Cyber-security research by ISPs: A NetFlow and DNS Anonymization Policy

Martin Fejrskov
Technology, IP Network and Core
Telenor A/S
Aalborg, Denmark
mfea@telenor.dk

Jens Myrup Pedersen
Cyber Security Network
Aalborg University
Aalborg, Denmark
jens@es.aau.dk

Emmanouil Vasilomanolakis
Cyber Security Network
Aalborg University
Copenhagen, Denmark
emv@es.aau.dk

***Abstract*—Internet Service Providers (ISPs) have an economic and operational interest in detecting malicious network activity relating to their subscribers. However, it is unclear what kind of traffic data an ISP has available for cyber-security research, and under which legal conditions it can be used. This paper gives an overview of the challenges posed by legislation and of the data sources available to a European ISP. DNS and NetFlow logs are identified as relevant data sources and the state of the art in anonymization and fingerprinting techniques is discussed. Based on legislation, data availability and privacy considerations, a practically applicable anonymization policy is presented.**

***Index Terms*—ISP, privacy, DNS, NetFlow, IPFIX, cyber-security, anonymization**

I. Introduction

Research in cyber-security is highly dependent on the availability of real-life traffic traces for a number of different purposes. When collecting these traffic traces, researchers and practitioners should consider aspects like legal requirements and the privacy risk involved. However, these topics may not be within the researchers' area of knowledge. This can have a number of undesirable consequences like increased project lead time, legal problems when sharing project data, or spending time on research that is irrelevant because it cannot be applied in practice. The purpose of this paper is to help researchers and practitioners avoid some of these pitfalls when collecting data at an ISP level.

To protect the privacy of the subscribers, European ISP legislation forbids the use of certain data, and sets anonymization requirements on data usage, requirements that also apply to positive use cases like cyber-security research. However, the legislation does not present which specific anonymization techniques must be used for specific data sources. Some studies of anonymization techniques and privacy risks focus broadly rather than on giving practical guidelines for specific use cases. Other studies investigate how specific data sources can present a privacy problem in different use cases, but do not consider if the data is already unavailable from a legal perspective or how to mitigate the privacy risk.

In this paper, we firstly identify the ISP data sources legally and technically available for research. Furthermore, we present a practically applicable and privacy-preserving anonymization policy, for NetFlow and DNS logs, that complies with the relevant ISP legislation. This allows researchers and developers to start with a focus on implementation rather than legislation when creating solutions targeted for ISP deployments.

Funded by Telenor A/S and Innovation Fund Denmark, 2020

This paper is organised in seven sections. Section II gives an introduction to the relevant legislation and anonymization requirements, and Section III provides an overview of the data sources often technically available to an ISP. Having limited the relevant scope to two data sources, Section IV presents related work on anonymization techniques and on subscriber fingerprinting based on anonymized DNS and NetFlow logs. Sections V and VI, build upon the knowledge derived from all previous sections to propose and discuss concrete anonymization policies for individual fields in NetFlow and DNS logs, thus providing the primary contribution of the paper. Lastly, Section VII summarizes and concludes the paper.

II. Legislation

To identify the legal opportunities and challenges, an overview of relevant legislation is needed, which will be the topic of this section.

A. ePrivacy Directive

The ePrivacy Directive [1] from 2002 and the related national implementations, regulate among other things how ISPs are allowed to handle data related to the subscribers data traffic. The 2009 update of the ePrivacy Directive does not contain any changes relevant to this paper.

Although the General Data Protection Regulation (GDPR) [2] is newer than the ePrivacy Directive, the latter is considered *lex specialis* to the GDPR. This means that the ePrivacy Directive overrides the GDPR in any situation that is specifically described in the ePrivacy Directive. Furthermore, as the ePrivacy Directive specifically regulates ISPs and their handling of subscriber data traffic, the GDPR is considered out of the scope of this paper.

Articles 5, 6 and 9 in the ePrivacy Directive set the following limitations relevant to this paper on processing a subscribers traffic or location data:

- Data *already* being processed for the purpose of transmission must be made anonymous before additional processing.

- Data *not* being processed for the purpose of transmission or as part of a value added service cannot be processed.
- Data can be processed for a specific value added service but only if consent is available.

In the context of this paper, "processing" means any form of storage, manipulation, forwarding etc. of customer IP traffic, location data etc. [2] In addition, "processing for the purpose of transmission" refers to processing needed to transfer IP packets (routing, switching), performing DNS lookups (caching, recursing), authenticating the subscribers, routing packets to the correct cell tower and similar operations [1].

As it is practically impossible to have all subscribers sign up to a value added service relating to cyber-security research (and thereby providing consent), using anonymized data is the only viable strategy.

B. Opinion on Anonymization Techniques

Various anonymization and pseudonymization techniques and their relation to the legal framework are described in "Opinion 05/2014 on Anonymization Techniques" [3]. "Opinion" documents contain the elaboration of a specific directive or regulation, and are considered recommendations, not legislation. This specific opinion is written to elaborate on the anonymization requirements in the Data Protection Directive, a predecessor to the GDPR, but is still applicable and relevant.

