Chapter 1

Aims and Scope

"Everything counts in large amounts"
Depeche Mode

1.1 Introduction

Finding information on the Web of today requires that one need only know the address of a popular search engine. With several keystrokes and the click of a mouse, a user can issue a query to one of the largest information repositories on the planet, only to have it processed and a response returned in what is often less than a second. Users of the Web have learned to expect nothing less from such services. Not only are they expected to be free, but there is incredible pressure amongst the bigger search engines to crawl as much of the Web as often as they can so as to be in a position to provide their patrons with the most relevant results possible (or at least better than the results offered by their competitors).

Maintaining a search engine that can scale with the complexity and vastness of a network the likes of the Web is not trivial. Indexing the Web requires that significant resources are made available to the search engine. At the very least, an effective search engine must have an army of crawlers (these are programs that index Web pages) which are constantly exploring new content in addition to updating their existing content. The cost and complexity of maintaining a system of this nature is such that there are only a few search engines in the world that can afford to be competitive. As a result, hundreds of billions of search queries belonging to hundreds of millions of online users are essentially funneling down to an extraordinarily small set of search engines (in the United States there are only five search engines of significance[1]).

The threat to privacy in this scenario is dire. On numerous occassions, privacy experts have highlighted these threats and warned of the potential for a privacy nightmare.

The more popular search engines have responded to these concerns with mitigations that are not apt. When perusing the privacy policies of two of the bigger search engines of today (Bing[2] and Google[3]), the mitigations provided in order to protect one's privacy (anonymising one's IP address and browser cookies after a period of time) contribute very little, if at all. Despite the legitimate concerns of the privacy community, the search engines have not engaged in a manner that is satisfactory. It seems that some do not even appreciate the gravity of the problem. Consider the following statement on privacy from Eric Schmidt, the CEO of the biggest search engine on the Web today (Google) in an interview with CNBC [20]: "If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first

[1] comScore Releases February 2009 U.S. Search Engine Rankings, http://wa.la/3E2

[2] Bing Privacy Supplement - http://privacy.microsoft.com/en-us/bing.mspx

[3] Google Privacy Center - http://www.google.com/privacy_faq.html

place."
Obviously, this statement is incredibly short-sighted. Whilst it may apply to some who are engaging in illegal activities and want it to be kept a secret, it does not apply to the individuals who have legitimate reasons for maintaining their privacy (those investigating antiretroviral drugs, abortion clinics, abuse et cetera).

The lack of commitment to privacy from the search engines has not gone unnoticed by the privacy community. There are a number of Privacy Enhancing Technologies (PETs) on the Web that aim to enhance and protect one's privacy. These include tools that are the result of projects conducted at universities as well as products shipped from multinational software corporations (the privacy options in Microsoft's Internet Explorer, for example).

In this work, we take a detailed look at privacy within the context of search. Given that there are a number of PETs which contribute to online privacy already, we ask if this is sufficient to protect one's search privacy. As we will discover, this is not the case. Despite the PETs that are available, when one considers the nature of search and what privacy in this context means, there is still work to be done. Upon identifying the search engines as entities that cannot be entrusted with one's search privacy, the work required can not come from them.

As a result, we will investigate potential solutions to the search privacy problem that do not rely on the cooperation of the search engines.

This work will contribute the following:

- We look at each of the Privacy Enhancing Technology (PET) categories and try to find an existing solution to the search privacy problem, i.e., using existing technology, can the search engine be eliminated as a threat to search privacy?

- Having found that the search engine remains a threat as long the original use case scenario for search is maintained, we confirm that users are sharing search queries in a manner that facilitates a network acting as a cache of search queries and results. We describe this network and discuss its contribution to preserving the search privacy of its users.

1.2 Outline

As mentioned above, the primary contribution of this work is the proposal of a search network. We build up to this by evaluating each PET category in terms of what it can deliver to search privacy. If the search engine can be eliminated as a threat to privacy using existing means, then the problem is solved. Of course, as we will see, this is not the case.

