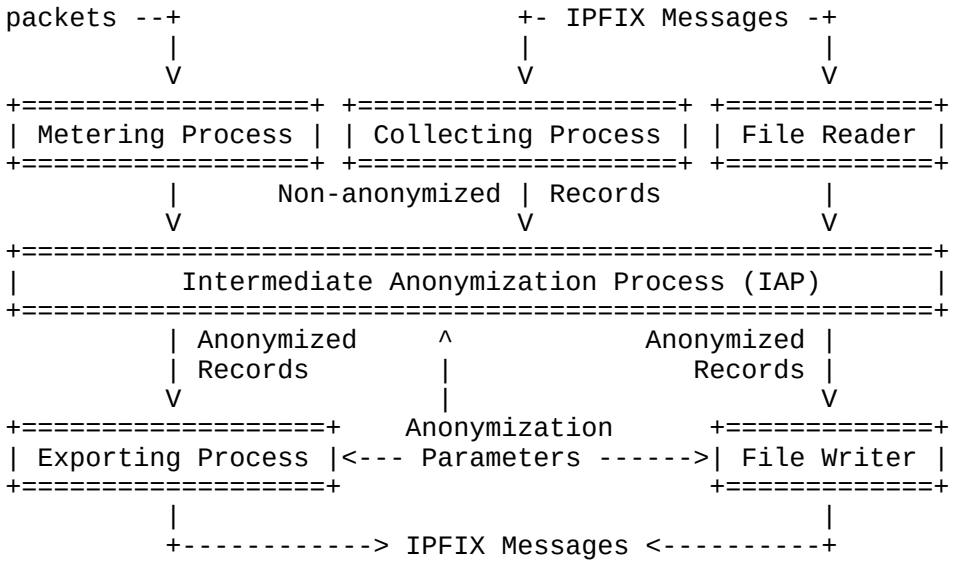


Figure 2: Data Flows through the Anonymization Process

Anonymization parameters must also be available to the Exporting Process and/or File Writer in order to ensure header data is also appropriately anonymized as in Section 7.2.3.

Following each of the data flows through the IAP, we describe five basic types of anonymization arrangements within this framework in Figure 3. In addition to the three arrangements described in detail above, anonymization can also be done at a collocated Metering

Process (MP) and File Writer (FW) (see Section 7.3.2 of [RFC5655]), or at a file manipulator, which combines a File Writer with a File Reader (FR) (see Section 7.3.7 of [RFC5655]).

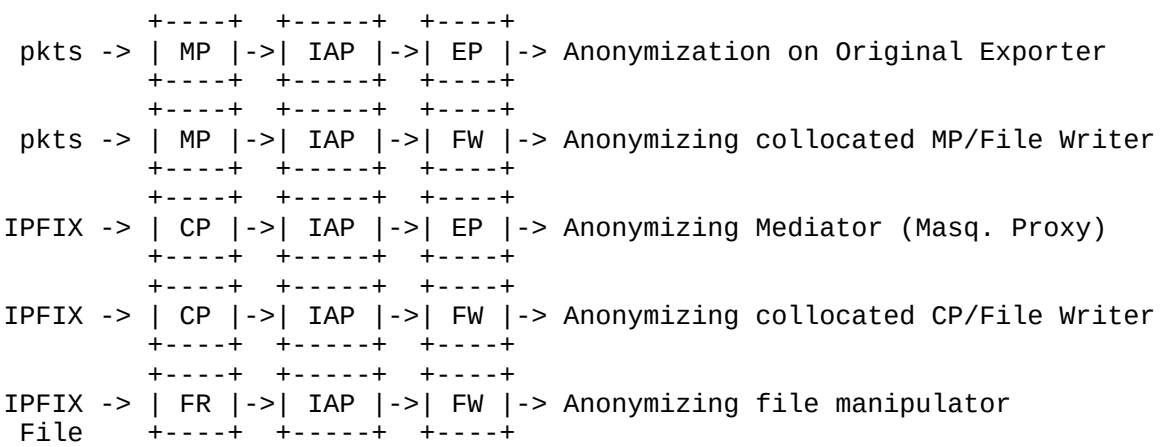


Figure 3: Possible Anonymization Arrangements in the IPFIX Architecture

Note that anonymization may occur at more than one location within a given collection infrastructure, to provide varying levels of anonymization, disclosure risk, or data utility for specific purposes.

