The Impact of Passive DNS Collection on End-user Privacy

Jonathan M. Spring, Carly L. Huth

***Abstract*—There are two distinct problems in determining the impact of passive DNS (pDNS) on end-user privacy. One is whether or not pDNS would allow the observer to reconstruct an individual end-user's DNS behavior. The other is if DNS behavior constitutes personally identifiable information (PII) or is otherwise legally protected. This paper develops a framework to discuss both aspects of the privacy issue. From the technical point of view, DNS sensor architecture is analyzed and a statistical model is developed to describe the sensor's ability to violate end-user privacy. To the other end, a review of various jurisdictions' privacy legislation is presented and analyzed in the context of DNS as a system and pDNS as a collection mechanism. In general, we find that pDNS, properly configured, does not violate end-user privacy.**

***Index Terms*—Passive DNS, Privacy and the DNS, Measurement studies**

I. Introduction

THE Domain Name System (DNS) is a nearly ubiquitous Internet protocol for carrying identifying information about machines, pages, and subsystems. It is an interesting and valuable target for network analysis and situational awareness for both its ubiquity and informational content. Broadly, there are two methods of collecting information from DNS — actively and passively. In the former, the observer generates queries to some number of name servers on specific topics of interest and analyzes aspects of the responses. In the latter, the observer merely records some subset of the DNS messages generated by others. Passive DNS (pDNS) monitoring has many advantages over active DNS monitoring, such as comprehensiveness, easier implementation, observation of previously unknown behaviors, stealthy observation, and not increasing the load on the name servers or network [1]. This document, however, will focus on assuaging a concern about capturing pDNS. That concern is the potential negative impact on end-user privacy. We find the impact to be between minimal and none.

There are two distinct problems in determining the impact of pDNS on end-user privacy. One is whether or not pDNS would violate an end-user's privacy by allowing the observer to reconstruct an individual end-user's pattern of DNS behavior. The other is if a set of DNS behavior constitutes personally identifiable information (PII) in the first place. The former is a technical question about system architecture and collection strategies. The latter is a policy question, and will potentially have a different answer per legal framework. A negative answer to either (the behavior is not recoverable, it is not PII) should demonstrate that pDNS does not significantly impact end-user privacy. This document is primarily concerned with elucidating the former: how pDNS data is technologically incapable of reconstructing end-user behavior. The latter is treated in a legal survey, highlighting starting points and broad themes in the legal argument; it is not to be considered an official legal opinion.

Jonathan Spring is with the Network Situational Awareness group at CERT, part of Carnegie Mellon University, Pittsburgh, USA, e-mail: jspring@cert.org.

Carly Huth is with the Enterprise Threat and Vulnerability Management group at CERT, part of Carnegie Mellon University, Pittsburgh, USA, e-mail: clhuth@sei.cmu.edu.

II. Technical considerations of pDNS privacy impact

There are four types of information present in pDNS collection:

1) DNS names sent as queries
2) DNS records sent in response to queries, i.e., positive answers
3) Negative answers to queries indicating no record exists
4) IP addresses associated with the systems asking and answering the queries

DNS, as a system, exists to distribute both positive and negative answers to queries. That information itself is public in much the same way the phone book is public. For a primer on DNS and an explanation of this concept in more detail, see Appendix A on page 6.

A. Identifying the parties to a query

In addition to any content of the DNS records, there are two IP addresses associated a DNS message: the questioner and the responder. The role played by sender and receiver switches if the message is a query or a response. Regardless, there is a different privacy impact for storing the IP address of either of these two roles.

1) The response issuer: The machine that issues DNS responses is not a private machine. DNS server names are publicly advertised in registrar WHOIS files and in the DNS itself (NS records) [2], and their IP addresses (A or AAAA records) are easily accessible given the name. DNS servers were long considered public resources until load considerations caused some administrators to restrict access. Obviously, in order for the DNS to function at all, the repositories of the information must be well known. Therefore the IP address of the machine that issued the response is not private information, and may be collected by pDNS.

2) On identifying the end-user: There are multiple possible configurations of pDNS sensors. This section discusses a general framework for describing the salient aspects of sensor placement for end-user privacy. Since the goal of this paper is to demonstrate that end-user privacy can be preserved, the argument demonstrates and uses such privacy-preserving configurations.

In the recommended architecture, collecting IP addresses of systems sending/receiving queries is not equivalent to identifying the end-user who issued the query [1], [3]. This issue is related to, but separate from, policy on whether device identification is PII. Due to the DNS protocol and implementation, there are various types of interference with collecting the identity of the querying entity. Figure 1 is a general diagram of entities involved in DNS resolution. Multiple external DNS resolvers can be involved in series, where the number of such resolvers, n, is in principle an integer from 0 to ∞. Dashed arrows indicate a transfer of data.

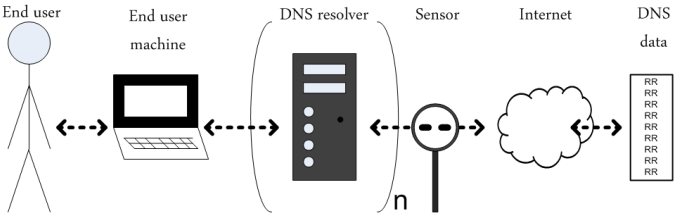

Figure 1. DNS resolution with pDNS sensor.

The operative question is what data is necessary to reconstruct a user's DNS behavior, perhaps with the further goal of reconstructing what network services the user has accessed. The following discussion will focus on uncovering the former; however note that reconstructing network behavior from DNS is an additional, separate issue.

There are two recurring themes in what makes attribution of DNS messages difficult: recursion and caching.

Recursion in the DNS protocol refers to successive resolvers issuing queries to collect information on behalf of the entity that originally asked the query. A recursive resolver is one that will issue queries on behalf of another resolver.

Caching is a mandatory feature of the DNS protocol. DNS resolvers store records locally for the duration specified in the time to live (TTL) of the record. The local cache is consulted first, to improve operational efficiency by reducing network queries issued.

Passive DNS is a network monitoring technique. This is important in attempts to reconstruct user sessions. Imagine that there is a sensor placed between the user's machine and everything else on the network. This sensor would hypothetically capture data that would be able to identify and trace what the user is doing. Of common protocols, this is probably least true about DNS. DNS is cached even on the local machine. Collecting DNS traffic data from the end-user machine, the observer would have no way to distinguish if the name was accessed more than one time during the TTL. The number of visits during the TTL is almost completely unknown. The upper bound is the transmission rate of the network, yet the lower bound is zero – the user need not ever visit a site that is looked up with DNS.