This document is structured as follows:

- Chapter 2 begins with a discussion of the definitions available for privacy. We don't try to redefine or provide our own definition in this chapter. Instead we arrive at an existing and applicable definition given the nature of the Web. We then discuss privacy of one's identity (anonymity) and move on to look at each category of PET available today.

- In chapter 3 we discuss search in more detail. Specifically, we look at its evolution through time and analyse the three key contributors to the search scenario of today: search users, search engines and crawlers. This leads to a discussion of search profiles, both online (the Web) and offline (the library). After highlighting a search privacy violation via AOL in 2006, we conclude this chapter by identifying which categories of PETs may be applicable in addressing the search privacy problem.

- Anonymity is the first category which showed promise in addressing the problem. In chapter 4 we look at this category and describe an attack against crowds based upon two principles, that of unlikely senders and recognising related requests. We then present and discuss the results of a simulated attack.

- We turn to the category of personal control in chapter 5. The Platform for Privacy Preferences (P3P) embodies the definition of this category and is examined in detail. We highlight two problems with this technology, that of trust and proxies, and investigate possible solutions.

- Private communication through encryption and obfuscation is examined in chapter 6. We take a closer look at technology which uses obfuscation in a bid to preserve one's privacy. We then discuss several problems with this approach and present countermeasures that can be taken by the search engine.

- We begin chapter 7 with an analysis of the database released by AOL in 2006. We use this to confirm that users are sharing a significant amount of search queries. We then leverage off of this fact and propose a search network built upon a Distributed Hash Table with the primary objective of preserving the search privacy of its users.

- This work is concluded in chapter 8. In light of the contributions made, we ask if search privacy has been enhanced and whether or not the search engine is still a threat to search privacy.

Chapter 2

Privacy Enhancing Technologies

Privatus - not in public life, past participle of privare, to release, deprive, from privus, single, alone

2.1 Introduction

In this chapter we examine several technologies that allow users to control their information on the Web; these are typically referred to as Privacy Enhancing Technologies (PETs). PETs range from simple standalone email delivery solutions to complex systems built upon multiple layers of cryptography and implemented across servers around the world.

Before presenting an analysis of technologies that preserve our privacy, we ask: what is privacy? This question may seem simple, but the answer is by no means trivial. In fact, for a term that almost everybody recognises immediately, it is somewhat surprising to note that there is no globally accepted definition [36, 120, 79].

The first goal of this chapter is not to answer this question ourselves, i.e., we will not offer our own definition of privacy. Instead, in the next section we discuss previous research in the field of privacy and look at some of the definitions that have been offered over time. This leads us to a definition of privacy that suits our needs when discussing privacy in the context of PETs on the Web of today.

Having made it clear what is meant by the term privacy when speaking of Privacy Enhancing Technologies, we then provide a brief discussion of anonymity followed by an analysis of some existing PETs.

2.2 Privacy

One of the earliest publications on privacy is an argument for the right to it by Warren and Brandeis [119]. They discuss the broadening of legal rights and equate the right to life with their definition of privacy: "the right to be let alone". Written during the rise of the printing press and the inevitable mass distribution of photographs, the authors emphasise that under common law, each individual may determine "to what extent his thoughts, sentiments, and emotions shall be communicated to others".

Introna et al [67] argue that this definition is far too restrictive. They discuss two counter examples: (1) if one were to use a telescope to observe your movements from a distance, this would in fact be leaving you alone, but is obviously not respecting your privacy and (2) there are certain entities/individuals that have a right not to leave you alone, for example, the tax service or your creditors. The authors comment that although privacy, or the absence thereof, is generally easy to identify, it is difficult to define. Even when given a definition, a counterexample can always be found. In arguing that conflicts between different stakeholders of a privacy claim may result in different conceptualisations of privacy, Introna et al describe the notion of privacy as “the freedom from the judgement of others”.