7.2. IPFIX-Specific Anonymization Guidelines

In implementing and deploying the anonymization techniques described in this document, implementors should note that IPFIX already provides features that support anonymized data export, and use these where appropriate. Care must also be taken that data structures supporting the operation of the protocol itself do not leak data that could be used to reverse the anonymization applied to the flow data. Such data structures may appear in the header, or within the data stream itself, especially as options data. Each of these and their impact on specific anonymization techniques is noted in a separate subsection below.

7.2.1. Appropriate Use of Information Elements for Anonymized Data

Note, as in Section 6 above, that black-marker anonymized fields SHOULD NOT be exported at all; the absence of the field in a given Data Set is implicitly declared by not including the corresponding Information Element in the Template describing that Data Set.

When using precision degradation of timestamps, Exporting Processes SHOULD export timing information using Information Elements of an appropriate precision, as explained in Section 4.5 of [RFC5153]. For example, timestamps measured in millisecond-level precision and degraded to second-level precision should use flowStartSeconds and flowEndSeconds, not flowStartMilliseconds and flowEndMilliseconds.

When exporting anonymized data and anonymization metadata, Exporting Processes SHOULD ensure that the combination of Information Element and declared anonymization technique are compatible. Specifically, the applicable and recommended Information Element types and semantics for each technique are noted in the description of the anonymizationTechnique Information Element in Section 6.2.2. In this description, a timestamp is an Information Element with the data type dateTimeSeconds, dataTimeMilliseconds, dateTimeMicroseconds, or dateTimeNanoseconds; an address is an Information Element with the data type ipv4Address, ipv6Address, or macAddress; and an identifier is an Information Element with identifier data type semantics. Exporting Process MUST NOT export Anonymization Options records binding techniques to Information Elements to which they are not applicable, and SHOULD NOT export Anonymization Options records binding techniques to Information Elements for which they are not recommended.

7.2.2. Export of Perimeter-Based Anonymization Policies

Data collected from a single network may require different anonymization policies for addresses internal and external to the network. For example, internal addresses could be subject to simple permutation, while external addresses could be aggregated into networks by truncation. When exporting anonymized perimeter bidirectional flow (biflow) data as in Section 5.2 of [RFC5103], this arrangement may be easily represented by specifying one technique for source endpoint information (which represents the external endpoint in a perimeter biflow) and one technique for destination endpoint information (which represents the internal address in a perimeter biflow).

However, it can also be useful to represent perimeter-based anonymization policies with unidirectional flow (uniflow), or non-perimeter biflow data. In this case, the Perimeter Anonymization bit (bit 2) in the anonymizationFlags Information Element describing the anonymized address Information Elements can be set to change the meaning of "source" and "destination" of Information Elements to mean "external" and "internal" as with perimeter biflows, but only with respect to anonymization policies.

7.2.3. Anonymization of Header Data

Each IPFIX Message contains a Message Header; within this Message Header are contained two fields which may be used to break certain anonymization techniques: the Export Time, and the Observation Domain ID.

Export of IPFIX Messages containing anonymized timestamp data where the original Export Time Message header has some relationship to the anonymized timestamps SHOULD anonymize the Export Time header field so that the Export Time is consistent with the anonymized timestamp data. Otherwise, relationships between export and flow time could be used to partially or totally reverse timestamp anonymization. When anonymizing timestamps and the Export Time header field SHOULD avoid times too far in the past or future; while [RFC5101] does not make any allowance for Export Time error detection, it is sensible that Collecting Processes may interpret Messages with seemingly nonsensical Export Times as erroneous. Specific limits are implementation dependent, but this issue may cause interoperability issues when anonymizing the Export Time header field.

The similarity in size between an Observation Domain ID and an IPv4 address (32 bits) may lead to a temptation to use an IPv4 interface address on the Metering or Exporting Process as the Observation Domain ID. If this address bears some relation to the IP addresses in the flow data (e.g., shares a network prefix with internal addresses) and the IP addresses in the flow data are anonymized in a structure-preserving way, then the Observation Domain ID may be used to break the IP address anonymization. Use of an IPv4 interface address on the Metering or Exporting Process as the Observation Domain ID is NOT RECOMMENDED in this case.

7.2.4. Anonymization of Options Data

IPFIX uses the Options mechanism to export, among other things, metadata about exported flows and the flow collection infrastructure. As with the IPFIX Message Header, certain Options recommended in [RFC5101] and [RFC5655] containing flow timestamps and network addresses of Exporting and Collecting Processes may be used to break certain anonymization techniques. When using these Options along anonymized data export and storage, values within the Options that could be used to break the anonymization SHOULD themselves be anonymized or omitted.