Both the Opinion and recital 26 in the GDPR make a clear distinction between pseudonymization and anonymization, and makes it explicit that a requirement from the ePrivacy Directive to anonymize certain data is not fulfilled by the use of pseudonymization.

Two main anonymization techniques are described:

- **Randomization** including noise addition and permutation techniques "alter the veracity of the data in order to remove the strong link between the data and the individual". As an example, an IP address (A) in a specific data record can be substituted with a random IP address (B), and the same IP address (A) in another data record can be substituted with a different random IP address (C).
- **Generalization** including aggregation (k-anonymity), L-diversity and T-closeness techniques "generalize, or dilute, the attributes of data subjects by modifying the respective scale or order of magnitude". As an example, the IP addresses in all data records can be replaced by a smaller IP prefix.

The differential privacy technique is also described, but as this technique requires the original data to be retained, this technique is not compliant with the anonymization requirement of the ePrivacy Directive.

The Opinion concludes that in most cases it is not possible to give minimum recommendations for parameters to use as each data set needs to be considered on a case-by-case basis.

C. Summary

The ePrivacy Directive mandates that only data already being processed by the ISP for the purpose of transmission can be used for cyber-security research, and the data can only be retained in an anonymized form. The Opinion on Anonymization Techniques details which anonymization techniques are considered compliant. The specific data sources to be used for cyber-security research must therefore be determined before further anonymization considerations can be made.

III. Data sources

The restrictions posed by legislation depends on the type of data that is to be processed. In this section, the data sources available to Telenor Denmark will be described as a representative example of data sources being generally available to an ISP. Table I summarizes the data sources, their content and their usage restrictions based on the presentation of legislation in section II. This will provide an overview of which data sources are both legally *and* technically available for cyber-security research, which can help researchers determine if their research can be applied legally in practice.

The data sources that require anonymization are described in more detail in the following sections. Note that all data sources containing personal identifiable information like IP addresses require anonymization or consent to be used. Omitted are those that are not relevant in relation to Internet cyber-security research, thus excluding for example the SMS/MMS service and non-Internet based telephony services. Section III-D summarizes and discusses possible use cases for the available logs.

A. Identity of the subscriber

a) IP assignment log: Assigning an IP address to a subscriber is handled by different components depending on the access type (DSL/fiber/coax/mobile). Each component can, however, create an accounting log entry containing the subscriber identity (DSL-number or IMSI) and the assigned/revoked IP address. In a Telenor Denmark context, the DSL-number is a 4-6 digit broadband customer identifier that (despite the name) enumerates both coax, fiber and DSL customers.

b) CGNAT log: If mobile subscribers are assigned private IP addresses, Carrier Grade Network Address Translation (CGNAT) functionality is used. CGNAT can operate just like regular Network Address Translation (NAT) except that the NAT is performed at the ISP premises rather than at the customer premises. Multiple customers thereby share the same public IP. The Telenor Denmark CGNAT device reserves a range of 64 ports (a "port block") to each private IP address. Upon assigning/revoking this port block, a CGNAT log entry is created containing private IP address, public IP address and port block. Notice that a log entry is not created for each TCP/UDP session, it is only created for each port block allocation. The use of NAT logs can be relevant when distinguishing between different mobile subscribers sharing the same IP address.

c) EPDG CDR log: In order to use Voice-over-WiFi service the mobile phone must create an IpSec tunnel towards the Evolved Packet Data Gateway (EPDG). The EPDG can create a log line containing the IMSI and the source IP

Name	Usage restriction	Contents
IP assignment log	Anonymized	IP address, IMSI/IMEI/DSL-number
CGNAT log	Anonymized	Private/public IP address, port block
Customer database	Contract/consent	Person name, geographical address, IMSI/DSL-number
Modem/router at customer	Contract/consent	Attached device name, MAC and IP
EPDG CDR log	Anonymized	IP address, IMSI, RAT type (WiFi)
Cell database	None	Geographical address, gain/height/tilt etc.
Mobility event log	Anonymized	IMSI/IMEI, RAT type, cell ID
NetFlow log	Anonymized	TCP/UDP/IP session information
DNS log	Anonymized	Source IP address and port, queried domain name and response
Layer 3-7 DPI	Contract/consent	IP address, malware type
PGW application log	Contract/consent	IMSI/IMEI, IP address, layer 7 specific information
PGW flow log	Contract/consent	IMSI/IMEI, TCP/UDP/IP session and layer 7 application enumeration

TABLE I
ISP data sources relevant for cyber-security research

address of the IpSec tunnel. This log is known as a Call Data Record (CDR), despite the fact that it is not the phone call, but the tunnel establishment that is logged. This provides two interesting pieces of information: First the fact that a phone is attached to WiFi rather than being completely offline. Second, it shows which broadband subscription the mobile phone is connecting from. This can be used to distinguish between an infected broadband subscriber and an infected mobile subscriber using a broadband subscriber's WiFi.