Solove [112] argues that “the right to be let alone” definition of privacy is far too broad a definition of privacy. In deconstructing the “nothing to hide” argument against privacy, Solove discusses the difficulties inherent in attempting to find a common core in privacy: if a conception of privacy is too broad, it may encompass too much and become unusable, alternatively if it is too narrow, it may be too restrictive and suffer the same fate. Solove proposes that instead of trying to conceptualise privacy using traditional methods (finding a common denominator), we must instead “understand privacy as a set of family resemblances” [111]. Furthermore, Solove notes that “privacy is not reducible to a singular essence; it is a plurality of different things that do not share one element in common but that nevertheless bear a resemblance to each other.”

Boyle [24] notes that privacy is an overwhelmingly large and nebulous concept. Having reviewed existing privacy literature, he still found himself confused as researchers used the term “privacy” to refer to different concepts. DeCew [36] notes that discussing the concept of privacy is difficult because on the one hand it appears we value it as a means to provide us with a sphere within which we can operate free from the interference from others, yet on the other hand this can be exploited to hide domination, degradation or physical harm.

McParland et al [84] point out that conceptual and operational confusion around privacy has resulted in researchers narrowing down definitions of privacy to fit their field of interest in a bid to avoid some of the confusion and, as a result, the confusion surrounding privacy remains undiminished.

Bellotti and Sellen [116] point out that any definition of privacy must be dynamic. They then define privacy as a subjective notion which may be influenced by culturally determined expectations and perceptions of one’s environment. Gavinson [55] describes privacy as the protection from being brought to the attention of others and notes that upon being brought to the attention of others, this will ultimately result in a further violation of privacy (due to further scrutiny).

Westin’s definition [120] (arguably one of the most popular) places emphasis on control over an individual’s information and describes privacy as “the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others.” This definition has lead to the widely used term “informational self-determination” and, as Ng-Kruelle et al [86] point out, is widely used in privacy related legislations.

In an argument against the right to privacy, Thomson [115] points out that one can lose *control* over one's information and yet still maintain one's privacy. She discusses an example where a neighbour has an X-ray device which has the ability to see through walls. If you were to come home and close your doors, the fact that the X-ray device exists next door means you no longer have control over your privacy, but your privacy is not violated until the device is used. She goes on to suggest that the right to privacy is itself a cluster of rights and that before we mark off something as a violation of the right to privacy, we should examine whether or not it is the right to something else that has been violated (or perhaps no violation has occurred at all).

2.2.1 Privacy in this research

The Web is founded upon a request/response protocol: a client makes a request to a Web server, this is processed and a response is then provided. Requests to a Web server can vary tremendously: sending/receiving email, shopping, viewing the news, watching a movie, looking up football scores, commenting on a paper or talking to a friend for example. Regardless of the activity, the smallest interactions on the Web result in the ability to store information related to each request. The most basic type of information includes the source from which a request originated (the IP address). Although an IP address only identifies a source of a query for a particular request at a specific time, it can be deemed to be information owned by an individual in the case where the IP identifies the individual (in the event of statically assigned IPs for example).

Consider the case of an individual who considers it a privacy violation to track his browsing habits across a particular site. Storing which pages were requested at what time, and by which IP would in effect compromise this user's privacy because the server would be indirectly tracking him. In this scenario if the user wanted to browse in private he would have to have a say over (or at least be aware of) which of his information is being stored and to what extent, i.e., he would be exercising control over his own information. By him not having control over his information, the server would keep logging his IP and he would not have the benefit of privacy.

As has been pointed out, the notion of control over one's information is emphasised in Westin's definition of privacy. Within the context of the Web and PETs, this can be succinctly captured as *the ability of the individual to protect information about himself* [58].

2.3 Anonymity

Privacy is not anonymity. Within the context of the privacy definition we have chosen to follow, anonymity is a form of privacy in the sense that one has control over one's identity [58]. Activities conducted in an anonymous fashion may still result in a violation of privacy.

