The Grey Web: Dataveillance Vision Fulfilled through the Evolving Web

Richard Gomer
University of Southampton
Southampton, UK
r.gomer@soton.ac.uk

Natasa Milic-Frayling
Microsoft Research
Cambridge, UK
natasamf@microsoft.com

m.c. schraefel
University of Southampton
Southampton, UK
mc@ecs.soton.ac.uk

ABSTRACT

Over the past three decades, Web has evolved from an information medium to an intricate economic ecosystem. Initially focused on supporting the transition from traditional business practices to e-commerce, the Web has given rise to new, purely Web based businesses. Aligned with the original vision and expectations of the 'free Web', they have provided free services but, over time, developed business models that leverage the user digital footprints and the user generated content to create economic value. With the use of computing technologies to analyze, aggregate, and share such data, individuals' privacy has been undermined and, with that, the their ability to shape their role in the digital society and beyond. The purpose of this paper is to instigate the dialogue around the critical societal issues that arise from the current Web economy and motivate research initiatives to assist with addressing them. We present three case studies that quantify the extent, rate, and pervasiveness of the user tracking on the Web. We use them to illustrate the determining aspects of the Web that have to be taken into account by the Web Science community. As researchers we aspire to understand the nature of the Web in depth and, based on that, propose designs and policies that are required to ensure that the Web is fit to be the underpinning of our societies and our digital future.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Commercial services; Web-based services; Sharing; K.4.1 [**Public Policy Issues**]: Privacy.

General Terms

Economics; Human Factors, Legal Aspects.

Keywords

Privacy; security; surveillance; dataveillance; policy; economics; regulation.

1. INTRODUCTION

Recent news about the degree of the National Security Agency (NSA) surveillance of the Web has caused international concern and challenge to the degree of governmental observation of our lives. The Web founder, Tim Berners-Lee has called for "better protection of internet users privacy" in a variety of fora[1], calling "spying" a "betrayal" of the Web[2]. We share the sentiment and fully support any effort to curtail such practices. We, as Web Science experts and affected citizens, however, want a deeper dialogue around the issue. This paper offers evidence towards this discussion. For example, if we consider the assertion of betrayal of "the Web", we must explore what is meant by "the Web" for which spying is a betrayal and, further, if "mass surveillance" is the hallmark betrayal of that entity, hasn't it already happened? We will show that the Web has evolved from its early days as a document web into its current state as a surveillance web, and that in the context of its growth and evolution, this outcome is both inevitable and unlikely to change without either radical intervention or paradigm shifting invention, neither of which seems currently likely.

Indeed, under the leadership of Berners-Lee, the World Wide Web Consortium has sponsored the "Do Not Track" (DNT) initiative.[3] DNT was seen as a voluntary standard proposed to ward off the Federal Trade Commission (FTC) intervention around the Web privacy legislation, and from which at least one large advertisers' group withdrew, putting the initiative in doubt.[4] More importantly, this withdrawal suggests that advertisers feel they need not be concerned about legislation curtailing their tracking activities. Why not? What is the significant difference between the surveillance and "threats to privacy" of the NSA vs. the user tracking embodied by Web businesses, the resellers and ad auctioneers like Doubleclick or Facebook as business as normal? Is it possible for the NSA to "betray" the Web, if the Web itself is already the Surveillance Web? To address these queries we look to Web prehistory that identifies the path to surveillance that any networked data system will develop given particular conditions - like the ones we now have for the Web.

In the pre-Internet, pre-Web era of 1988, the Communications of the ACM published an article by Roger Clarke, titled "Dataveillence", that is eerily prescient [3]. "IT technologies", he warns, "already make information sharing between governments and private companies easy". Indeed, computing systems have essentially replaced the sophisticated and expensive physical and electronic surveillance by enabling highly automated, cheaper and systematic collection of data about people. People's behaviour is

[1] Tim Berners-Lee: UK and US must do more to protect internet users' privacy. Friday 22 November 2013 08.40 GMT http://www.theguardian.com/technology/2013/nov/22/tim-berners-lee-internet-privacy-surveillance-censorship

[2] Tim Berners-Lee: NSA spying is 'betrayal' of world wide web. The Week http://www.theweek.co.uk/uk-news/nsa/55961/tim-berners-lee-nsa-spying-betrayal-world-wide-web

[3] Tracking Preferences Extension. Jan 28, 2014 http://www.w3.org/TR/2014/WD-tracking-dnt-20140128/

[4] Do-Not-Track Show Will Go On at W3C Advertising Age. Sept 20, 2013 -- For Now. http://adage.com/article/privacy-and-regulation/track-show-w3c/244285/

monitored through the increasingly intensive data trails that their behaviour is generating.