B. Mobile location

a) Cell Database: Information about the geographical location, frequency, antenna gain/height/tilt, topography etc. of all cells is available in a central database. This can be used to estimate the coverage area of a specific cell.

b) Mobility event log: Phone mobility on 4G is handled by the Mobility Management Entity (MME) component, and this component can create a log line for each mobility event containing the subscriber identity (IMSI/IMEI), the destination cell identity (a 5-6 digit number) and the destination Radio Access Technology (RAT, 2G/3G/4G).

C. Internet activity

a) NetFlow log: The routers of the backbone network can emit NetFlow/IPFIX records. Most ISPs have equipment capable of doing this, but the specific implementation varies. ISPs may emit NetFlow logs from all routers or no routers, and may use varying levels of sampling/aggregation.

b) DNS log: Most subscribers (both mobile and broadband) use the ISP's DNS resolvers for name resolution. A log entry can, depending on the logging method, contain the client source IP/port, the query and the response. The authoritative DNS servers are considered less relevant for the topic of this paper, as traffic from ISP subscribers will in most cases be visible at the DNS resolvers as well.

D. Summary

This section outlines the different data sources available to Telenor Denmark as an example of a typical ISP, and identifies if consent or anonymization is required for data usage. Specifically for cyber-security research, the point of focus is the Internet activity (described by DNS and NetFlow logs) rather than the location or the subscriber identity. The DNS and NetFlow logs must be anonymized before use, and this is the topic of the rest of the paper.

IV. Related work

The previous sections argue that of the data sources technically and legally available to an ISP, NetFlow and DNS logs are the most interesting to cyber-security research. Having limited the scope, it is now relevant to identify existing, related work on DNS and NetFlow anonymization and fingerprinting. First, we provide a few notes on terminology and a general overview of related work. Afterwards, we discuss relevant papers in NetFlow and DNS respectively.

A. Terminology and overview

Many papers describe topics relating to anonymization, privacy and fingerprinting, so in order to discern which papers are the most relevant, an introductory note on terminology and preconditions is needed:

- Aggregation vs. generalization: Some papers use the terms generalization and aggregation interchangeably or with different definitions. For this paper, the terminology applied in RFC6235 will be used [4], and only generalization approaches are considered to preserve utility.
- Anonymization vs. pseudonymization: A brief look at existing literature, including taxonomy papers, shows that the distinction between anonymization and pseudonymization required by legislation is not often used, as typically the term "anonymization" is used for pseudonymization techniques as well.
- Anonymization must be applied before data analysis: Some techniques such as (k,j)-obfuscation [5] are based on a statistical analysis of the entire data set to be obfuscated, thus requiring all data to be stored in a non-anonymized form prior to release, which is not in line with legal requirements.

The goal of this paper is to provide a *DNS and NetFlow* anonymization policy stating which *anonymization technique* should be applied for individual *protocol fields*, while taking the *privacy risk* and ISP related *legislation* into consideration when focusing on the cyber-security research use case. Related

Aspect	Related Work							
	3	4,6	7	8,9,10,11	12	13	14	15
NetFlow or DNS		X	X	X	X	X		X
Anon. techniques	X	X	X	X			X	
Protocol fields		X				X	X	X
Privacy risk	X	X	X		X	X		
Legislation	X		X					

TABLE II
Notable related work and aspects in focus

works cover some but not all of these aspects, as illustrated by Table II.

B. NetFlow

A good introduction to the topic of passive internet measurement in general, including many aspects ranging from a legal overview to lessons learned on various practical deployment work is written by the authors of [7]. One of the lessons learned is that considering legislative aspects is a time consuming and complicated process, a problem that this paper attempts to address.

RFC 6235 provides a thorough walk-through of anonymization and pseudonymization options for the individual fields of the IPFIX protocol [4]. The paper categorizes various anonymization techniques into different classes, however, only the classes named ”generalization” (such as truncation) or ”set substitution” (such as noise addition) can be considered anonymization rather than pseudonymization techniques [3]. The paper does not provide any specific suggestions such as the length to be used for IP address truncation or on how much the precision of a timestamp should be degraded.

A comprehensive survey of anonymization techniques and 25 tools is written by the authors of [6]. The paper also discusses the relevance of anonymizing different fields in the different protocol layers in a network packet capture. The paper concludes with a number of statements like ”The port number should not be anonymized as it will have a big impact on the usefulness of a network capture and cannot be directly used for identification” [6] and ”Currently, in an environment without completely trusted parties, it is not recommended to share complete anonymized datasets. The current protection against re-identification is still inadequate.” (due to the large amount of context available in complete datasets) [6]

C. DNS

Two papers show that it is possible to perform user fingerprinting based on the domain name part of DNS logs [8], [9]. However, no suggestions on how to anonymize the DNS logs in data storage / mining environments are provided.