Pfitzmann and Koehntopp [94] define anonymity as "the state of being not identifiable within a set of subjects"

Various papers [80, 58] discuss the positive and negative aspects of anonymity on the Internet. In favour of anonymity is the extended support for members of support groups (rape victims, recovering alcoholics and others), whistleblowing, refereeing for academic conferences, anonymous tips to investigative journalists et cetera. Disadvantages of anonymity include exploiting email services to spam the masses, launching massive denial of service attacks and illegally distributed copyrighted software.

Goldberg et al [58] divides anonymity into two categories:

> Persistent anonymity - this has a user making use of a pseudonym which can be used to link a number of transactions to each other, but not to the person behind the pseudonym. This may include communicating via email under a different identity.
>
> One-time anonymity - refers to independent transactions that cannot be linked together. An example may include commenting on blogs through an anonymisation service under randomly chosen names.

Pfitzmann et al [95] discuss exactly what forms of anonymity are possible on a network:

> Receiver anonymity - referring to the receiver of a message. The sender may be known, the message itself may be observed but the receiver of the message is anonymous.
>
> Sender anonymity - referring to the entity from which the message originated. This is the same as receiver anonymity except it applies to the sender only.
>
> Unlinkability of Receiver and Sender - this form of anonymity hides the relation between the sender and the receiver of a message.

There are a number of anonymity models that provide anonymity in one form or another. Of the most popular is that of the mix machines. Chaum

[26] employs the use of mix machines to delay the delivery of encrypted messages so as to hide the source of the message (typically used for email).

The mix approach does not require a trusted central authority. In fact, it is assumed that anyone in the system may learn the source, destination and representation of all messages in addition to having the ability to insert, modify or remove messages. The mix system has its roots in public key cryptography and begins with a user attaching a random string of bits to the message being sent and encrypting this with the receiver's public key. This is then combined with the receiver's address and another random string of bits, encrypted with the mix machine's public key and delivered to the mix machine.

The aim of the mix machine is to deliver the message to B in a manner that does not reveal that it came from A. Since anyone on the mix network can observe the source and destination of all messages, the mix machine cannot simply forward the encrypted message to B since any observer would note that a message from A went to the mix, and onto B. As a result, A must be communicating with B. Instead the mix machine only outputs messages in batches where each message is uniform in length and ordered in a manner that cannot be predicted. While a single mix machine is thought to be ideal, in practice a number of mix machines feed into one another to make for mix cascades. This cascade of machines makes it significantly more complex to perform traffic analysis. Wright et al [122] provides an overview on a host of different mix implementations.

2.4 Privacy Enhancing Technologies

In this section, we discuss a number of PETs on the Web today. Each PET is categorised into one of the key PET categories identified by Olivier [89, 91]:

> Private communication - this category consists of technologies that provide the ability to communicate content only to the recipient(s) specified, regardless of who is listening.
>
> Anonymity - these technologies are either based on Chaum's mix or implemented in one way or another through the notion of a proxy.
>
> Personal control - defined as "the use of technology to ensure that an individuals personal information is only used in a manner commensurate with the individuals privacy policy."
>
> Organisational Safeguards - similar to personal control, but focused on the organisation: "the use of technology to ensure that the organisation

> complies with its own privacy policy as well as the preferences of the individual".

There are a number of papers that provide an overview of PETs [64, 58, 56, 57, 82, 109]. Each of the technologies discussed falls into at least one of the categories listed above.

2.4.1 Private Communication

Olivier [89] highlights private communication as fundamental to the other categories since the absence of it would weaken almost all of the remaining solutions. For example, consider the case of sending your medical details online to your doctor. When these details are not encrypted, the fact that your doctor has privacy controls in place to protect your data and your privacy is of little consequence if your Internet Service Provider is logging all content processed through it (and in turn logging your medical data).

Goldberg [57] highlights the following private communication technologies:

> Pretty Good Privacy (PGP) [108] - widely available software that leverages off of public-key cryptography [103]. It is popular in email because it allows a sender to easily encrypt/sign the contents of an email in a manner that guarantees: (1) only the intended recipient(s) can read the email, (2) the integrity of the message is preserved (encrypted or not) and (3) non-repudiation from the sender's side, i.e., only the sender could have written the message.