Clarke also references US work from the 60's that chose to abandon a national data center scheme in the interests of protecting privacy. He laments that three conditions only are required for any system to become a surveillance system: (1) a range of services processing personal data, (2) with these services connected on a shared telecommunications network, and (3) with the data consistently identifiable, that national data center, and hence the dissolution of privacy, effectively exists. He warns that two features must have already been in place for privacy to have been protected: intrinsic and extrinsic controls. He argues that intrinsic controls such as a consumer watch dog or even competitor are not possible when "surveillance activities are undertaken in a covert manner." Extrinsic controls are laws, economic policies and "natural" rejection by the society of unpalatable practices – all of which require again these practices to be visible..

If we apply Clarke's metrics to current history, we see such unpalatability of the now very visible NSA surveillance, but not around the more pervasive daily surveillance we do not see or somehow take as given, even if we might not be able to explain its mechanisms. In this paper, we quantify at least some of them. In the current Web era, we are tracked at every click in a "covert manner" that Berners-Lee and most of us abhor about the NSA breaches. Likewise, we can be re-identified from our traces through the Web with as little as three pieces of information [23]. Multiple groups collect and trade in this information. Clarke's "Extrinsic controls" are likewise impossible now as predicted, in the absence of "comprehensive information privacy laws" being in place. Legislation like EU's "cookie laws", requiring that companies to declare that they use cookies [29], is hardly useful when few of us know what a cookie is, or what that cookie may be capable of enabling. There are no laws that other tracking that does not leave traces – we describe below – is also being used.

Our research has looked at the prevalence of the user surveillance on the Web that fuels Web businesses. We have found it to be almost ubiquitous and dominated by a small number of Web entities [4]. In less than 30 clicks we get unknowingly ensnared in the widespread surveillance nets of all top 10 trackers. Each net is woven of a large number of Web sites that have given them permission to track the visitors in exchange for market intelligence and targeted ad campaigns. We are left without policies and legislations to give us a voice in the matter. There are no even sensible measures to enforce transparency around such activities, leaving us without means to develop strategies to deal with them. It is through this void, the lack of intrinsic and extrinsic controls that Clarke's has warned us about, that the Web has transformed itself into a *surveillance Web*. From this stems the deep irony: the "betrayal of the Web" ascribed to NSA escapes any meaningful argument; the Web that could be betrayed in that manner no longer exists.

This observation begs another important question to which we alluded above: has this transformation of the Web been a "natural" inevitable evolution of a data-oriented technology? Given Clarke's description of dataveillence, we can see it as inevitable, whether by nature or nurture. Thus, we are faced with a dilemma: to go forward with the dataveillence status quo, or try to get the Berners-Lee's Web back. If the latter, do we stand a chance to succeed with the currently enshrouded Web infrastructure? Or must we start again; this time with intrinsic and extrinsic controls established from the start?

We raise these questions with a full understanding that they are beyond a single research paper and already contemplated to various degrees by researchers in Web privacy, personal identity and Web commerce. Our goals for this paper are: (1) to offer quantified evidence of the Surveillance Web, and (2) to situate the discussion of the Surveillance Web in terms of the Web Science agenda.

Towards the first goal, we present three case studies that we conducted to quantify tracking that the users are exposed to while engaging in three basic Web activities: browsing, search, and sharing Web content within social media. As Clarke suggests, if we see it, we can explore intrinsic control. Second, based on the study results, we offer suggestions where in Web Science this evidence might be carried forward to support explorations of three important Web aspects around extrinsic controls: Web engineering, regulations, and economics.