The authors of [11] describe the best privacy practices for DNS operators. Authenticity and confidentiality mechanisms like DNSSEC and DNS-over-TLS are described, but the section detailing how to protect data at rest focuses mainly on data minimization, IP address anonymization and TCP/TLS related features.

The implications of using only requests for the top n most popular host names for identity fingerprinting, as well as using only requests for anything but the n most popular host names is discussed by the authors of [8]. This is relevant in the context of cyber-security research as this idea can be used for data minimization, thus decreasing the privacy risk.

Field	Technique	Specifics
Bytes/packets	Precision degradation	NetFlow 1:n sampling
Start/end time	Reverse truncation	Remove AM/PM info
IP addr.(no NAT)	Truncation	Truncate to /24 prefix
IP addr.(CGNAT)	None	-
IP addr.(Infrastr.)	None	-
IP addr.(external)	Truncation	Truncate to /16 prefix
IP protocol	Binning	TCP+UDP+ICMP/”other”
ICMP type+code	None	-
Port (no NAT)	None	-
Port (CGNAT)	Truncation	Truncate to /2 prefix
Port (Infrastr.)	None	-
Port (external)	None	-
TCP flags	None	-

TABLE III
NetFlow anonymization policy assuming 64 port block based CGNAT

Bloom filters [16] rely on hash functions to store domain names in an irreversible way. While this provides good privacy, it also reduces the utility of the stored data, as data can then only be used to search for *already known* malware related domain names. This excludes for example domain names created by a Domain Generation Algorithm. While this can be sufficient from an operational perspective, it is less interesting to a cyber-security researcher, and therefore Bloom filters will not be considered further in this paper.

D. Summary

Related work does not provide a concrete answer on how an anonymization policy could be implemented, but does provide some good directions. The most specific input to an anonymization policy is provided by RFC 6235, which suggests specific techniques like truncation, but not directions on the truncation length. Based on these directions, section V and VI will provide a suggestion for a legally compliant anonymization policy suitable for ISP cyber-security research.

V. NetFlow anonymization policy

Based on the directions offered by legislation and related work on anonymization of NetFlow described in the previous sections, this section will provide a suggestion for a legally compliant anonymization policy suitable for ISP cyber-security research. The section describes the choice of protocol field/features, elaborates on the choice of anonymization technique for the individual fields, and concludes by providing the pseudo-code implementing such a policy.

A. Choice of features

The IPFIX features most typically used for cyber-security research [17] are listed in Table III along with the suggested anonymization policy. ICMP type/code and TCP flags are also added to table. The following paragraphs describe the considerations for each field noted in the table.

B. Feature anonymization details

a) Total bytes and packets: The total count of bytes and packets in a TCP/UDP session can be used for user profiling and for attacks against other anonymization techniques [4]. Moreover, it can under some circumstances be used as part of an algorithm to determine which web sites are visited [15].

The discussion may, however, be less important in practice, as NetFlows are typically sampled 1:n when collected by an ISP for performance reasons. The sampling also automatically provides a precision degradation of packet and byte counts, which is considered a valid method of anonymization for that field [4]. From a performance perspective, network equipment vendors consider $n <= 512$ a very low sample rate. This order of magnitude for sampling seems likely to be sufficient for anonymization purposes although to the best of our knowledge, no research has been conducted on quantifying this.

b) IP addresses: The authors of [10] conclude that if any other type of IP address anonymization technique than truncation is applied, re-identification of a host in NetFlow traffic is possible when active fingerprinting techniques are applied. If IP address truncation is applied, other fields may still be able to identify the host, though.

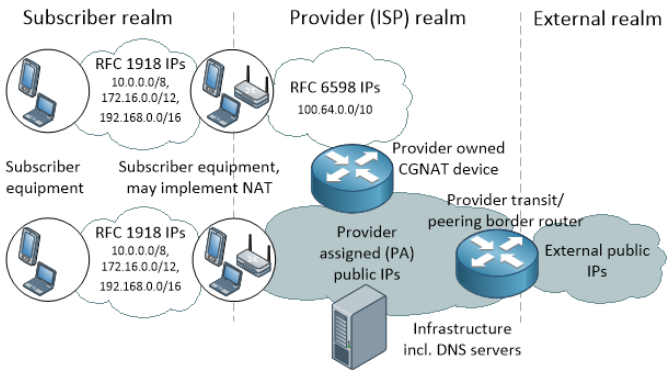

Fig. 1. *Classification of IP addresses*

From the perspective of an ISP capturing NetFlow at the border routers, 4 different categories of IP addresses are relevant to describe separately, as illustrated in Figure 1: Some Provider Assigned (PA) IP addresses are allocated directly to subscriber equipment (typically broadband routers at customer premises), some PA IP addresses are allocated for use on the outside of the CGNAT device and some PA IP addresses are allocated to ISP infrastructure, including the DNS servers, content caches, routers etc. IP addresses outside the ISP/provider realm (more specifically IP addresses not announced through Border Gateway Protocol (BGP) by the ISP outside the ISPs Autonomous System (AS)) will simply be referred to as ”external” IP addresses.