> Secure Sockets Layer (SSL) [51, 118] - a predecessor to Transport Layer Security (TLS) [40, 41]. Since there are a number of unknown entities that may be involved in delivering messages from a Web server to a client (routers/transparent proxies et cetera), SSL and TLS are communication protocols supported by most popular browsers that guarantee the integrity of messages between applications on a network in addition to encrypting its contents and authenticating the parties involved (typically only the Web server is authenticated).

> Off-the-Record Messaging (OTR) [23] - in contrast to PGP, OTR is a communications protocol which uses encryption to provide repudiation, forgeability and perfect forward secrecy [61]. Perfect forward secrecy guarantees that the key used for a session cannot be learned if the public/private key set from which it was derived is compromised. Repudiation is provided because messages between parties are never

proved to be from any single party (digital signatures are only used in creating a session). Forgeability is delivered in the form of malleable encryption, i.e., ciphertext can be modified to make meaningful changes to the plaintext even if the key used to generate the ciphertext is not known. These properties ensure that once a conversation is over, nobody can verify the authenticity of any of the messages and since the messages could have been forged, the participants are guaranteed confidentiality. An example of where these properties would be of interest is the scenario where journalists are communicating with informants.

Olivier [89] notes that steganography is receiving renewed interest as a private communication technology. In a paper clarifying the field of steganography, and discussing the limits thereof, Anderson and Petitcolas [9] point out that steganography is not to be confused with cryptography. The latter transforms a message so that it is indecipherable to those who intercept it whereas the former deals with hiding the very existence of the message so that it is only known to the intended recipients. A simple example of steganography includes setting the least significant bits of image pixels to the bits of the message one wants to hide [117].

Howe et al [66] focus on the privacy of a user's search profile and introduce the TrackMeNot (TMN) extension to the Firefox Web Browser. Essentially, TMN assumes that the organisation that will be receiving and storing a user's search queries cannot be trusted. As a result, the aim of this technology is to thwart the logging activities of major search engines through obfuscation. This is achieved by issuing a number of false search queries over time on the user's behalf. In doing so, the real search queries issued by the user would form a small subset of the total queries issued, supposedly leaving the search profile unusable in so far as applying to the user it represents.

TMN has no third party dependencies and uses dynamic query lists to generate queries that are unique to the user in question. The component starts with a dictionary of random terms, seeded from a public source that is then used to issue queries. The search result of each query is analysed and a subset of terms is non deterministically extracted and used to replace terms already in the dictionary. As a result, the terms used and queries generated become unique to the user in question, in addition to looking very much like a real user that is searching. Although the statistical significance has not been proven, at the very least this extension generates a significant amount of noise in a manner that would make extracting the genuine queries (those explicitly issued by the user) cumbersome.

2.4.2 Anonymity

Federrath [47] illustrates the development of modern PETs through a timeline which begins with two important contributions:

- 1978, Public-key encryption. A technology that has already featured in the private communication category, this form of encryption is an important building block in anonymising technologies.
- 1981, Chaum's mix. Discussed in a previous section, most anonymity models are based either on the mix or the notion of a proxy.

Since Chaum first proposed the mix, there have been a number of anonymisation technologies modelled on top of it and the addition of a proxy [99, 42, 100, 80, 33, 1, 69, 100, 90]. The technologies discussed in this section form a small and well known subset of this literature. As a result, we only discuss core concepts and provide relevant references to the reader in the event that more detailed information is required.

Anonymising proxies

Levine and Shields [80] define a proxy as a single server that accepts connections from an *initiator* of an anonymous connection (party A) and forwards them on to the *responder* (party B). The assumption made when dealing with an anonymising proxy is that it will deliver messages from party A to party B but will not disclose to party B that party A is the source, i.e., party B will see the anonymising proxy as the source (this is sender anonymity, as discussed in section 2.3).