In the rest of this paper, we first motivate our focus on the Web as the economic ecosystem driven by the data-exchange business and then, in the background and the related work section discuss the types of data traces that are involved in obscure tracing of the Web users. In Case Studies we present three typical use scenarios on the Web and the level of surveillance that the users are exposed to. Finally, we situate these observations in terms of a proposed Web Science discussion about the surveillance Web as the reality that we need to take into account when conducting research, devising methods to investigate individuals attitudes and societal preferences, and consider issues of design, engineering and policy to support or change the status quo.

2. MOTIVATION

The Web has been conceived as an information medium that enables transfer of data and information among organizations and individuals. It has since evolved into an intricate and multifaceted environment that enables ubiquitous access to digital content, communication and social interactions by the consumers and organizations. From the engineering point of view, the Web comprises a distributed computing infrastructure that uses HTTP protocols as the basis for network communication and an information architecture that uses hypertext as a unifying mechanism for content organization and access. The latter is the dominant aspect of the user experience on the Web as the user accesses the content through client applications such as Web browsers by specifying the URLs of the sites that wish to access or by following URLs provided by services and Web sites. Hyperlinks, resolved by the DNS referencing system, are the key to the information access and flow.

This explicit information exchange between users and content providers is dwarfed by the scale of data gathering at various levels of the infrastructure, communication protocols, and specific service access on the Web that are not visible to the individuals. Indeed, every process and every interaction with the Web leaves a digital footprint that can be persisted by the providers of the services. That ranges from the low level network data captured by the intermediaries that facilitate Web traffic in general, such as DNS services and Internet Service Providers (ISPs) to the usage data collected by the Web servers that host and provide content.

Generally, usage data of computing systems can help with resolving engineering issues and improving usability of the services. However, the Web has evolved to include data collection with a much broader scope and purpose. Digital footprints associated with specific services, and information inferred from such data, have become the common source of economic value within the Web ecosystem. Methods, such as third party tracking via cookies, have been devised to gather information about individuals browsing habits and interests across Web sites. Browser add-ons such as search toolbars continuously send information about the user's activities online. In essence, the information is gathered in exchange for the service.

More importantly, the value of user generated data is now driving the design of new types of services that, essentially, provide a digital real estate within which the interaction of individuals occur. In such cases both the user generated content and the usage information are accessible to the service providers. In fact, in many instances, the consent forms that the users agree to explicitly claim the ownership and rights to the data use by the service providers.

While we agree that there is some ambiguity as to the rights of the consumers, it is safe to assume that the usage and the ownership of the data disclosed through the Web infrastructure is in the hands of those who own it. Privatization of the Internet infrastructure in 1990s [22] has played a role in the rapid development of the Web economy that we have now.

One of the main purposes of this paper is to dispel any illusion of the Web as simply a 'free for all' information medium. As researchers we have a mandate to dive deeper into the phenomena of the Web and look at it from the engineering, economic, and social perspective.

Our studies confirmed that the complexity of the Web technologies and the lack of transparency in the design of services and client applications have left consumers in the dark about the collection and use of their digital footprints.

Not surprisingly, we often hear reports that people do not care about privacy [2, 5], putting forward the argument that if they did care they would use ad blockers and cookie blockers and stop posting personal details on social networking sites. Our case study that presents how people's responses change when they see what tracking looks like led us to think that most of us operate from the perspective "if we don't see it, it's not happening". This is consistent with our understanding of human psychology in which salience and availability play a significant part [6]. There was a time when the notion of invisible "germs" was inconceivable and it was hard to motivate people to wash their hands, even among the medical staff. People did not perceive a correlation between illness and lack of hygiene. It's still unclear if unseen Web interactions have similar effects. Our study offers the basis for situating the issues of privacy preferences within a research agenda that explores how our requirements and expectations about Web interactions changes as we develop mechanisms that reveal the unseen surveillance activities.

Furthermore, our empirical analysis of online services and cookie based tracking reveal the scope, the scale and the pervasiveness of such practices. Finally, we confirm that the Web of information has been transformed into the Web of data trading that leaves consumers without say and without choice except to opt out from the Web usage altogether. Thus, we are strongly motivated to extend the debate beyond the narrow technological field of the Web to the societal values, preferences, and specific measures that can be taken to enforce them if the Web is to become the essential underpinning of our societies and our digital future.