The uncertainty introduced due to caching is related to some parameters. More uncertainty is introduced by a larger TTL for the queried name's answer. This obviously follows, since a larger TTL is a larger window in which to access the resource without a DNS footprint, thus a longer period of uncertainty. A longer chain of recursion also increases uncertainty from caching because each resolver in the chain will cache the response.

Since pDNS is network monitoring, there are two locations one could place a pDNS sensor. To monitor the internal behavior of a small set of clients, one could monitor the internal interface of the organization's recursive name server. The other option is to monitor on the outside, either at the organization's boundary with the Internet or farther up the recursive chain of DNS resolution [3], [4]. Monitoring at the Internet service provider (ISP) or top-level domain (TLD) name server level would increase the scope and abstraction of the collection; this abstraction makes the collection yet higher in the network, increasing the value of n in Figure 1. Later it will be demonstrated that a larger value for n increases the uncertainty in identifying a user session. Outside monitoring is also sufficient for useful large scale DNS analysis. Often such sensor placement eases analysis because it forces more manageable data volumes [1].

Since DNS queries are in practice resolved hierarchically,[1] the querying entity is naturally anonymized. Before a query/response from an end-user machine passes a sensor, one or more other resolvers process the query and re-issue it with their IP address. The recursive nature of DNS makes identifying the time that a query was issued by the user somewhat uncertain as well, because the sensor detects when the query was issued by the recursive resolver.

With only one end-user machine making queries, these confounding conditions would not be extraordinarily difficult to overcome. Yet with multiple users, each user beyond the first to issue a query will not generate a detectable message, because the answer will already be in the cache. This effect lasts for the duration of the TTL of the DNS entry. This makes it impossible to determine the number of users who have issued a certain query. Since the sensor is above the resolver, there is also no unique identifier which can be used to associate with a particular end-user machine.

There are additional non-network-based factors that further frustrate identification of the end-user via DNS queries with the discussed sensor architecture. These are discussed in Appendix A-D on page 7.

[1] According to [5], DNS queries can be resolved recursively or iteratively. Iterative resolution must be supported, whereas recursive iteration may be supported. In practice, DNS queries go through at least one recursive resolver at the organizational level. In the context of pDNS, the proposed aggregation point for the queries is just above such a resolver. For these reasons, at least one level of recursion, and thus obfuscation of the IP address, is assumed.

B. Formalizing end-user identification probability

These constraints make some formalization of the problem possible. Variables will be used as follows:

$a := single\,resource\,record$
$R := the\,set\,\{a_1, a_2, \ldots, a_b\}\,in\,a\,message$
$\mathfrak{r} := the\,set\,of\,all\,\{R_1, R_2, \ldots\}\,during\,observation$
$m_i := A\,given\,user\,machine$
$M := all\,m_i\,serviced\,by\,a\,DNS\,resolver$
$o_d := the\,sensor\,at\,DNS\,server\,d$
$n := number\,of\,resolvers\,between\,m_i\,and\,o_d,\,as\,Figure\,1$
$P_i := \{R_0, R_1, \ldots, R_k\}\,such\,that\,each\,R\,was\,sent\,to\,m_i$
$p\,(o_d \rightarrow P_i) := probability\,that\,o_d\,correctly\,guesses\,P_i$

Using these terms some relationships from the previous discussion can be modeled as follows, where f and g are functions, $\propto$ denotes the "proportional to" relation, $\|\ \|$ denotes cardinality of a set, and R_{TTL} is the TTL of the resource records in question.

$$p\,(o_d \rightarrow P_i) \propto \frac{1}{f\,(n)} * \frac{1}{g\,(\|M\|\,, \|\mathfrak{r}\|)} \tag{1}$$

$$f\,(n) \propto R_{TTL} \tag{2}$$

The nature of f and g are important; they are almost certainly monotonically increasing. The number of resolvers adds noise. Attributing the data to potentially more users also increases the solution space. Neither of these would make it easier for a sensor to attribute a set of records to a machine.

The simplest case is a single machine's internal resolver. In this case, it would be trivial to identify the source of the traffic. Even so, correlating that set with user activity is not trivial. See appendix A-D on page 7 for a discussion of those issues.

A more realistic case for a system with one recursive resolver. The value of n becomes 2; each client retains its resolver, and the primary name server serves many clients. Calculation is complicated by the fact that DNS servers execute TTL control and caching at the level of resource records, not messages. However, the logical items we desire to reconstruct are the messages the end-user received, which are sets of resource records.

In the case of two resolvers, given any single message, the chance that a given machine is responsible for R is proportional to $1/m$. Due to caching, the sensor will only observe that R once during the TTL. During this time, no further observations of R will be made to improve attribution. One could attribute more intelligently than random chance. For a simplistic example, if machine m_i looks up example.com, one could posit the same machine was more likely to look up foo.example.com shortly thereafter. This is not very helpful in practice. Even if clustering is successful, the sensor would still not be able to attribute this cluster to a particular identifier. Furthermore, if one of the domains (e.g. foo.example.com or example.com) were already cached, the sensor would not be able to correlate such queries temporally in the first place because it would be answered from the name server's cache and would not pass the sensor.

The difficulties grow geometrically for each caching resolver before the sensor, and the number of possible machines to attribute. Each resolver will cache queries, so it is still the case that the sensor will only detect one instance of R, no matter how many machines request it within the TTL. In equations 1 and 2, $f(n)$ is a function that increases with more resolvers due to increased noise and the impact of caching.

The function f will also be related to the TTL of the records which the observer is attempting to attribute. This is due to caching. The longer the records are cached, the longer the period which the observer is ignorant of their activity. Extremely short TTLs would lessen the effect of caching. Short TTLs are observed, but they are more the exception than the rule. These measurements are reported in Appendix B on page 7.