Whether PA and external IP addresses should be subjected to the same truncation length is discussed in [12] based on a ”risk vs. utility” analysis. Choosing the ”sweet spot” with the most utility preserved, this would be equivalent to truncating PA IP addresses to their /24 prefix and external IP addresses to their /16 prefix.

Truncation of prefixes is considered an implementation of k-anonymity [3]. Extensions to k-anonymity like l-diversity and t-closeness require an analysis of the data distribution before anonymization. These extensions are therefore not immediately implementable in practice.

c) Timestamps: Several papers discuss host fingerprinting based on ICMP Timestamp Requests/Replies and the TCP Timestamp option, as these contain a timestamp that originates from the host. However, neither DNS or NetFlow logs contain a host originated timestamp, as only the timestamps from the NetFlow/DNS log capture devices are logged. This timestamp can be used in an injection attack to identify a host in a traffic trace with pseudonymized (permuted) IP addresses [13].

In the case where only one subscriber in a truncated IP prefix is actively generating traffic at a specific period of time (e.g., during nighttime), this subscriber does not benefit from IP address truncation. To preserve anonymity, the precision of the timestamp could be reduced to for example an hour or a minute. These approaches are typically infeasible for research, as the order of events is not preserved. An approach not described by the authors of [4] or other known sources is to simply remove AM/PM information from the timestamp. This approach has the advantage compared to traditional precision reduction that the order of all interrelated NetFlow/DNS events that do not cross the AM/PM time boundary is preserved.

d) IP protocol: The IP header protocol field is not considered privacy sensitive by any known papers, the authors of [6] even omit the discussion of the field entirely. RFC 6235 suggests using the binning technique such that 4 bins are used: TCP, UDP, ICMP and ”all other protocols”, an approach which seems suitable for cyber-security research as well.

e) ICMP type/code: ICMP messages and their payload are widely used for OS fingerprinting by tools such as Nmap. The methods typically involve differentiating using TTL or some other IP field, however a specific method creates ICMP requests using illegal combinations of type and code values, and the ICMP response code can then in some cases reveal the OS family [18]. To anonymize this, the code field could be omitted from the logs. However, doing so comes with a significant drawback, as it will obviously also hide any malware using the technique for OS detection.

Note that when using NetFlow logs from an ISP, the OS family revealed will typically not reveal the end users' operating system. Instead, it will reveal either the OS family of the subscribers' modem/router/firewall or the CGNAT device deployed by the ISP. Therefore, the reasonable compromise for cyber-security research seems to be not to anonymize this field.

f) Ports: The authors of [14] conclude that anonymizing ports or IP addresses, as opposed to anonymizing other individual fields, have the biggest impact on the utility of the data. However, the risk and the risk/utility trade-off is not discussed in the paper. Not much research was found that quantifies the risk of host fingerprinting based on port numbers when IP

addresses are truncated in practice. The authors of [13] provide a short note describing that injected flow patterns are no longer recognizable under certain anonymization policies. However, they do not describe a systematic approach or conclusion for this. This is likely caused by the fact that much attention has already been given to properly randomizing TCP port numbers to avoid Denial-Of-Service and Man-In-The-Middle attacks [19]. The authors of [6] conclude that the port number should not be anonymized as it will have a big impact on the usefulness of a network capture.

g) TCP flags: TCP flags can be used for OS fingerprinting using a technique similar to the one described for fingerprinting using ICMP type and codes: Specific flags in a request can trigger an OS-specific flag combination in the response. Analyzing TCP flags is key in detecting malware employing DDoS SYN attacks and other attack types.

As with the ICMP type and code, the OS family revealed by TCP flags will typically not reveal the type of CGNAT device deployed by the ISP. Therefore, the reasonable compromise seems to be not to anonymize this field.

h) NAT: Most ISPs implement CGNAT for at least a subset of their subscribers, so that one IP address contains traffic from more than one subscriber. Many port allocation schemes exist, and it is beyond the scope of this paper to describe all. However, from an anonymization perspective, two different consequences of introducing CGNAT can be relevant: *decreasing the truncation length of the IP address* and *increasing the truncation length of the port*.

In a CGNAT scheme where a single RFC6598 IP address is shared by for example 32 subscribers (5 bits) by random port assignment, the PA IP address truncation length could be reduced from a /24 (256 addresses,8 bits) to a /29 (4 addresses,32+5-8=29 prefix) to preserve utility as the expected amount of hosts grouped will then be the same.

In a scheme where the port allocation is not random, but based on a range of ports being reserved for a particular host, or where initially randomly assigned ports are heavily reused for the same subscriber, the port information must be truncated using the same methodology as the IP addresses. For example, if a port block of 64 ports (6 bits) are reserved for each user, and anonymization equivalent to a /24 IP prefix (256 addresses, 8 bits) is desired, the port number must be reduced to 16-6-8=2 bits. However, the PA IP address truncation can then be reduced to a 32+16-6-8=34 prefix, effectively making the anonymization of the PA IP address unneeded.