3. BACKGROUND

In this section we present a brief overview of tracking technologies that are used in various online scenarios. Most straightforwardly, the users' activities and data are logged and processed by entities that provide services to the users. Information about the user interests and habits is typically used to personalize the service or facilitate marketing campaigns and deliver targeted advertising. In general, one cannot assume that the use and resale of user profiles is restricted to the specific usage scenario. In addition, there has been little awareness and understanding of online tracking practices facilitated by cookies [8, 20]. This lack of awareness is partially due to the obscurity of the standard Web browser interface that neither informs us about cookies installed on our computer nor makes explicit the mechanism by which they enable the trackers to follow our interactions from one Web site to another [14, 20].

3.1 Third Party Cookies

Browser cookies are text files of small size that are installed on the user's computer at the time the browser is connecting with a Web site. They can be thought of as badges that the sites can use to label an individual. Cookies were originally intended to deal with the stateless nature of the Web [9] and enable a better experience for users as they move between the pages on a website. Such cookies are normally referred to as first party cookies since they are created by the Web site that the user is visiting. However, the use of cookies has evolved. By the same mechanism, other domains that are referred to by the first party website, such as advertisers or analytics providers, can install cookies on the user's computer. Such cookies area referred to as third party cookies.

Installing third party cookies on a site has a dual effect. First the third party can monitor repeat visits to the first party website by the same user. Second, the third party may reach the same agreement with other sites, thus creating the basis for tracking the user across multiple participating sites. By enlisting companies across market sectors the third party can track individuals in different contexts, from online shopping to gaming, entertainment, or search for health and financial advice. Furthermore, it can pass on this information to other partners who may be interested in specific types of tracking and marketing practices.

Figure 1 shows our analysis of user logs, based on a week-long logging of Web sites and tracking domains. As the user visits, for example, forbes.com, the user is immediately enlisted in the network of participating Web sites and trackers. The collection of information and placement of targeted ads is enabled by a tracker domain such as doubleclick.com.

3.2 Blocking Tracking Cookies

Various Web browsers provide different levels of control over cookie exposure. Almost all standard browsers offer basic features to block cookies of certain types or cookies from specific site. Thus, individuals who are aware of the tracking practices can assert a level of control over their exposure to tracking. Furthermore, by using specialized add-ons designed to support management of Browser cookies, the users can gain access to more information about the tracking domains. Such is the Ghostery add-on (www.ghostery.com) available for the Internet Explorer, Firefox, Safari, Opera, and Google Chrome browsers.