Of course, an implicit assumption is that party A trusts the anonymising proxy since although A is anonymous to B, he/she is not anonymous to the anonymising proxy. The Lucent Personalized Web Assistant (LPWA[1]) [52] and Anonymizer.com[2] are technologies built upon this premise. Anonymizer.com is simple, easy to use and, as Goldberg [57] remarks, is one of the few commercially successful anonymity technology providers. Anonymous-proxies.com[3] on the other hand, maintains a private database of anonymous proxies on the Web which is accessible upon paying a monthly fee.

[1] LPWA, available online at http://www.bell-labs.com/projects/lpwa

[2] Anonymizer - Anonymous Proxy, Anonymous Surfing & Anti Spyware, available online at http://www.anonymizer.com

[3] Available online at http://anonymous-proxies.com/index.html

Crowds

Much like anonymising proxies, crowds [100] offers various degrees of sender anonymity through the notion of blending into a crowd. The degree of anonymity, with respect to the sender of a message, is defined as an informal continuum ranging from *absolute privacy* to *probable innocence* and eventually *provably exposed* (refer to [39] for more detail on quantifying anonymity).

Each member in a crowd (a jondo) receives and forwards messages on behalf of any other member in the crowd, i.e., messages are proxied from one jondo to another. When a jondo decides to send its first message via the crowd, a random path must be configured to serve as the route taken for all messages sent by that jondo from that point onwards. The route is established as follows:

1. The sender picks a random jondo and forwards the message to him.

2. This jondo then flips a coin which determines whether to forward the message to yet another jondo (in which case this step is repeated) or to send the message to the intended recipient.

Upon receiving a response from the intended recipient, the message is routed back along the path to the initial sender of the message.

The authors of crowds define three types of attackers:

- A local eavesdropper is defined as someone that can monitor all communication from a single computer (jondo). Though no sender anonymity is offered from this type of attack (since the attacker can see all messages initiated from the computer), Reiter and Ruben describe how receiver anonymity (the end server) tends towards *beyond suspicion* as the size of the crowd tends towards infinity.

- Collaborating crowd members are seen as jondos that pool their information regarding the crowd together in an effort to expose crowd members. An attack is described where collaborators on a path of jondos used to send a message try to determine if their immediate predecessor on the path is the sender. Reiter and Rubin conclude that *probable innocence* is guaranteed for sender anonymity as long as $n >= 3(c + 1)$ where n is the size of the crowd and c the number of collaborating jondos. Wright et al [124, 123] formally define this type of attack (the *predecessor attack*) and examine it within a broader scope. In [125], Wright et al formally prove that the predecessor attack is successful against all existing anonymous protocols.

- The end server: Reiter and Rubin make a strong case for sender anonymity against an attack from the end server. They suggest that the anonymity of the initial sender of the message is *beyond suspicion*.

Onion routing

Onion routing delivers private communication within the context of a public network [99, 59]. Sender anonymity is provided in the form of a dynamic path of mix machines between the sender and recipient. A data structure representing an onion is made up of a number of encrypted layers and constructed by the initiator of the message. Each layer of the onion represents a mix machine that will form part of the path that the message will take from the sender to the recipient. As the onion makes its way through the network, each router on the path extracts a layer away in order to determine where to send it next. Encryption at each layer ensures that no router on the path can know the full path that the message will take.

Tor[4] is an example of an anonymising network that is based on second generation onion routing [42]. In 2006, it was estimated that Tor had approximately 450 server nodes participating in its network [92], although as of November 2008 Tor Status services[5] pointed to as many as 1,239 nodes on the network.

Essentially, Tor provides a high degree of unlinkability between the sender and receiver of a message even in the case of compromised mixes. Note that Tor is susceptible to predecessor and intersection attacks. In an intersection attack, the attacker repeatedly tracks which nodes are active through time in an effort to determine who is communicating with which recipients. By exploiting the fact that onion routers may fail or leave the network, an attacker knows that any path of communication that is still functional after a node has left, does not include this node. Similarly, new additions to the network will not be active in any existing paths. Berthold et al [22] provide a comprehensive list of the types of attacks that are possible against mix-based networks the likes of Tor.