C. Pseudo-code: a NetFlow anonymization policy

The pseudo-code listed in Listing 1 implements the anonymization policy summarized in Table III assuming sampling by the NetFlow emitter. Lines 2-3 remove AM/PM information, lines 8 and 14 truncate IP addresses to /8 and /16 prefixes, line 10 truncates the port number to a /2 prefix for customers with NAT (assuming 64 port range based CGNAT). It is noteworthy that the implementation can be made with basic operations. This allows a high level of performance, which is required for ISP deployments. Searching for an IP address in a list of prefixes (lines 7, 9 and 11) should also be implemented effectively. This is considered trivial, assuming a small amount of non-overlapping prefixes is used, and therefore it is omitted for readability. Finally, lines 22-23 implement binning of protocol information into 4 different bins.

Listing 1. NetFlow anonymization policy.

```
1  def anontimestamp(timestamp t)
2      if t.hour >= 12:
3          t.hour = t.hour-12
4      return t
5
6  def anonipport(int32 ip, int16 port)
7      if ip in listOfSubscriberAssignedPrefixes:
8          ip = ip & 0xFFFFFF00
9      else if ip in listOfCGNatPrefixes:
10         port = port & 0xC000
11     else if ip in listOfInfrastructPrefixes:
12         // do nothing
13     else: //external
14         ip = ip & 0xFFFF0000
15     return ip,port
16
17 starttime = anontimestamp(starttime)
18 endtime = anontimestamp(endtime)
19 srcip,srcport = anonipport(srcip,srcport)
20 dstip,dstport = anonipport(dstip,dstport)
21
22 if protocol != (ICMP or TCP or UDP):
23  protocol = 0
```

VI. DNS ANONYMIZATION POLICY

Similar to the previous section but focusing on DNS rather than NetFlow, this section will provide a suggestion for a DNS anonymization policy, and provide the pseudo-code implementing such a policy.

A. Choice of features

The choice of features for DNS based cyber-security research is very diverse [20]. Moreover, whereas NetFlow is only a format for logging passively collected flow properties, DNS is a service used (and potentially attacked) by subscribers. This calls for a more full-featured approach to logging than focusing on a few specific fields. One example is that a protocol violation could be made intentionally by a client to attack the DNS service, and this can only be discovered if the specific field containing the violation is logged.

A DNS packet typically consists of a header section and a query section, and response packets also include one or more sections containing the answer to the query. The answer sections can contain a number of different resource records (RRs). The content, typically an IP address or domain name, and the interpretation of the query and answer RRs depend on flags in the header as well as on which specific type of information is queried.

The increased field diversity and inter-dependency makes DNS log anonymization more complicated than NetFlow log anonymization: Table IV lists a number of fields for which an anonymization policy can be directly described, whereas

Field	Anonymization tech.	Specifics
Timestamp	Reverse truncation	As NetFlow
Client IP address	As Netflow	As NetFlow
Client TCP/UDP port	As Netflow	As NetFlow
DNS header	None	-
DNS response TTL	Binning	5 predefined bins

TABLE IV
Content independent DNS anonymization policy

Opcode	Class	Type	Domain anon. technique
Not Query	-	-	None
Query	Not IN	-	None
Query	IN	Common types	Minimization
Query	IN	Uncommon types	None

TABLE V
Content based DNS anonymization policy

Table V lists the type dependent anonymization techniques. The lines of the tables are elaborated in the following.

B. Feature anonymization details

a) Timestamp, Client IP address and TCP/UDP port: For these fields, the anonymization policy also used for the similar fields NetFlow packets is chosen. To the best of our knowledge, no research is made that indicates that DNS and NetFlow logs should be subject to different anonymization requirements relating to these fields.

b) DNS header: The DNS header consists of a number of identifiers, response codes and flags. Many of these are needed to parse the non-header components, and no fields contain directly personal identifying information. The randomness of the Message ID has, like the randomness of the TCP/UDP source port number, been subject to scrutiny to prevent Man-in-the-Middle attacks, so this field is expected to be properly randomized to not represent a privacy risk.

c) TTL: The TTL value found in the answer sections of a DNS packet could, together with the timestamp, be used to determine that two clients requested the same RR, as these would have the same TTL. Nevertheless, it is unknown whether this can be practically exploited for subscriber fingerprinting. The bins $[0, 1), [1, 100), [100, 300), [300, 900), [900, \infty)$ are found to be relevant for cyber-security research [21], and therefore this technique is chosen.