3.3 Other Tracking Practices

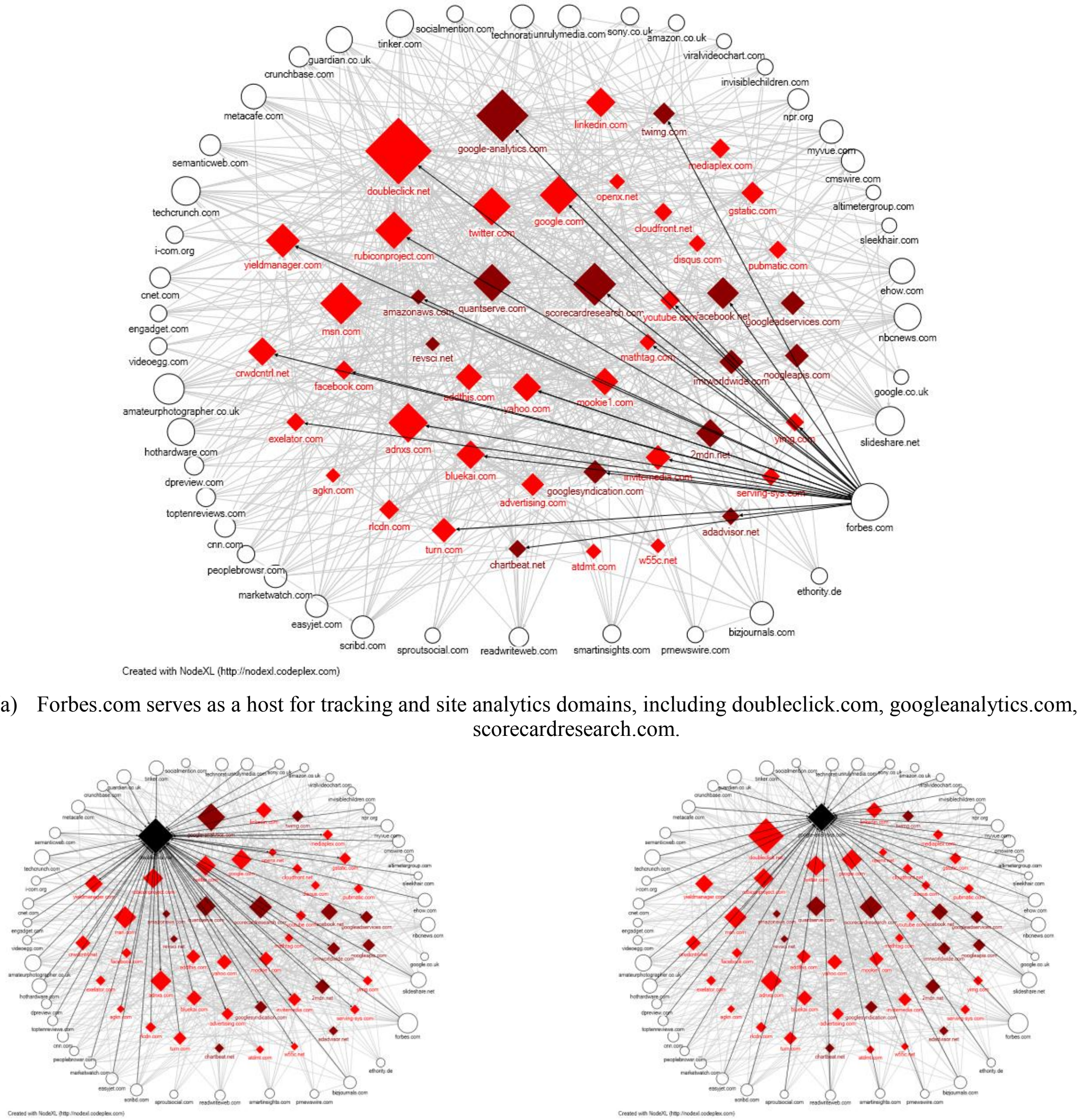
Cookies are only one kind of tracking. We have focused on these because, while often unobserved by users, they are easily detected. There are other kinds of tracking that leave little or no trace, suggesting that our case studies will already be only partially indicative of how wide-spread and deep tracking may be.

Tracking techniques can be split into two categories: stateful, and stateless [12]. Stateful mechanisms rely on storing some information on the users' computer and so HTTP cookies fall within this category. Other stateful mechanisms include Flash "LSOs" (which operate similar to cookies), HTML5's "local storage" API or more subtle techniques that make use of browser caching mechanisms.

Stateless mechanisms, by contrast, do not rely upon any kind of stored state. Instead, they rely on identifying a unique (or at least very uncommon) combination of properties related to a particular device. This could include, for instance, the list of installed fonts, the "clock skew" of the device (the number of milliseconds that the device's clock deviates from the true time), the list of installed plugins or the user-agent string that identifies the browser and operating system version. Combined, these properties can provide enough entropy to uniquely identify a particular device with enough stability to do so multiple times.

3.4 Online Advertising

Web sites who subscribe to an ad exchange service embedded code (e.g., Javascript) on their pages that connects to the ad



(a) Forbes.com serves as a host for tracking and site analytics domains, including doubleclick.com, googleanalytics.com, and scorecardresearch.com.

(b) Network of web sites and third party domains in the doubclick.com tracking network

(c) googleanalytics.com domain connects only to the web sites. It does not involve other third parties.

Figure 1. Network of Web sites (black) and tracking companies (red and purple) that are involved in installing cookies on the user computer. Red colored third party domains involve their partner domains. Graphs include nodes with eigenvector centrality of 0.0032 and above.

networks and loads adverts into the Web page at the time a page is rendered by the browser. In that process, ad networks may store or retrieve cookies containing a persistent user identifier. Such cookies are referred to as third party cookies, in contrast to the first party cookies that are delivered by the Web site itself. The latter are commonly used to support log-in and multi-page browsing on the site. As the user visits other sites associated with the same ad network, this third party cookie is used by the ad network to identify the user pseudonymously. In this way the ad network obtains, processes, and accumulates data about the user's online activity in real time.