The Exporting Process Reliability Statistics Options Template, recommended in [RFC5101], contains an Exporting Process ID field, which may be an exportingProcessIPv4Address Information Element or an exportingProcessIPv6Address Information Element. If the Exporting

Process address bears some relation to the IP addresses in the flow data (e.g., shares a network prefix with internal addresses) and the IP addresses in the flow data are anonymized in a structure-preserving way, then the Exporting Process address may be used to break the IP address anonymization. Exporting Processes exporting anonymized data in this situation SHOULD mitigate the risk of attack either by omitting Options described by the Exporting Process Reliability Statistics Options Template or by anonymizing the Exporting Process address using a similar technique to that used to anonymize the IP addresses in the exported data.

Similarly, the Export Session Details Options Template and Message Details Options Template specified for the IPFIX File Format [RFC5655] may contain the exportingProcessIPv4Address Information Element or the exportingProcessIPv6Address Information Element to identify an Exporting Process from which a flow record was received, and the collectingProcessIPv4Address Information Element or the collectingProcessIPv6Address Information Element to identify the Collecting Process which received it. If the Exporting Process or Collecting Process address bears some relation to the IP addresses in the dataset (e.g., shares a network prefix with internal addresses) and the IP addresses in the dataset are anonymized in a structure-preserving way, then the Exporting Process or Collecting Process address may be used to break the IP address anonymization. Since these Options Templates are primarily intended for storing IPFIX Transport Session data for auditing, replay, and testing purposes, it is NOT RECOMMENDED that storage of anonymized data include these Options Templates in order to mitigate the risk of attack.

The Message Details Options Template specified for the IPFIX File Format [RFC5655] also contains the collectionTimeMilliseconds Information Element. As with the Export Time Message Header field, if the exported dataset contains anonymized timestamp information, and the collectionTimeMilliseconds Information Element in a given Message has some relationship to the anonymized timestamp information, then this relationship can be exploited to reverse the timestamp anonymization. Since this Options Template is primarily intended for storing IPFIX Transport Session data for auditing, replay, and testing purposes, it is NOT RECOMMENDED that storage of anonymized data include this Options Template in order to mitigate the risk of attack.

Since the Time Window Options Template specified for the IPFIX File Format [RFC5655] refers to the timestamps within the dataset to provide partial table of contents information for an IPFIX File, Options described by this Template SHOULD be written using the anonymized timestamps instead of the original ones.

7.2.5. Special-Use Address Space Considerations

When anonymizing data for transport or storage using IPFIX containing anonymized IP addresses, and the analysis purpose permits doing so, it is RECOMMENDED to filter out or leave unanonymized data containing the special-use IPv4 addresses enumerated in [RFC5735] or the special-use IPv6 addresses enumerated in [RFC5156]. Data containing these addresses (e.g. 0.0.0.0 and 169.254.0.0/16 for link-local autoconfiguration in IPv4 space) are often associated with specific, well-known behavioral patterns. Detection of these patterns in anonymized data can lead to deanonymization of these special-use addresses, which increases the chance of a complete reversal of anonymization by an attacker, especially of prefix-preserving techniques.

7.2.6. Protecting Out-of-Band Configuration and Management Data

Special care should be taken when exporting or sharing anonymized data to avoid information leakage via the configuration or management planes of the IPFIX Device containing the Exporting Process or the File Writer. For example, adding noise to counters is useless if the receiver can deduce the values in the counters from Simple Network Management Protocol (SNMP) information, and concealing the network under test is similarly useless if such information is available in a configuration document. As the specifics of these concerns are largely implementation and deployment dependent, specific mitigation is out of scope for this document. The general ground rule is that information of similar type to that anonymized SHOULD NOT be made available to the receiver by any means, whether in the Data Records, in IPFIX protocol structures such as Message Headers, or out of band.

8. Examples

In this example, consider the export or storage of an anonymized IPv4 dataset from a single network described by a simple Template containing a timestamp in seconds, a five-tuple, and packet and octet counters. The Template describing each record in this Data Set is shown in Figure 4.