d) Uncommon opcodes, classes and types: Request and response messages containing an Opcode of any other value than "query" (such as "status" or "update"), query messages of any other class than IN (such as Chaos and Hesiod) and IN class query messages of any other type than the 15 most common types (see below) are represented by the first two lines and the last lines in Table V. A smaller data sample collected at Telenor Denmark suggests that traffic in these three categories represent misconfigured equipment, malformed packets and spurious requests with an empty response. This type of traffic does not seem to be the result of human Internet usage behavior and is therefore not likely to represent any privacy risk. However, as mentioned initially, the traffic may represent an attack initiated by malware, and therefore the data is still relevant to retain.

e) Common types: On Telenor Denmark's resolvers, the 15 most common query types in the IN class are A, AAAA, A6, CNAME, PTR, MX, TXT, SRV, NAPTR, NS, SOA, DS, RRSIG, DNSKEY and NSEC3. Resource records of these types typically consist of a QNAME component (the name queried) and an RDATA component (the response to the queried name). Either of these components can contain an IP address, a domain name or a string of text containing either of the two, such as SRV or TXT records. It is clear that any anonymization policy applied to an RR must be applied to both the QNAME and RDATA components, as one component can typically be derived from the other by issuing a new DNS request, thus breaking the anonymization.

f) Domain name: As described in Section IV, the queried domain name can be used to fingerprint subscribers, and the only known anonymization strategy is data minimization. The authors of [8] suggest two minimization strategies: Omitting the most or least popular hostnames. From a cyber-security research perspective, omitting the least popular hostnames severely decreases data utility. As an example, botnets based on Domain Generation Algorithms (DGAs) are likely to be rendered undetectable.

The authors of [8] argue that omitting the most popular hostnames would have only a limited effect on fingerprinting risk, though the effect increases when the 500-1000 most popular hostnames are omitted. However, it is questionable if this result applies on an ISP network in 2020. The paper analyses data from approximately 3600 users on a campus network in 2010, where removing the 1000 most popular hostnames is equivalent to removing 51,2% of all queries. Nevertheless, on Telenor Denmark's network having around 1.7 million subscribers in 2020, the same percentage of queries relates only to 15 domains and associated subdomains[1]. This suggests that significant data minimization (removing $> 50\%$) could decrease the fingerprinting risk. If the omitted domain names represent domains that are less interesting from a cyber-security research perspective, the utility of the data can be preserved while decreasing the fingerprinting risk.

OS fingerprinting can be avoided using the same technique, by simply adding known OS-specific domain names and IP addresses to the list of omitted domains. This includes for example captive portal detection mechanisms (such as resolving "connectivitycheck.gstatic.com"), proxy detection (resolving the "wpad" hostname), etc.

C. Pseudo-code: a DNS anonymization policy

The pseudo-code listed in Listing 2 implements the anonymization policy summarized in Tables V and IV. The anonymization functions for timestamps and client IP address/-port (lines 1 and 2) can be found in Listing 1. Line 11 represents the binning of the TTL value, but the implementation of the function itself is left out for brevity. Lines 5-7 clear

[1]Specifically: apple.com, facebook.com, akadns.net, google.com, googleapis.com, snapchat.com, akamaiedge.net, fbcdn.net, icloud.com, apple-dns.net, doubleclick.net, gstatic.com, netflix.com, microsoft.com and googlevideo.com.

Listing 2. DNS anonymization policy.

```
1  timestamp = anontimestamp(timestamp)
2  ip,port = anonipport(ip,port)
3
4  if header.opcode == Query:
5   if query.class==IN and query.type in
        commonTypes:
6     if any in commonDomainList in query.name:
7      query.name = ""
8
9   foreach rr in answerSectionsOfPayload:
10   if rr.class==IN and rr.type in commonTypes:
11    rr.ttl  = integerBinning(listOfIntervals)
12    if any in commonDomainList in rr.name:
13     rr.name = ""
14     rr.data = ""
15    if any in commonDomainList in rr.data:
16     rr.name = ""
17     rr.data = ""
```

the queried domain name if it matches or is a sub-domain of the domain names listed in commonDomainList. Lines 12-17 perform the same operation on the Answer RRs, which includes searching for the domain name in both the question (rr.name) and response (rr.data) part of the RR. For brevity, the Answer payload section is considered to also include the Additional and Authoritative sections.

The DNS anonymization pseudo-code is clearly more computationally heavy than the NetFlow anonymization pseudo-code due to the use of string operations. This is to some extent mitigated by the list of common domains being short.

VII. Conclusion

It has previously been unclear what traffic data an ISP has available for cyber-security research, and under which legal conditions it can be used. This paper attempts to address this by presenting relevant legislation and data sources, and by presenting an anonymization policy for the relevant data.

The EU ePrivacy Directive puts strict requirements on which data can be used by ISPs. Only data that is already used for the purpose of transmission can be used for other purposes, and then only when anonymized. If use of other data and/or use of data in a non-anonymized form is desired, an explicit consent from the subscriber is required. We present the relevant data sources available to a typical ISP, using Telenor Denmark as example, and argue that *DNS and NetFlow* data are identified as relevant to cyber-security research and as technically and legally available data sources under the condition that the data is anonymized before further processing. We elaborate by proposing anonymization policies (in the form of pseudo-code) for DNS and NetFlow log data.