Figure 1 illustrates the tracking mechanism: shows Web sites a user visited over 10 days (circular nodes) and the third party domains that were referred to during site visits (red diamond nodes). In the user's visit to everydayhealth.com, we observe that the visited page referred to a number of third parties which delivered adverts and installed cookies on the user's computer. Among them is the tracking domain doubleclick.net which is associated with other Web sites that the user visited (see Figure 2b). Every time the user visits such a site, the user's action is known to doubleclick.net. That information becomes the basis for behavioral targeting as it captures user's activities across Web sites.

4. RELATED RESEARCH

4.1 Behavioral Tracking and Responses

Online Behavioral Advertising (OBA) aims at inferring users' intent, preferences, habits, and interests from their online activities and selecting personal ads to present to the user. In many instances, the OBA providers, such as audiencescience.com and audiencetargetting.com offer retargeting of ads [10]. Ad retargeting involves placing an ad related to the Web site on the pages of subsequently visited sites and extending the exposure to the ad over time.

Privacy concerns related to the user tracking led to OBA approaches that reduce the scope of user information that is shared during ad targeting. Among such methods is user modeling on the client side, i.e., within the browser.

The system Adnostic by Toubiana et al. [25] uses a browser extension that incorporates behavioral targeting algorithm based on a local database of browsing history, not shared with external parties. Similar attempts towards privacy protecting techniques have been explored by Langheririch et al. [11] and Tomlin [24]. In 2011, Riderer et al. [18] proposed an alternative mechanism of transitional privacy, allowing the user to decide what personal information is released and put on sale while receiving compensation for it.

4.2 Analysis of Tracking Practices

Krishnamurthy and Wills [8] examine technical aspects of data aggregation by the third parties and, through longitudinal observations of the techniques and entities involved in the tracking practices, show that the market is consolidating towards strong dominance by a few companies.

Furthermore, Roesner et al. [20] differentiate among 5 third-party tracking practices based on the mechanism they use to manipulate the browser state. They designate them as: (1) analytics, for within site monitoring using a third party (e.g., Google Analytics), (2) vanilla, for cross site monitoring (e.g., DoubleClick), where a third party stores and aggregates user data, (3) forced tracking, for using pop-ups or similar mechanisms to force the users to visit the tracker's site, (4) referred, for negotiating with a service or a cross-site tracker to provide the unique user identifier, and (5) personal, for embedding a tracker (e.g., Facebook 'Likes' widget) that the user visits directly. They select 1,000 Web sites from the Alexa service (www.alexa.com) to observe the tracking practices and show that on most of the observed sites the users are tracked by multiple parties which combine different tracking practices. The coverage of the Web sites by the trackers varies with few of them playing a dominant role with large coverage.

Krishnamurthy and Wills's 2009 longitudinal study of Web tracking practices on the Web [8] shows the historical evolution of the major tracking entities. More recent work by Roesner, Kohno, and Wetherall 2012 [20] provides more detailed differentiation of tracking mechanism and statistical analysis of the tracking domains. They used Alexa rankings of sites from September 19, 2011 and observed tracking practice associated with 2500 pages from the top 500 international Web sites. The average number of trackers on 1655 pages, from 457 domains that embed at least one tracker, is over 4.5. Of these, 1469 pages include at least one cross-site tracker. Overall they identified a total of 524 unique tracker companies and estimated that several trackers can each capture more than 20% of a user's browsing behavior.