The other function to which p is related is $g\,(\|M\|\,, \|\mathfrak{r}\|)$. This function draws on the reasoning that a larger search space makes it less likely that one can attribute a certain set to a particular machine. The function is partly determined by the fact that one must choose the correct combination of R to assign to each m_i. However, it is not clear how many elements to assign to each machine. Due to caching, it is not known how many machines to assign to each record, i.e. how many duplicates there should be of a given R. The observer knows the unique $\{R_1, R_2, \ldots\}$, not the total count for each R that occurred during the observation period. Therefore, each R could be assigned to from 1 to all, i.e. $\|M\|$, machines. Some intelligence could be applied to determining the search space; however the covering sample size would be that any machine could be responsible for any set of R. Any number of records could be assigned to each machine, in theory, including the null set because it would not be detectable in pDNS if a particular machine were active or not. This means that any combination of presence or absence of each record could be assigned to a single machine. We define the function X as the set of all sets of messages that could be assigned to machine m_i, such as$\{\{\}\,, \{R_p\}\,, \{R_p, R_q, ...\}\,, ...\}$. Therefore, the number of sets of records potentially assigned to a machine m_i is as follows:

$X\,(\mathfrak{r}, i) := combinations\,of\,R\,from\,\mathfrak{r}\,possible\,to\,link\,to\,m_i$

$$\|X\,(\mathfrak{r}, i)\| = 2^{\|\mathfrak{r}\|} \tag{3}$$

This number of possible combinations could be assigned to each machine associated with the name server at the sensor. Therefore, the size of the set of possible assignments for all machines would be defined by:

$$\sum_{i=1}^{i=\|M\|} \|X\,(\mathfrak{r}, i)\| = \|M\| * 2^{\|\mathfrak{r}\|} \tag{4}$$

Even if the function g is very efficient at reducing this sample size to attribute records to machines, this is a huge solution space to attempt to work with. Current sources of passive DNS information routinely report over 100 million unique resource records per day.[2] Additionally, since pDNS does not detect individual machine activity, the value of M is not precisely known. There may be peculiar circumstances in which individual machines can be counted in DNS alone, but

[2]As measured by the authors between Mar 1 and June 30, 2011.

this is not the majority case. If one cannot detect if a given machine is active or not, one cannot gain a precise value for M.

There are architectural features that limit or prevent smart correlation of R with the correct machine m_i — i.e. the assignment is essentially independent of each other assignment. One might assume that given the occasional relatively unique request the observer could associate all other activity originating from the same host as the unique message. This is not the case. First, there is no unique identifier carried by the messages with which to associate the messages. As described previously, the IP address of the requesting machine is not maintained in subsequent messages, only that of the resolver. One could investigate the queries that a particular public resolver has answered, but it is not clear that this aids identification of the end-user. Due to caching, the other requests will not pass the sensor at a time correlated with a known message. Since messages cannot be correlated based on time or identifier, the search space is not easily reduced from its astronomical size in the recommended pDNS architecture.

These technical features of passive DNS collection make it extremely unlikely that an observer would be able to correlate resource records with the originating machine. The built-in obfuscation and omission of data, as well as the hierarchical nature of DNS, the sensor placement architecture, and the vast, independent search space all combine for this effect.

III. Legal considerations of pDNS privacy impact

This section constitutes a legal survey of issues surrounding pDNS collection, however it is not a legal opinion. The first aspect to survey is whether DNS information or requests for it are considered personally identifiable information (PII). If so the privacy implications of collecting it would be significantly affected. Before discussing DNS as PII, PII must be defined. This is not an easy task, and no authoritative definition yet exists. See appendix C on page 8 for a summary of various authorities' definitions of PII. This section examines how well the different aspects of the DNS exchange are described by the PII definitions and possible legal concerns regarding pDNS besides PII.

A. Is the data collected through pDNS considered PII?

Two clear issues can be characterized with respect to PII and pDNS. First, do any of the types of information collected with pDNS constitute a type of information that is protected under definitions of PII? This question will be addressed by considering the traditional and NIST Special Publication definitions in the United States as well as international definitions, as discussed in Appendix C on page 8. Secondly, if the information as collected through pDNS, if it is PII, can be used to consistently link to an individual (or under some definitions, a small, well-defined group of people). As discussed in section II on page 1, pDNS can collect four types of information: DNS queries, DNS messages' positive answers, negative answers, and IP addresses of the parties to the communication.

First, with respect to DNS queries, it is unlikely that any part of the end user's name or identification number would be directly revealed absent willful user disclosure as described in appendix A-D1 on page 7. Therefore, queries do not constitute "personal information" under any U.S. State definition, nor do they constitute PII under any of the surveyed U.S. federal laws. However, queries could conceivably reveal the end-user's activities, which would be considered a type of information that falls under the NIST Special Publication's definition of PII. In this manner, queries may fall under some international definitions of PII, even being considered 'sensitive information' in some instances. This would only be the case if the queries could be linked to the parties of the query. Sections II-A and II-B discuss the technical feasibility of such identification in a properly instrumented collection environment, and finds it extraordinarily unlikely. Therefore it is unlikely that pDNS will constitute PII under any definition, due to this lack of ability to identify or link to the end-user.

Passive DNS also may reveal organizational IT information, as described in appendix A-C on page 7. This consideration is not generally of consequence to a discussion of PII, because business information is only considered PII if it can be used to identify an individual or sometimes a small group of people. The organization itself is not protected under PII legislation.

Passive DNS also collects negative and positive answers. Appendix A-B on page 6 outlines the multiple types of negative answers, noting that they are similar in character. Quite simply, negative answers do not reveal any information that is not revealed by queries, i.e. that something looked for some name. The statutes surveyed deal with information that could potentially identify an individual; the above analysis finds that queries will not, and negative answers do not contain additional information identifying an individual. Therefore, since queries are not considered PII, neither are negative answers.

Positive answers primarily return the host's IP address. Under the traditional U.S. State definition of personal information, IP addresses are unlikely to be considered "personal information" as it does not contain part of an individual's name in conjunction with an identification number or financial information. The host's IP address is also not likely to be considered PII under any current U.S. federal laws. However, in NIST SP 800-122, IP addresses can be consider PII, if it "consistently links" to an individual or "small, well-defined group of people" [6]. While it is usually a third party provider that is publishing the positive answers, it is possible that the positive answer IP addresses could be linkable to a small-well defined group of people. In cases where a small group or individual is publishing the positive answers, positive answer IP addresses could be analogized to business telephone numbers. Business telephone numbers are considered a type of PII under the NIST definition. SP 800-122 gives the example of an organization that publishes a telephone directory. NIST explains that while the information in the telephone directory is considered PII, the organization would not need to preserve confidentiality, but would be tasked with protecting integrity and availability of the information [6]. Therefore, even in the case where the IP address in answers is linkable to an

individual/small group, the information would fall under this particular category of PII, where confidentiality is not required.