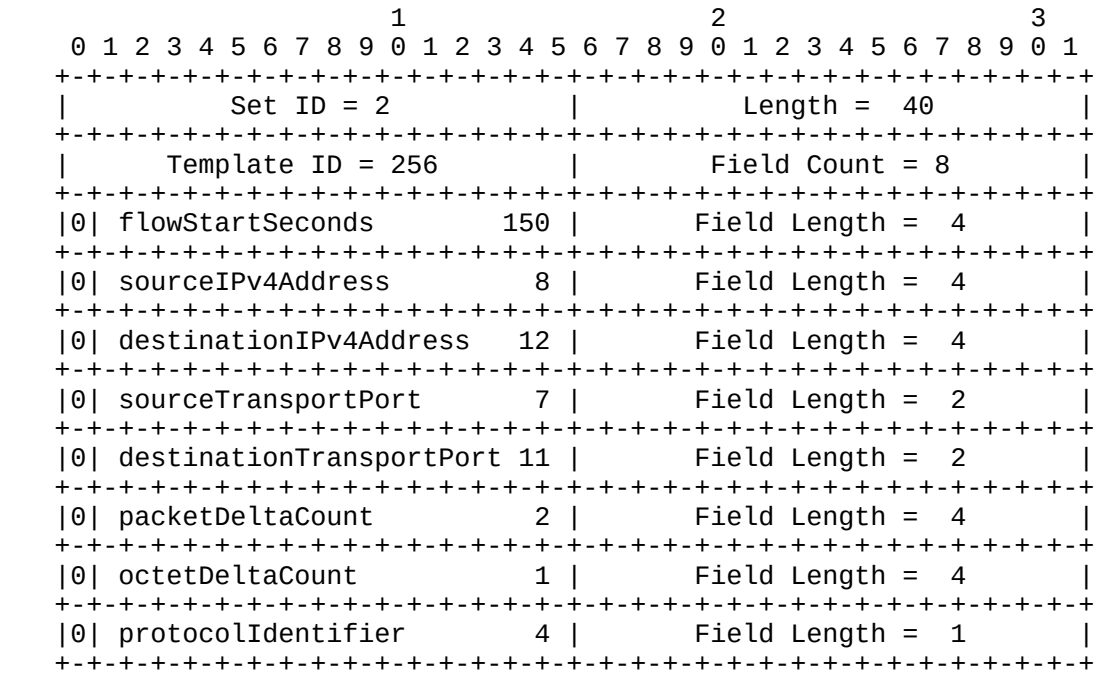


Figure 4: Example Flow Template

Suppose that this Data Set is anonymized according to the following policy:

- IP addresses within the network are protected by reverse truncation.
- IP addresses outside the network are protected by prefix-preserving anonymization.
- Octet counts are exported using degraded precision in order to provide minimal protection against fingerprinting attacks.
- All other fields are exported unanonymized.

In order to export Anonymization Records for this Template and policy, first, the Anonymization Options Template shown in Figure 5 is exported. For this example, the optional privateEnterpriseNumber and informationElementIndex Information Elements are omitted, because they are not used.

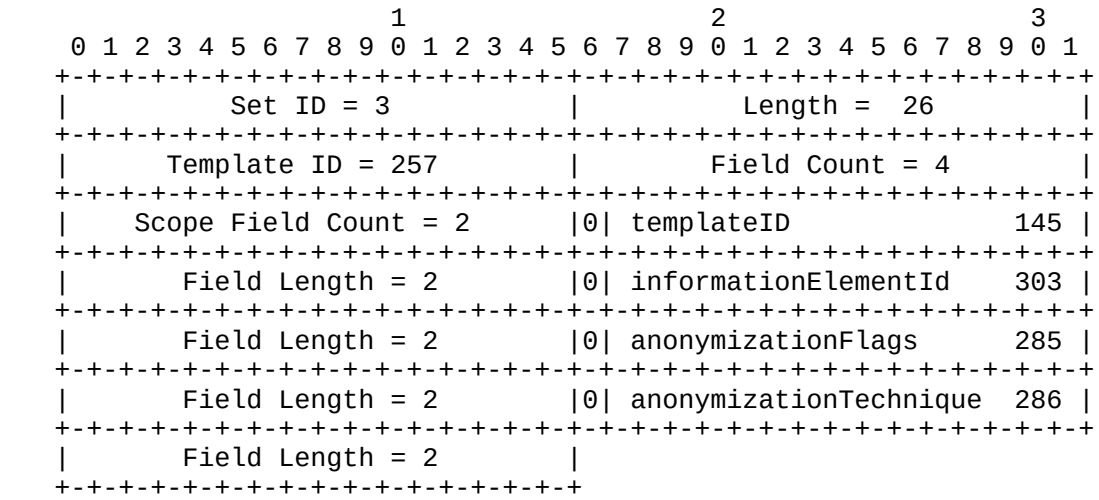


Figure 5: Example Anonymization Options Template

Following the Anonymization Options Template comes a Data Set containing Anonymization Records. This dataset has an entry for each Information Element Specifier in Template 256 describing the flow records. This Data Set is shown in Figure 6. Note that sourceIPv4Address and destinationIPv4Address have the Perimeter Anonymization (0x0004) flag set in anonymizationFlags, meaning that source address should be treated as network-external, and the destination address as network-internal.