2.4.3 Personal Control

We highlight several technologies that provide personal control using separate methodologies: P3P, IE8 InPrivate Blocking and Cookie managers.

[4] An anonymous Internet communication system - http://tor.eff.org/

[5] Available online at http://torstatus.blutmagie.de/

P3P

The Platform for Privacy Preferences (P3P) framework allows Web users to automate the protection of their privacy. Web sites publish P3P policies clearly describing what they intend to do with the data they collect. Web users then compare these policies to their own set of privacy preferences. Users can express their privacy preferences either through the APPEL preference language [78] or through the XPath-based preference language [6]. Provided that the P3P policy published by the Web site is acceptable (they align with the user's privacy preferences), the user may browse the site without interruption.

P3P has been the focus of much research and criticism [30, 29, 32, 65, 114, 104]. Common criticism of P3P includes that it is too hard to use and that there is no means of enforcement, i.e., a Web site with suspicious privacy practices has no incentive to declare what it is doing. Unless legislation codifies privacy protection of this nature, sites with bad privacy practices can simply mimic the policies of sites with good practices. With this in mind, the World Wide Web Consortium (W3C) describes P3P as a technology that is "*complementary to laws and self-regulatory programs that can provide enforcement mechanisms*" [31].

IE8 InPrivate Blocking

The Web has been designed in a manner such that navigation to a single URL may result in fetching content from a number of disparate sources. For example, loading the Wall Street Journal[6] in a browser will result in content delivered directly from the Wall Street Journal's Web servers in addition to references to at least two separate sources for advertisements. Typically, the browser loading the site will then make requests to each of the ad servers for the content in question.

Although primarily not an advertisement blocker, the IE8[7] InPrivate Blocking feature scans for third party sources that are referred to a predefined number of times during the course of navigation through the Web (advertisements typically fall into this category). If a single source exceeds this threshold, then it will be marked as having the ability to track a user's surfing habits (by logging his/her IP) and, as a result, although Web sites may still refer to them for content the browser will no longer request content from these servers. Figure 2.1 illustrates this process.

[6] Available online at http://online.wsj.com

[7] IE8, available online at http://www.microsoft.com/ie8

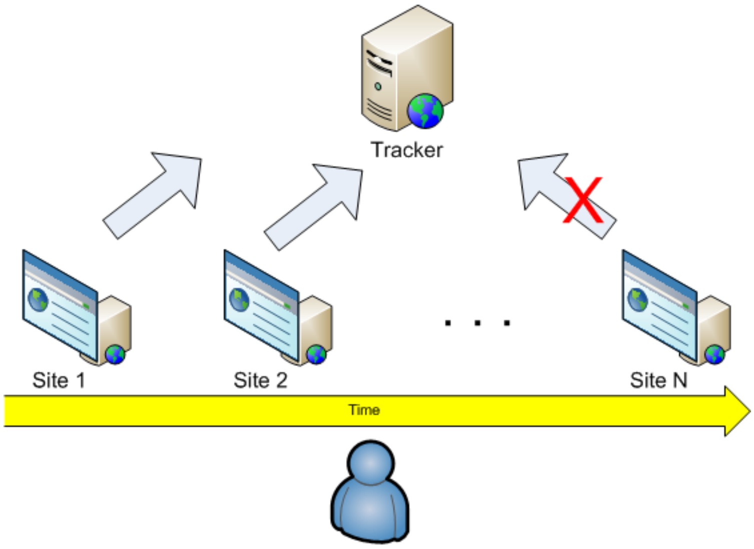

Figure 2.1: A user browses the Web and navigates to several different sites. Each of these sites refer to the same third party Web server. This could be for third party content, counters, advertisements et cetera. In this example, the IE8 InPrivate Blocking feature is enabled and has a threshold of N. On the Nth occurrence of referring to the same third party site from a number of different sites, the feature marks the third party site as being in a position to track the user in question. Subsequent requests to this third party are therefore no longer addressed, i.e., when a Web site refers to this server for content, IE8 will not retrieve it.