As in the NIST definition, public information is often still considered personal data by other nations. For example, the EU states that when, "personal data are made public... the data subject is not deprived of protection... he is guaranteed such protection by law in accordance with the fundamental principles of the right to privacy." Again, as in the case of the NIST document, personal data in the EU can sometimes be shared, for example in a telephone directory [7]. The international opinion on the inclusion of IP addresses as PII is still under debate. Some nations, such as Switzerland, clearly protect IP addresses as personal data, when the address can identify an individual [8]; however this is not usually the case with positive answers. Therefore, positive answers will likely not be considered as PII under the EU's Article 29 Working Party, as it also states that an IP address is considered personal data if an individual can be identified [9].

Canada's Personal Information Protection and Electronic Documents Act would likely not consider A record's IP addresses as PII in most instances because business information is excluded from Canada's definition of personal information [10]. Overall, in most instances positive answers will not constitute PII. However, in a small number of cases, it is possible that positive answers could constitute PII, when an individual could be identified. Whether additional work can be done on that information will depend on the nation in question. Many guidelines to the collection and use of PII indicate that PII must be collected and used for a specific purpose. When considering whether an additional use is within that purpose, the question is, "what would a data subject have reasonably expected to happen to his or her data at the time the data were obtained?" [11]. This question may help to define whether or not the additional use of positive answers which could identify an individual would currently be appropriate as well as guide future treatment of the issue. It is also worth noting that the intended use of the data effects the determination of acceptable collection; for example, in the EU collection for scientific research has fewer restrictions.

A final consideration in this area is that positive answers are likely not the focus of PII statues, as illustrated in a recent pDNS study proposal from the EU [12]. The main issue of concern highlighted in this study proposal is clearly the protection of the end users' information, and not the information contained in the answer.

The final categories of information that can be collected from pDNS are both DNS server and client IP addresses. While DNS server IP addresses do not constitute PII because the addresses do not link to individuals, client IP addresses merit a more detailed discussion. As described above, IP addresses, while likely not considered PII by U.S. State or federal laws are specifically discussed by the NIST Special Publication 800-122. Under the NIST Publication, IP addresses constitute PII if the address can be linked to an individual or "small, well-defined group of people." In fact, many of the nations that contain broad definitions of PII would consider IP addresses as PII at least in some contexts.

The collection of user IP addresses requires some caution, as illustrated in a recent pDNS study proposal from the EU [12]. This proposal explicitly states that it would not collect user IP addresses as part of the study. Omission of these addresses is one factor which leads the authors to conclude that the collected data does not fall within the definition of personal data. IP addresses are considered personal data by the EU if the address can be linked to an individual, although not all member nations have followed this opinion. In addition, both Canada and Australia's governments have discussed IP addresses as qualifying as PII in some instances [13], [14]. However, it is questionable whether or not sender/receiver IP addresses, as collected through pDNS, can actually be used to identify an individual end-user, thereby constituting PII. As addressed in sections II-A2 and II-B on page 3, because the hierarchical resolution of DNS queries leads to a natural anonymization of the querying entity, it is unlikely an individual or small group will be identified if the sensor is maintained as described. As the user cannot be identified from their IP addresses as collected through a pDNS collection architecture, such IP addresses are likely not considered PII.

Finally, the NIST Special Publication 800-122 contains an example which may be relevant to some organizations. An organization, such as an employer, often has access to internal usage of IP addresses, access logs, Uniform Resource Locator (URL), and website information. 800-122 notes that the combination of this information would in fact be linked PII. While this information is considered by NIST to be a low level of harm if confidentiality is breached, it is still considered PII. One could imagine pDNS data may be one part of a similar set of linkable data. Therefore it is important to note when forming policies that the combination of different pieces of data can create linkable PII.

B. Other Legal Considerations

Even if the information collected through pDNS is not considered PII, there are a few other laws which may be applicable to the collection of information through pDNS. For example, the European Union regulates the collection and storage of traffic data, which is defined as, "any data processed for the purpose of the conveyance of a communication on an electronic communications network..." [15]. The data which pDNS collects is likely considered data processed for the conveyance of communications and thus these EU regulations may apply to collecting pDNS data. The traffic data directive requires that traffic data be erased or anonymized when it is no longer being used for transmission. As the procedure of resolving DNS queries leads to a natural anonymization of the querying entity this requirement may already be fulfilled, at least with respect to the end-user. In fact, the EU's directive on traffic data explicitly states that they find the regulations do not conflict with "procedures on the Internet as the caching in the domain name system." While the U.S. does provide a few regulations surrounding ISPs, the focus is more on the protection of the electronic communications themselves, rather than the data used to process the communications. In this way, the EU provides broader protections than the U.S. However, U.S. entities, the government in particular, should consider

the possible implications of the Electronic Communications Privacy Act, as amended by the U.S. Patriot Act, and in particular the Pen Register Act.

Another area of law with possible implications is the law on behavioral tracking, also known as 'e-tracking.' Nations are beginning to be concerned about business tracking consumers' online behaviors. In the United States, this area is regulated through the Federal Trade Commission, where they focus on the accurate implementation of organizations' privacy policies. In the EU, the e-tracking area is regulated through the E-Privacy Directive, which requires user consent prior to tracking [15], [16], [17]. However in both cases these regulations apply to web traffic, which more explicitly identifies end-users and keeps the state of their communications in addition, unlike DNS. Therefore, the applicability of new developments in this area of law is not likely to effect pDNS collection any further than the regulations regarding PII already discussed.

IV. Conclusions

In the recommended sensor configuration, the technical architecture of DNS prevents passive DNS collection from accurately identifying the machine ultimately connected to individual messages. Furthermore, individual messages are disassociated from end-user behavior, making behavior identification difficult, if not impossible, even if this message-to-machine connection were possible.

Even if both of these difficulties were removed, it is not clear that DNS activity would be legally protected. The definitions of what types of data are protected vary. In many of these cases, data is permitted to be collected given certain safeguards are implemented. Given the technical architectural attributes discussed, it appears that DNS happens to impose these safeguards as a component of its functioning.

Therefore, passive DNS data should be able to be collected as described without negatively impacting the privacy of the end-users who helped generate the data.

Appendix A
DNS record content

The content of the DNS record is a subset of the information in a DNS message. There are different issues regarding different aspects of this content. However, the privacy impact of such information is driven by the definition of what the DNS is. The domain name system is, first and foremost, a system for publishing and distributing the content of DNS records to anyone who asks for them.