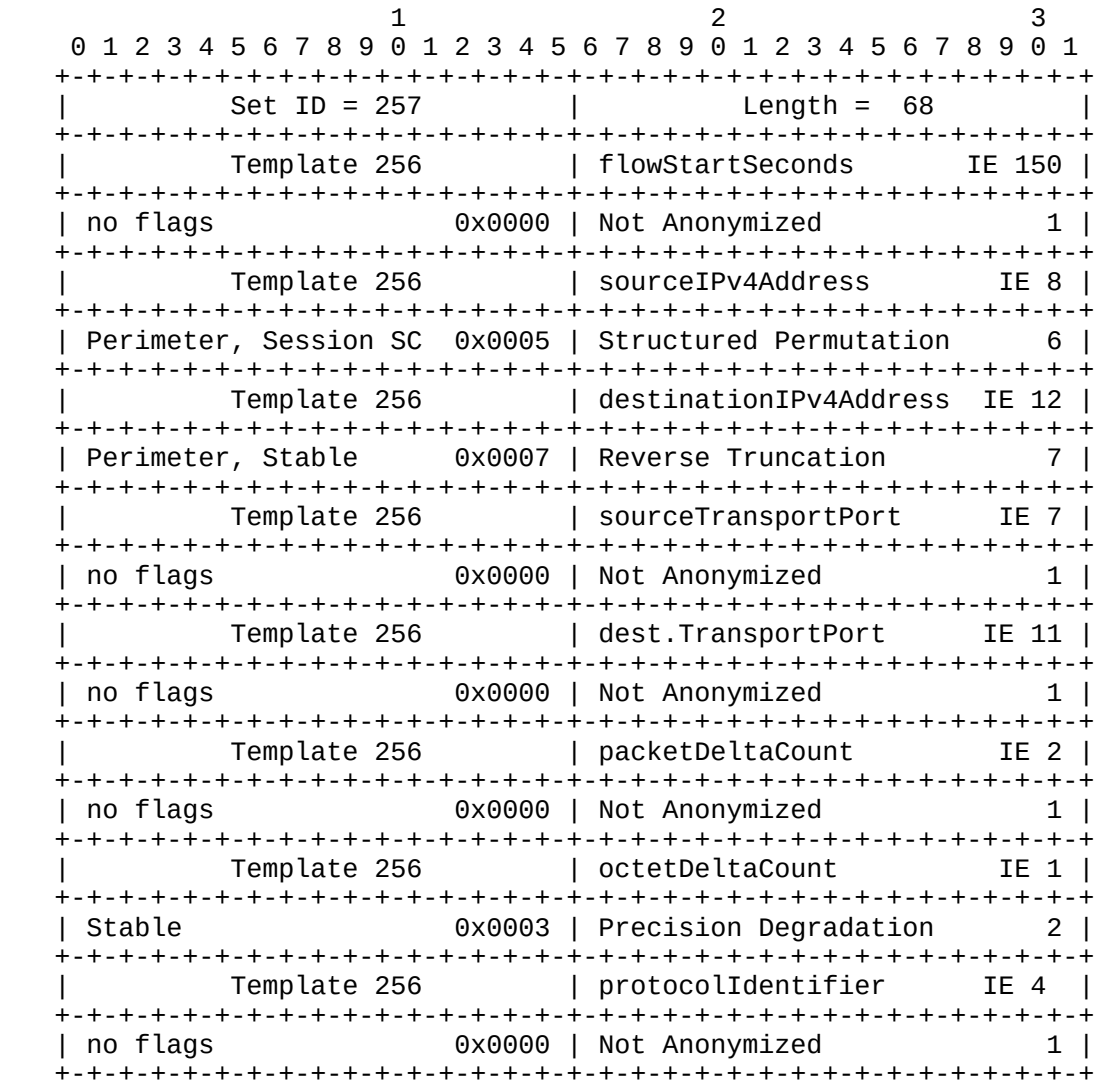


Figure 6: Example Anonymization Records

Following the Anonymization Records come the Data Sets containing the anonymized data, exported according to the Template in Figure 4. Bringing it all together, consider an IPFIX Message containing three real data records and the necessary templates to export them, shown in Figure 7. (Note that the scale of this message is 8-bytes per line, for compactness; lines of dots '. . . . . ' represent shifting of the example bit structure for clarity.)