The proposed anonymization policies make use of various techniques for generalization, such as truncation of IP addresses, precision degradation of timestamps, data minimization on collected DNS logs etc. as mandated by legislation and suggested and inferred by best practices and related work. The pseudo-code implements the anonymization in a computationally inexpensive way such that application at ISP-scale traffic rates is possible. The anonymization policies and related pseudo-code are considered the primary contribution of this paper, giving researchers and developers a concrete and technically focused starting point when creating solutions targeted for deployment in ISPs.

References

[1] The European Parliament and of the Council, "Directive 2002/58/ec (the ePrivacy directive)," 2002. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32002L0058

[2] ——, "Regulation (eu) 2016/679 (the General Data Protection Regulation, GDPR)," 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[3] The Working Party on the Protection of Individuals With Regard to the Processing of Personal Data, "Opinion 05/2014 on Anonymisation Techniques," 2014. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[4] E. Boschi and B. Trammel, "Ip flow anonymization support, rfc 6235," 2011. [Online]. Available: https://doi.org/10.17487/RFC6235

[5] D. Riboni, A. Villani, D. Vitali, C. Bettini, and L. V. Mancini, "Obfuscation of sensitive data for incremental release of network flows," 2014. [Online]. Available: https://doi.org/10.1109/TNET.2014.2309011

[6] N. Dijkhuizen and J. Ham, "A survey of network traffic anonymisation techniques and implementations," 2018. [Online]. Available: https://doi.org/10.1145/3182660

[7] W. John, S. Tafvelin, and T. Olovsson, "Passive internet measurement: Overview and guidelines based on experiences," 2009. [Online]. Available: https://doi.org/10.1016/j.comcom.2009.10.021

[8] D. Herrmann, C. Banse, and H. Federrath, "Behavior-based tracking: Exploiting characteristic patterns in dns traffic," 2013. [Online]. Available: https://doi.org/10.1016/j.cose.2013.03.012

[9] D. W. Kim and J. Zhang, "Deriving and measuring dns-based fingerprints," 2017. [Online]. Available: https://doi.org/10.1016/j.jisa.2017.07.006

[10] D. Sauter, M. Burkhart, D. Schatzmann, and B. Plattner, "Invasion of privacy using fingerprinting attacks," 2009. [Online]. Available: https://pub.tik.ee.ethz.ch/students/2008-HS/MA-2008-22.pdf

[11] S. Dickinson, B. Overeinder, R. van Rijswijk-Deij, and A. Mankin, "Recommendations for DNS Privacy Service Operators," 2019. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-ietf-dprive-bcp-op-07

[12] M. Burkhart, D. Brauckhoff, M. May, and E. Boschi, "The risk-utility tradeoff for ip address truncation," 2008. [Online]. Available: https://doi.org/10.1145/1456441.1456452

[13] M. Burkhart, D. Schatzmann, B. Trammell, E. Boschi, and B. R. Plattner, "The role of network trace anonymization under attack," 2010. [Online]. Available: https://doi.org/10.1145/1672308.1672310

[14] K. Lakkaraju and A. Slagell, "Evaluating the utility of anonymized network traces for intrusion detection," 2008. [Online]. Available: https://doi.org/10.1145/1460877.1460899

[15] S. E. Coull, M. P. Collins, C. V. Wright, F. Monrose, and M. K. Reiter, "On web browsing privacy in anonymized netflows," 2007. [Online]. Available: https://doi.org/10.5555/1362903.1362926

[16] R. van Rijswijk-Deij, G. Rijnders, M. Bomhoff, and L. Allodi, "Privacy-Conscious Threat Intelligence Using DNSBLoom," 2019. [Online]. Available: http://dl.ifip.org/db/conf/im/im2019/189282.pdf

[17] D. C. Ferreira, M. Bachl, G. Vormayr, F. Iglesias, and T. Zseby, "A meta-analysis approach for feature selection in network traffic research," 2017. [Online]. Available: https://doi.org/10.1145/3097766.3097771

[18] Nmap.org, "Tcp/ip fingerprinting methods supported by nmap," 2019. [Online]. Available: https://nmap.org/book/osdetect-methods.html

[19] M. Larsen and F. Gont, "Recommendations for transport-protocol port randomization, rfc 6056," 2011. [Online]. Available: https://doi.org/10.17487/RFC6056

[20] M. Singh, M. Singh, and S. Kaur, "Issues and challenges in dns based botnet detection: A survey," 2019. [Online]. Available: https://doi.org/10.1016/j.cose.2019.05.019

[21] L. Bilge, S. Sen, D. Balzarotti, E. Kirda, and C. Kruegel, "Exposure: a passive dns analysis service to detect and report malicious domains," 2014. [Online]. Available: https://doi.org/10.1145/2584679