A. Positive answers

The actual information in DNS records is public by definition. It is to be rendered upon request by the name server; issuing answers is its reason to exist [5]. These public answers are hardly optional, they are necessary for much of the Internet to function as we expect. This information is what is captured by positive answers. In addition to being public these positive answers are analytically useful; for example see [18], [19]. The content of DNS messages is a separate issue from determining the parties to a query, as discussed in section II-A2. Since the contents are already published by name servers, the positive answers are public.

B. Negative answers

A large proportion of DNS requests are never fulfilled, yet these queries are also of interest to analysts. When analyzing pDNS data, identifying evidence of non-existence, or negative answers, is an analytic objective. These unanswered queries do not contain inherently public information — since they are unanswered, their target is not in the DNS. It is proper behavior for the name server to indicate the non-existence of a query, however. In this respect, negative answers are just like positive ones — the name server's job is to publicly render the information upon request.

There are multiple types of negative answers and generally the information collected by any of these methods is similar in character. They can be inferred from unanswered queries or observed from NXDOMAIN messages. When using the DNS Security Extensions (DNSSEC), Next SECure (NSEC) records or NSEC3 records are used to authenticate non-existence. Since these are record types, operationally they appear as a type of valid answer to a query.

NXDOMAIN messages are an explicit notification that a domain does not exist. However, there is no assurance that an NXDOMAIN message is accurate or authentic, which there is with NSEC and NSEC3 records. Although the collection methodologies need to be different for NXDOMAIN and NSEC records, they carry similar information. Name servers could be configured to not issue a response for a non-existent domain, or may not issue one in error. Therefore, the set of negative answers collected by these responses may not be complete.

In the case of NSEC and NSEC3 records, non-existence is actually demonstrated by a valid, positive response in the DNS. This blurs the line between negative and positive answers. NSEC records captured this way must be considered public, just as positive answers are. They are served and stored exactly as any other public record type. This fact supports the classification of other types of negative answers as public, since the information content is essentially the same.

Negative answers do not need to be explicitly collected. If one has a complete set of positive answers and a set of queries, then all non-positive queries would receive negative answers. The set of non-existent domains is infinite, but pDNS collects demonstrations of non-existence. This is related to queries and positive answers. Since each query is reproduced in the positive answers, and the positive answers are publicly available, these negative answers are also publicly available. The identity of the querying entity is a separate issue, but the content of negative answer is not different in character from positive answers.

Negative answers can be collected via various techniques. These use the public positive answer set, and in some cases are a subset of the positive answers. Since negative answers are publicly deducible, there is no additional privacy impact to collecting them via pDNS.

C. DNS answers and network structure

One could be concerned about allowing collection of a complete set of positive answers on internal organizational structure. Importantly, this is not an end-user privacy issue at all, but it is an institutional issue that warrants being addressed. On the one hand, WHOIS records have published institutional information as a necessary part of registering for Internet names for many years [2], and the purpose of DNS is to advertise these internal addresses for external use. On the other hand, revealing organizational structure may seem analogous to problems which faced DNSSEC. DNSSEC's NSEC records allow for efficient enumeration of all the names in the zone by anyone who could query the name server as a side-effect of the desired function [20]. A DNS zone is composed of all the names for which one name server is responsible, and is usually defined in one zone file. By the letter of the RFCs this would not be an issue. However, there were enough practical issues with implementing NSEC to justify a new, obfuscated record format, NSEC3. Some organizations agree with the original RFC and operate NSEC records, however NSEC3 provides another option.

Emphatically different from the NSEC problem, pDNS does not allow zone enumeration by arbitrary parties. First, it is not on the demand of the observer — the observer can only collect those names which are resolved publicly. This ability to issue intelligent queries after each answer is a necessary feature of the concern about NSEC. That only public resolutions are collected is enforced by the organization itself. An organization chooses where to place the pDNS sensor, and should do so on the external or Internet facing interface of its DNS server, as recommended by [1]. As an additional measure, the collecting organization should provide the data only to trustworthy entities that enter an agreement not to abuse it. A similar practice has been used to share zone files, which is a requirement [21]for the generic top-level domains (gTLD) granted an anti-abuse and nondisclosure agreement [22].

Negative answers are often the result of misconfigurations somewhere on the network , which can provide a significant amount of network structure information. This configuration information does not contain PII, however it can be considered sensitive system configuration information. Malicious software on the network is another source of these unanswered queries. In general, organizations should monitor the DNS for such misconfigurations and infections in order to rectify them. However, there is conceptually no added exposure from pDNS due to this sort of data — it is already leaving the network if it is detected by pDNS on the border of the network. Autonomous System (AS) 112 is established to capture this traffic [23]. It has been demonstrated that traffic to AS112 publicizes such misconfigurations [24]. Given this current state of affairs, pDNS does not additionally expose the organizational structure.

D. On the users

In addition to barriers to identifying a particular machine with particular DNS traffic, as discussed in section II-A on page 1, correlating human behavior with the observed machine behavior is not straightforward. There are several factors that contribute to this additional barrier. First, if transient network addressing is used, such as network address translation (NAT) and/or dynamic host configuration protocol (DHCP), then attribution to an individual machine is impossible, without further information outside DNS, even if the IP address could be obtained. These services disrupt the one-to-one mapping between IP addresses and machines that identification relies upon.

Additionally, correlation to human end-users is obfuscated by the fact that multiple people may use one machine. Further, legitimate web pages may contain content from more than a dozen different domains, all of which the machine would need in order to issue a DNS query based on one user's actions. Determining which query was issued by the user and which by the machine as ancilliary results is not straight forward and further complicates attribution. Finally, many DNS lookups are completely automated, such as checking for software updates. The user may never have configured or consented to such checks; in the common case of malicious software, the user certainly has no hand in the queries the machine is made to issue.

Disentangling each of these layers would require information that is not contained in the DNS, such as what user was logged in to a machine or the router's historical NAT table. Such NAT state data is volatile — it will be lost at power down unless explicitly logged. To the author's knowledge, no one shares their NAT state data. Since these various information sources are not available, pDNS would be unable to attribute activity if any of these concerns come into play.