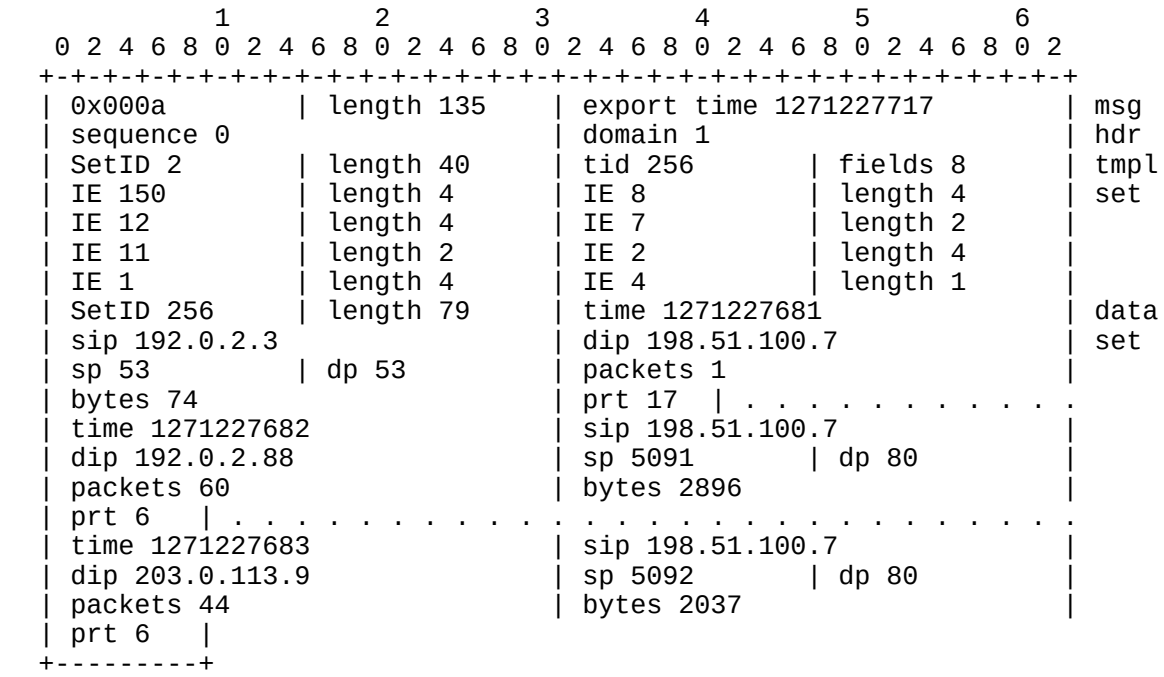


Figure 7: Example Real Message

The corresponding anonymized message is then shown in Figure 8. The Options Template Set describing Anonymization Records and the Anonymization Records themselves are added; IP addresses and byte counts are anonymized as declared.