Cookie managers

Cookie managers grant control to the user in so far as determining which cookies are accepted from a Web server and how long they will be valid for. Cookies are a means for Web servers to persist small bits of information to a user's machine via the Web browser. This information can be used to enrich the Web experience. Examples may include setting a cookie that is a unique identifier which links to a shopping cart of items on the Web server. A user returning to a Web site need not restock the shopping cart since the Web server will be able to match the user to a cart through the unique identifier.

In analysing different types of PETs, Hall [64] describes cookies as neither good nor bad. Much of the criticism surrounding cookies is targeted at the nature in which they can be persisted, i.e., they can be written to disk without the user's knowledge or approval.

Cookie managers are built into most modern day Web Browsers and can be configured to include the following:

- deny all cookies
- prompt the user to accept/deny all cookies
- accept certain cookies by default
- clear all cookies upon closing the browser

2.4.4 Organisational Safeguards

Organisational safeguards serve as a means to ensure that organisations adhere to the commitments made in their privacy policies.

Consider the Internal Revenue Service (IRS) as an example. To safeguard the tax returns of tax payers, tax assessors need not be granted access to all tax returns all of the time. The organisation could implement mechanisms that only allow tax returns to be assessed by assigned individuals and only for the period for which an assessment is valid. Of course, there are exceptions (in the event of an audit, for example). Furthermore, in so far as issuing tax refunds to tax payers, the department responsible may only require access to their bank details and not the tax return in question. As a result, the IRS could store the bank details in one database and the tax returns in another, with separate DBAs assigned to each.

Olivier [91] highlights a common reservation with any of the technologies that fall into this category: the doubt that businesses, unless forced, will implement them at all.

E-P3P

Using the Web, organisations can declare their privacy policy in a machine readable fashion (via P3P files) or in a separate section of their Web site (typically in natural language). The privacy policy of an organisation essentially states which data will be collected from the individual and what the organisation intends to do with it.

The presence of a privacy policy alone does not guarantee that data collected from the individual will be used as promised. Not even P3P provides the means to enforce this within an organisation. Once data from the individual has been collected, it is up to the employees in the organisation to ensure that the data is used in a manner that is in accordance to the organisation's privacy policy. Given the size and complexity of some organisations, this is not always possible.

With this in mind, the Platform for Enterprise Privacy Practices (E-P3P) provides a system within which the promises made by an organisation, from a privacy perspective, can be formalised and enforced internally [75, 14, 18]. Essentially, E-P3P provides a means to automate an organisation's privacy guarantees by ensuring that data flow and usage practices within the organisation that involve an individual's data never conflict with the organisation's privacy policy.

Hippocratic databases

Agrawal et al [5] use the Hippocratic Oath as a basis for arguing that databases should assume responsibility for the privacy of the data maintained within them. They propose re-architecting databases to include the privacy of data as a fundamental tenet implemented through ten core principles. Essentially, the individual that owns the data and the database that collects it must specify privacy policies that detail the purposes for which data can be collected/used in addition to the recipients that can use parts of it. Agrawal et al note that Hippocratic databases go hand in hand with P3P in so far as the enforcement of an organisation's privacy policy. The XML structure of a P3P policy essentially describes what data will be collected, by whom, for what purpose and for how long. This notion of purpose and retention in P3P are directly applicable within the context of Hippocratic databases.

Rutherford et al [107, 106] take the notion of Hippocratic databases further when highlighting the privacy concerns inherent in log files. They then investigate the extent to which Hippocratic database principles can be applied in a Hippocratic log file architecture.

2.5 Conclusion

We began this chapter by highlighting a number of definitions for privacy. Having presented the notion of privacy as a concept for which there is no globally accepted definition, we eventually presented a clear definition of what is meant by privacy within the context of this research: " *the ability of the individual to protect information about himself.*"

After discussing anonymity as a form of privacy, we then discussed a number of PETs available on the Web today. Each technology presented comes with its own strengths and weaknesses. In the next chapter, we highlight these characteristics in an effort to identify which PET best addresses the search privacy problem.