1) Willful user disclosure : There is user-generated content captured by pDNS. Since users are unpredictable, it is not possible to completely prevent user data from being captured. It seems impossible for the user to do it accidentally, which is why it is termed "willful." However, it is conceivable that a user would issue a query for a domain such as "Name:.John.Doe.SSN:.123-45-6789.home:.123MainSt.com." If the observer were collecting queries or negative answers it would enter the pDNS data. This is roughly equivalent to the user uploading the information to a public blog or even printing flyers and throwing them out the window. Any system which captured user input would have this problem. Users don't generally do this, and there doesn't seem to be any reason why anyone but the user would be held responsible for the user's odd behavior.

APPENDIX B
TTL STUDIES

To verify that TTLs are in fact long enough to practicably frustrate reconstruction of DNS behavior, TTL distribution was measured by using the Security Information Exchange (SIE), a common and relatively inclusive pDNS source [25]. The most common TTL values within unique resource records are summarized in Table I.

Table I represents a summary of pDNS data collected between March 18th and July 19th, 2011. The unique resource

Table I
MOST COMMON TTLS OBSERVED OVER 124 DAYS

TTL (s)	# of unique non-PTR rrsets	% of $1.75 * 10^{10}$ non-PTR records	# of unique PTR rrsets	% of $3.31 * 10^{9}$ PTR records
3600	2610500679	14.94%	374252094	11.31%
86400	2195700156	12.56%	1732873726	52.34%
60	1915912925	10.96%	1574403	0.05%
172800	1612894359	9.23%	141989511	4.29%
5	1192856851	6.82%	38011	0.00%
14400	983796852	5.63%	21261011	0.64%
300	752680349	4.31%	14496264	0.44%
7200	595854664	3.41%	298764407	9.02%
30	594663554	3.40%	156344	0.00%
900	505761436	2.89%	19118338	0.58%
20	450407183	2.58%	7000048	0.21%
600	397056674	2.27%	12800846	0.39%
10800	310257303	1.78%	40182773	1.21%
0	306139110	1.75%	509735	0.02%
1800	296136323	1.69%	9403114	0.28%
230	291800738	1.67%	586	0.00%
2100	235450087	1.35%	496	0.00%
1000	228648657	1.31%	300138	0.01%
1	222117469	1.27%	44826	0.00%
43200	173979556	1.00%	250701732	7.57%

record sets, as determined by the data fields of name, record type, TTL, and record data, were considered for each day. Then the TTL values were extracted from these records and counted. The total number of records considered was 20,788,711,654. Of these, it was found that pointer (PTR) records behaved significantly differently, and so the roughly 3.3 billion of these are listed separately. The median TTL is one hour (3600 s) for both all records and the non-PTR records. The median PTR record value is 86400 s (one day). The distribution appears to follow a power law, with the average TTL value, in seconds, for non-PTR records equal to 53140, for PTR records equal to 88815, and for all records equal to 58821. These results are consistent with the assertion that caching of records is a significant source of noise in pDNS data.

APPENDIX C
DEFINING PII

This appendix is a legal survey of relevant PII legislation. In any jursdiction both domestic and foreign law must be examined when considering if a set of DNS behavior constitutes PII. Currently, there is a lack of international consensus about the treatment of PII. In the European Union, the protection of PII is considered a fundamental right [26, Art. 8]. In contrast, the United States relies on a patchwork of state and federal regulations to protect PII. Meanwhile some organizations, such as the Asian-Pacific Economic Cooperation (APEC), focus on the practicalities of transferring PII between nations. The differences in the treatment of PII are reflected in the diverse definitions of PII established by different states, nations and nation groups. It is important to note that not all nations use the same terminology (e.g. PII, personal data, personal information), but the underlying purpose of protecting certain details about an individual provides the basis for comparison between nations. This section will present an international survey of definitions of PII, with an emphasis on categories of PII that may be relevant to pDNS collection.

A. United States

The United States' definition of what information constitutes PII may be evolving into a broader concept. However, the traditional view of what is considered protectable information about a person is still reflected both in state and federal laws [27, p 25]. With respect to state laws, forty-six states have data breach notification laws, and each of these laws contains a definition of "personal information" [28]. Although these definitions contain differences in the types of information included, a general trend in the definitions is observable. Often, states define personal information as part of an individual's name in combination with a social security number, a driver's license number or information for access to a financial account (such as a credit card number and access code), for example [29]. Beyond this basic information, some states include other identifiers in their definition of personal information. Medical data, birth date, employment information are all examples of additional identifiers used in some states such as [30], [31]. Interestingly, North Carolina considers Internet account numbers and Internet identification names to be personal information if and only if the information allows for access to financial information [32].

The United States also has federal laws which reflect what has traditionally been considered protectable information about an individual. However, these federal laws only cover certain entities and no federal regulation specifically addresses DNS data collection. The relevant laws include the Privacy Act of 1974, the Health Insurance Portability and Accountability Act of 1996, the Gramm-Leach-Bliley Act, and the Communications Act.

The Privacy Act of 1974 protects the privacy of an individual's records which are held by the federal government. The Privacy Act defines a record as a collection of information, such as education or medical history, which includes an individual's, "name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a photograph" [33].

The Health Insurance Portability and Accountability Act of 1996 (HIPPA) discusses individually identifiable health information. This type of identifiable information includes mental and physical health information as well as "common identifiers" such as name, address, and social security number [34]. The information is only considered individually identifiable health information if it identifies or could reasonably be used to identify an individual.

With respect to financial information, the Gramm-Leach-Bliley Act requires that financial institution protect nonpublic personal information [35]. Protected information includes the name, address, account number, and encompasses customer lists based on nonpublic personal information. Multiple exceptions exist that allow for the disclosure of nonpublic personal information under the Gramm-Leach-Bliley Act, including business uses within the institution and customer consent.

Another example of the traditional protection of personal information comes from the Communications Act. The Com-

munications Act requires telecommunications carriers to protect customer proprietary network information (CPNI). CPNI includes details of the call (time, date, duration), services to which the customer has subscribed, and account information (including social security number) [36]. While the terminology may differ among the laws, there are commonalities that constitute protectable information. Common identifiers such as name and social security number appear in all of the laws, in addition to financial and medical information.

B. U.S. National Institute of Standards and Technology (NIST)

Recent documents have illustrated that regulators in the United States may be considering a broader view of what constitutes PII [27]. In one indication of this view, the National Institute of Standards and Technology (NIST), which develops guidelines for federal agencies, published Special Publication 800-122. [6, sect. 2] defines PII as "any information about an individual maintained by an agency." This expansive definition of PII includes not only traditionally protected information (social security number, name) but also information that could logically be associated with an individual ("linked" or "linkable" information) such as race, religion or geographical indicators. In addition, the NIST definition of PII includes information that can be used to trace an individual's activities (such as audit logs). One type of PII that is relevant to pDNS data collection, IP addresses, is specifically noted in the NIST publication. 800-122 specifically states that an IP address can be considered PII if it "consistently links to a particular person or small, well-defined group of people."