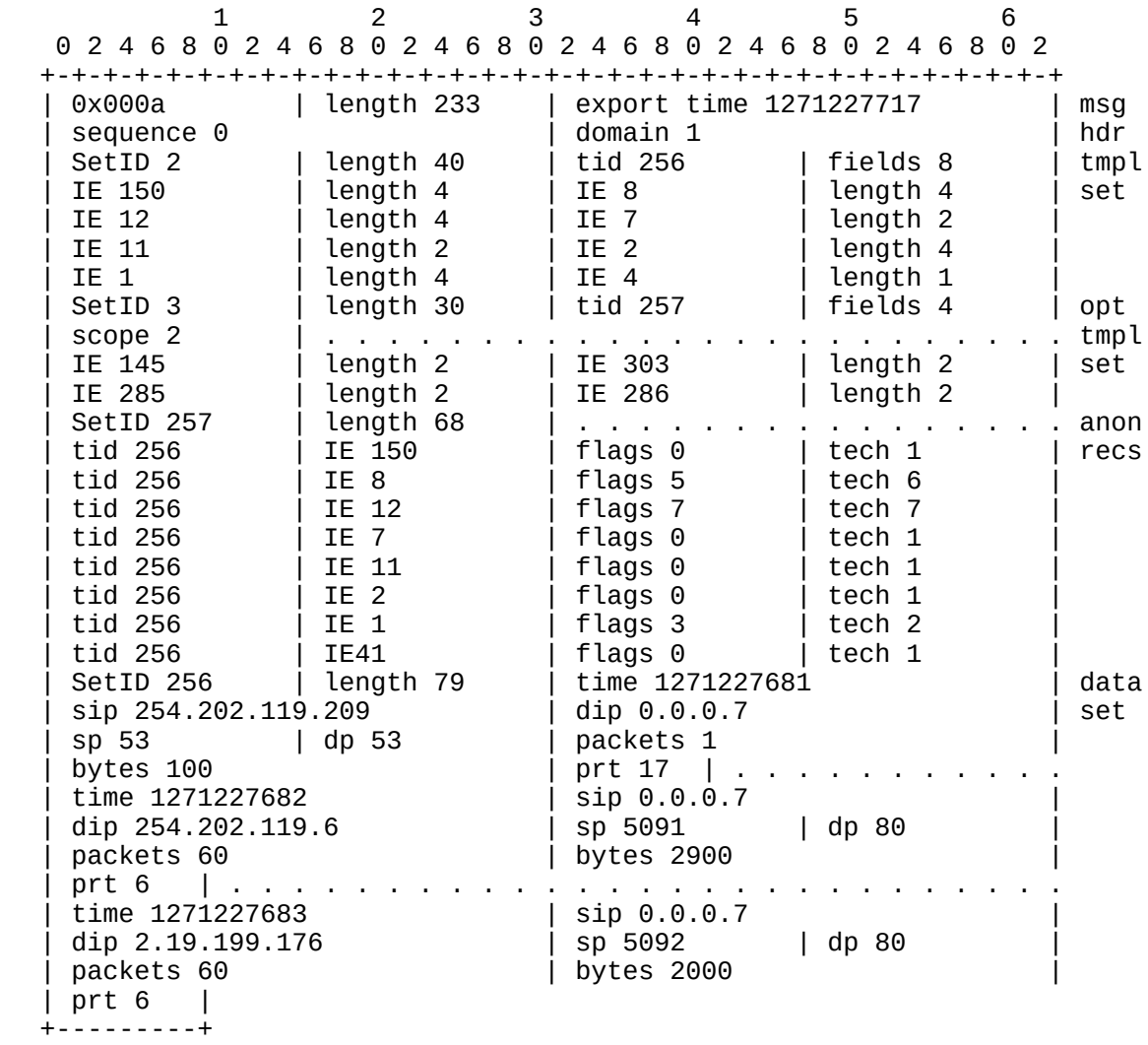


Figure 8: Corresponding Anonymized Message

9. Security Considerations

This document provides guidelines for exporting metadata about anonymized data in IPFIX, or storing metadata about anonymized data in IPFIX Files. It is not intended as a general statement on the applicability of specific flow data anonymization techniques. Exporters or publishers of anonymized data must take care that the applied anonymization technique is appropriate for the data source, the purpose, and the risk of deanonymization of a given application.

Research in anonymization techniques, and techniques for deanonymization, is ongoing, and currently "safe" anonymization techniques may be rendered unsafe by future developments.

We note specifically that anonymization is not a replacement for encryption for confidentiality. It is only appropriate for protecting identifying information in data to be used for purposes in which the protected data is irrelevant. Confidentiality in export is best served by using TLS [RFC5246] or Datagram Transport Layer Security (DTLS) [RFC4347] as in the Security Considerations section of [RFC5101], and in long-term storage by implementation-specific protection applied as in the Security Considerations section of [RFC5655]. Indeed, confidentiality and anonymization are not mutually exclusive, as encryption for confidentiality may be applied to anonymized data export or storage, as well, when the anonymized data is not intended for public release.

We note as well that care should be taken even with well-anonymized data, and anonymized data should still be treated as privacy sensitive. Anonymization reduces the risk of misuse, but is not a complete solution to the problem of protecting end-user privacy in network flow trace analysis.

When using pseudonymization techniques that have a mutable mapping, there is an inherent trade-off in the stability of the map between long-term comparability and security of the dataset against deanonymization. In general, deanonymization attacks are more effective given more information, so the longer a given mapping is valid, the more information can be applied to deanonymization. The specific details of this are technique-dependent and therefore out of the scope of this document.