C. European Union (EU)

The broader view of PII indicated in the NIST publication is more closely associated with the European Union's definition of PII than the United States' traditional view. The EU's Directive 95/46/EC, known as the Data Protection Directive (DPD), defines personal data as any information "relating to an identifiable natural person" [37, Art 2a]. Under the DPD, a person is identifiable either directly or indirectly by physical, cultural, mental, economic and social factors. In addition, the DPD specifically limits the processing of certain 'special categories.' These protected types of information include race, political opinions, union membership, religious beliefs, health information and sex life [37, Art 8].

Each member nation in the EU is responsible for the implementation of the DPD in the country's own laws. Spain, for example, defines personal data as "any alphanumeric, graphic, photographic, acoustic or any other type of information pertaining to identified or identifiable natural persons," excluding business information [38]. In addition to variations in definitions, member nation's courts have addressed the application of personal data to different types of information. One type of information with application to pDNS, IP addresses, has been a contentious issue. The EU's Article 29 Working Party has stated that IP addresses are considered personal data if an individual can be identified [9]. However, in the United Kingdom, static IP addresses can be considered personal data, but dynamic IP addresses are not [39]. Also, a German court has stated that IP addresses are not considered personal data [39].

D. Russia

Russia's federal law on personal data also contains a broad concept of PII, defining personal data as, "any information referring to a particular individual or which can be used to verify an individual identity (hereinafter, "individual concerned") including his/her surname; given name; patronymic; year, month, day and place of birth; address; marital status; social and property status; education; occupation; and income level" [40]. Similar to the EU's DPD, Russia also limits the collection of particular types of PII, specifically data about an individual's health and private life.

E. Canada

Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) also defines PII broadly; personal information is defined as, "information about an identifiable individual" excluding business information about an employee [10]. Similar to the EU's definition, the PIPEDA also includes a definition of sensitive information; however it is less specific than the EU's definition. In contrast to the DPD's explicit categories of sensitive information, the Canadian definition of 'sensitive information' is dependent on context, although the PIPEDA always includes medical and income records (not a category of special information in the EU's DPD).

F. South America

A push for harmonization with the European Union's DPD and subsequent legislation is taking place in South America in the form of the Ibero-American Network of Data Protection (RIPD) [41]. While the RIPD's objectives include fostering data protection, no laws specifically address DNS data collection. However, much more so than Europe and North America the laws in South America are not uniform and there are many draft laws still under consideration. The definition of what is considered PII also varies by nation. For example, Mexico's definition of personal information includes not only traditional information such as address and email but information such as religious, political, and philosophical beliefs [42]. On the other hand, the definition in Panama is somewhat more limited. In Panama, some information is considered "confidential" including medical information, criminal history, and correspondence including "audiovisual mail" [43].

G. Asia-Pacific Economic Cooperation (APEC)

Another multi-nation organization, APEC, presents an alternative model to data protection. However, while the model differs, the definition of personal data under the APEC model is similar to that of the EU model. APEC's definition includes, "any information about an identified or identifiable individual" [44]. This is also similar to [45]. With respect to APEC member nations, while no nation surveyed specifically address pDNS, they do address the protection of personal information. Australia's definition of personal information includes both

fact and opinion, regardless of source, on an individual that is identified or identifiable [46]. Australia's definition of sensitive information is similar to the EU's DPD, including sensitive categories such as race, beliefs, sexual orientation and health information. Japan is also an APEC member nation; its data protection law describes personal information as information that can identify a living individual by name, date of birth, or "other description" [47]. While it is unclear what constitutes other descriptions, Japan's definition does agree with most international definitions in including both information that can directly identify an individual and information that, when combined with other information, could identify an individual.

On the other hand, China, while a member of APEC, does not have a specific data privacy law from which to draw a definition of PII. However, China does maintain privacy protection in a few laws, including general protections for the privacy of electronic correspondence and network users (both with major exceptions) [48, art. 3]. In addition, a Chinese national law has been amended to provide stronger privacy protections for national identity cards, which contain information such as name, gender, and home address [49]. India, which is not a member of APEC, also does not have a specific data privacy law. India does, however, have the Data Security Council of India (DSCI), which is studying both American and European Union legislation and produced Best Practices for Data Protection [50].

H. Middle East and Africa

Considerations of PII in the Middle East and Africa are more difficult to discern, however there are a few data points available. With respect to Africa, the South African Bill for the Protection of Personal Information, which is not yet in force, defines personal information as information relating to an identifiable person, including race, gender, ethnicity, beliefs, criminal history, medical information, personal opinions, private correspondence, email address and any identifying number [51]. With respect to pDNS, it is possible that IP addresses would be included as an identifying number, if it could be used to identify an individual. With respect to the Middle East, Dubai's Data Protection Law of 2007 defines both PII, any information relating to individual, and sensitive information, including race, sex life, and religious and political beliefs [52].

ACKNOWLEGMENT

This work was created with the funding and support of the Department of Homeland Security under the Federal Government Contract Number FA8721-05-C-0003 between the U.S. Department of Defense and Carnegie Mellon University for the operation of the Software Engineering Institute, a Federally Funded Research and Development Center. CERT® is a registered mark owned by Carnegie Mellon University.

REFERENCES

[1] F. Weimer, "Passive DNS replication," in *17th Annual FIRST Conference on Computer Security Incident Handling*, 2005.

[2] L. Daigle, "RFC 3912 - whois protocol specification," *http://www.ietf.org/rfc/rfc3912.txt*, 2004.

[3] D. Plonka and P. Barford, "Flexible traffic and host profiling via dns rendezvous," *SATIN 2011*.

[4] R. Rasmussen, "Practical usage of passive dns monitoring for e-crime investigations," *SATIN 2011*.

[5] P. Mockapetris, "RFC 1034 - domain names - concepts and facilities," *http://www.ietf.org/rfc/rfc1034.txt*, November 1987.

[6] SP800-122, "Guide to protecting the confidentiality of personally identifiable information (PII)," NIST, Tech. Rep., April 2010.