When releasing anonymized data, publishers need to ensure that data that could be used in deanonymization is not leaked through a side channel. The entire workflow (hardware, software, operational policies and procedures, etc.) for handling anonymized data must be evaluated for risk of data leakage. While most of these possible side channels are out of scope for this document, guidelines for reducing the risk of information leakage specific to the IPFIX export protocol are provided in Section 7.2.

Note as well that the Security Considerations section of [RFC5101] applies as well to the export of anonymized data, and the Security Considerations section of [RFC5655] to the storage of anonymized data, or the publication of anonymized traces.

10. IANA Considerations

This document specifies the creation of several new IPFIX Information Elements in the IPFIX Information Element registry available from the IANA site (http://www.iana.org), as defined in Section 6.2. IANA has assigned the following Information Element numbers for their respective Information Elements as specified below:

- Information Element number 285 for the anonymizationFlags Information Element.
- Information Element number 286 for the anonymizationTechnique Information Element.
- Information Element number 287 for the informationElementIndex Information Element.

11. Acknowledgments

We thank Paul Aitken and John McHugh for their comments and insight, and Carsten Schmoll, Benoit Claise, Lothar Braun, Dan Romascanu, Stewart Bryant, and Sean Turner for their reviews. Special thanks to the FP7 PRISM and DEMONS projects for their material support of this work.

12. References

12.1. Normative References

[RFC5101] Claise, B., "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information", RFC 5101, January 2008.

[RFC5102] Quittek, J., Bryant, S., Claise, B., Aitken, P., and J. Meyer, "Information Model for IP Flow Information Export", RFC 5102, January 2008.

[RFC5103] Trammell, B. and E. Boschi, "Bidirectional Flow Export Using IP Flow Information Export (IPFIX)", RFC 5103, January 2008.

[RFC5655] Trammell, B., Boschi, E., Mark, L., Zseby, T., and A. Wagner, "Specification of the IP Flow Information Export (IPFIX) File Format", RFC 5655, October 2009.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5735] Cotton, M. and L. Vegoda, "Special Use IPv4 Addresses", BCP 153, RFC 5735, January 2010.

[RFC5156] Blanchet, M., "Special-Use IPv6 Addresses", RFC 5156, April 2008.

12.2. Informative References

[RFC5470] Sadasivan, G., Brownlee, N., Claise, B., and J. Quittek, "Architecture for IP Flow Information Export", RFC 5470, March 2009.

[RFC5472] Zseby, T., Boschi, E., Brownlee, N., and B. Claise, "IP Flow Information Export (IPFIX) Applicability", RFC 5472, March 2009.

[RFC6183] Kobayashi, A., Claise, B., Muenz, G., and K. Ishibashi, "IP Flow Information Export (IPFIX) Mediation: Framework", RFC 6183, April 2011.

[IPFIX-PERSTREAM] Claise, B., Aitken, P., Johnson, A., and G. Muenz, "IPFIX Export per SCTP Stream", Work in Progress, May 2010.

[RFC5153] Boschi, E., Mark, L., Quittek, J., Stiemerling, M., and P. Aitken, "IP Flow Information Export (IPFIX) Implementation Guidelines", RFC 5153, April 2008.

[RFC3917] Quittek, J., Zseby, T., Claise, B., and S. Zander, "Requirements for IP Flow Information Export (IPFIX)", RFC 3917, October 2004.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[RFC4347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security", RFC 4347, April 2006.

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

[Bur10] Burkhart, M., Schatzmann, D., Trammell, B., and E. Boschi, "The Role of Network Trace Anonymization Under Attack", ACM Computer Communications Review, vol. 40, no. 1, pp. 6-11, January 2010.

[Mur07] Murdoch, S. and P. Zielinski, "Sampled Traffic Analysis by Internet-Exchange-Level Adversaries", Proceedings of the 7th Workshop on Privacy Enhancing Technologies, Ottawa, Canada, June 2007.

Authors' Addresses

Elisa Boschi
Swiss Federal Institute of Technology Zurich
Gloriastrasse 35
8092 Zurich
Switzerland

EMail: boschie@tik.ee.ethz.ch

Brian Trammell
Swiss Federal Institute of Technology Zurich
Gloriastrasse 35
8092 Zurich
Switzerland

Phone: +41 44 632 70 13
EMail: trammell@tik.ee.ethz.ch