[7] "Directive 97/66/ec of the european parliament and of the council of 15," December 1997.

[8] Switzerland, "Federal data protection and information commissioner," http://www.edoeb.admin.ch/aktuell/01688/index.html?lang=en.

[9] EU, "Article 29 data protection working party, opinion 1/2008," April 2008.

[10] Canada, "Personal information protection and electronic documents act (PIPEDA)," s.C. 2000, c. 5.

[11] Ireland Data Protection Commissioner, "case study 8/99, telecommunications company - electronic publication of telephone directory on the internet and cd-rom," http://www.dataprotection.ie/viewdoc.asp?m=&fn=/documents/caseStudies/99cs8.htm.

[12] R. Clayton and P. Ryan, "The legal compliance of a luxembourg-based passive dns project," iSBN: 978-2-87971-109-6 (In press).

[13] Office of Australian Privacy Commissioner, "Does anti-piracy = anti-privacy?" *Privacy Matters Newsletter*, vol. 3, no. 1.

[14] Office of Privacy Commissioner Canada, "Privacy in the social media age remarks," in *Legal Services of the Communications Security Establishment Canada*, april 20, 2011, http://www.priv.gc.ca/speech/2011/sp-d_20110420_cb_e.cfm.

[15] "Directive 2002/58/ec of the european parliament and of the council of 12," July 2002.

[16] "Regulation (ec) no. 2006/2004 of the european parliament and of the council of 27," October 2004.

[17] "Directive 2009/136/ec of the european parliament and of the council of 25, amending directive 2002/22/ec," November 2009.

[18] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "EXPOSURE: Finding malicious domains using passive DNS analysis," *Proceedings of the Annual Network and Distributed System Security (NDSS)*, February 2011.

[19] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou II, and D. Dagon, "Detecting malware domains at the upper DNS hierarchy," in *20th Usenix Security Symposium*, San Francisco, CA, 2011.

[20] B. Laurie, G. Sisson, R. Arends, D. Blacka *et al.*, "RFC 5155 - dns security (dnssec) hashed authenticated denial of existence," *http://www.ietf.org/rfc/rfc5155.txt*, February 2008.

[21] ICANN, *New gTLD Agreement Specifications*, http://www.icann.org/en/topics/new-gtlds/agreement-specs-clean-19sep11-en.pdf, September 19, 2011, specification 4, section 2.

[22] Verisign, "Tld zone file access program," http://www.verisigninc.com/en_US/products-and-services/domain-name-services/grow-your-domain-name-business/analyze/tld-zone-access/index.xhtml, May 11, 2011.

[23] "AS112 project," http://public.as112.net, June 19, 2011.
[24] E. Wright, "Security implications of anycast routing," Master's thesis, Carnegie Mellon University. Information Networking Institute, 2008.
[25] J. Spring, L. Metcalf, and E. Stoner, "Correlating domain registrations and DNS first activity in general and for malware," in *Securing and Trusting Internet Names 2011*, 2011.
[26] "Charter of fundamental rights of the european union. 2000/c 364/01," 2000.
[27] T. Shaw, *Information Security and Privacy: A Practical Guide for Global Executives, Lawyers and Technologists*. ABA, 2011.
[28] National Conference of State Legislators, "State security breach notificaiton laws," http://www.ncsl.org/default.aspx?tabid=13489, October 2010.
[29] Fla., "Stat. section 817.5681," http://www.leg.state.fl.us/Statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=0800-0899/0817/Sections/0817.5681.html.
[30] Ark., "Code section 4-110-101," http://www.lexis-nexis.com/hottopics/arcode/.
[31] N.D., "Cent. code section 51-30-01," http://www.legis.nd.gov/cencode/t51c30.pdf.
[32] N.C., "Gen. stat section 75-65," http://www.ncga.state.nc.us/EnactedLegislation/Statutes/HTML/BySection/Chapter_75/GS_75-65.html.
[33] The Privacy Act of 1974, "5 u.s.c. section 552a."
[34] The Health Insurance Portability and Accountability Act of 1996, "P.l. no.104-191 (42 u.s.c. section1302d-2)."
[35] U.S.C. section 6801-6809, Federal Trade Commission, "The gramm-leach-bliley act privacy of consumer financial information," http://www.ftc.gov/privacy/glbact/glboutline.htm.
[36] "Communications act of 1934, section 222(a)."
[37] "Directive 95/46/ec of the european parliament and of the council," 24 October, 1995.
[38] Spain, "Royal decree 1720/2007," modifying 15/1999.
[39] Law Group of Pinsent Masons LLP, "Data protection quarterly," no. 22.
[40] Federal Law of 27 July 2006 No. 152-FZ, "On personal data," unofficial translation: http://www.mofo.com/docs/mofoprivacy/Federal%20Law%20of%2027%20July%202006%20N152-, updated by Federal Law No. 261-FZ.
[41] "Ibero-american network of data protection," http://www.redipd.org/la_red/Historia/index-iden-idphp.php.
[42] Mexico, "Decree law issue of privacy personal for the federal district, number 434," 2008.
[43] Panama, "Law no. 6," January 22, 2002.
[44] Asian-Pacific Economic Cooperation, "Apec privacy framework," http://publications.apec.org/publication-detail.php?pub_id=390.
[45] Organization for Economic Co-operation and Development, "Oecd guidelines on the protection of privacy and transborder flows of personal data," http://www.oecd.org/document/18/0,3746,en_2649_34255_1815186_1_1_1_100&&en-USS_01DBC.html#part1.
[46] Australia Office of the Federal Privacy Commissioner, "Guidelines to national privacy principles," September 2001.
[47] Japan, "Act on the protection of personal information (act no. 57 of 2003)," http://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf, 2003.
[48] China, "Measures for the administration of internet e-mail services," http://www.chinaitlaw.org/?p1=print&p2=060604172055.
[49] Hunton and Williams LLP, "New chinese legislation includes provisions protecting personal information," http://www.huntonprivacyblog.com/2011/11/articles/new-chinese-legislation-includes-provisions-protecting-personal-information/, November 8, 2011.
[50] Data Security Council of India, "About us," http://www.dsci.in/taxonomypage/1.
[51] Republic of South Africa, "Protection of personal information bill," http://www.justice.gov.za/legislation/bills/B9-2009_ProtectionOfPersonalInformation.pdf, 2009.
[52] UAE, "DIFC law no. 1 of 2007," http://dp.difc.ae/legislation/dp_protection/#8.