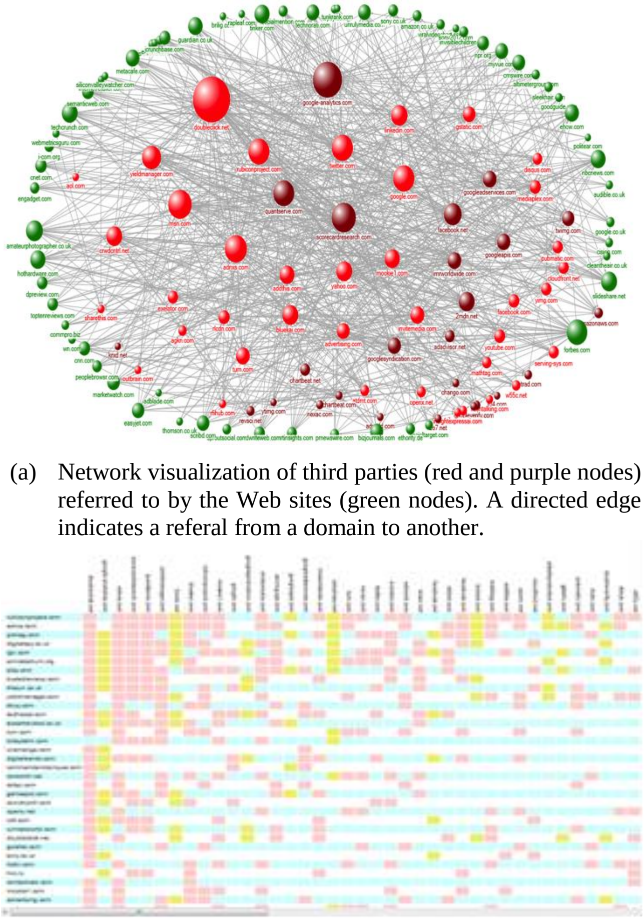
4.3 URL sharing in Social Media

One of our case study is particularly concerned with the use of social media, specifically Twitter, to propagate Web content by sharing URLs in tweets. Twitter users share tens of millions of Web links every day [17] and this trend is expected to increase over time. Market research has already shown that social networking sites are becoming a major driver of traffic to many Web sites [1]. Schonfeld [21] reported that for some sites, Facebook and Twitter drive, respectively, 44% and 29% of the traffic. Rodrigues at al. [19] found that, on average, a URL is spread by 3 users and reaches 843 users.

Considering the URL propagation patterns, Rodrigues et al. [19] differentiate among three types of participants: the initiators, the spreaders, and the receivers. The propagation ends with the receivers. Data analysis showed that nearly 90% of all URLs were propagated only by initiators without involving any spreaders, which indicates the importance of the initiators in the content discovery. Thus, it is not surprising that the Twitter cascades tend to be much wider than they are deep. However, those URLs that involved spreaders would gain 3.5 times larger audience showing that multi-hop propagation of URLs significantly increases the reach of the Web content.

Wu et al. [27] considered 5B tweets generated over a 223 day period in 2010. Among them were 260 million tweets with bit.ly URLs. Their analysis showed that roughly 50% of URLs in Twitter are generated by only 20,000 elite users (0.05% of the user population). Among them the media produces the most information while the celebrities are most followed. Furthermore, different types of content exhibit very different lifespans with media-originated URLs being predominant among short-lived URLs while those originated by bloggers are prevalent among long-lived URLs. In fact, the longest-lived URLs are related to media content such as videos and music that seems to be repeatedly rediscovered by the Twitter users and thus persist within Twitter.

Our study complements this research by focusing on the invisible network of trackers that the users of social media are exposed to while visiting the shared URLs. We explore how much of the personal and social information can be revealed to the trackers just from the visits to the URL domains.



(a) Network visualization of third parties (red and purple nodes) referred to by the Web sites (green nodes). A directed edge indicates a referal from a domain to another.

(b) Information about third parties in the tabular form. Pink−TP that uses cookies; Yellow−TP that does not use cookies; Blue/White indicates no tracking.

Figure 2. Aggregate information about third party (TP) involvement with visited first party (FP) domains during user browsing based on user log collected over a period of 7 days.

5. CASE STUDIES

We have conducted three studies around the surveillance Web that provide empirical evidence of its scale and reach as well as qualitative insight into the beliefs and expectations of Web users in relation to the surveillance of their own Web browsing activity.

5.1 Browsing and Cookie Invisibility

As part of a study evaluating real-time in-browser visualization of surveillance activity, we obtained browsing logs for 14 users. These logs were used to create visualizations that were shown to participants during semi-structured interviews in order to capture their reactions to the revelation of the surveillance networks.

Example visualisations are shown in Figure 2. Each participant was shown the two visualisations and was provided with a verbal explanation of each. We first checked the participants' understanding by asking them questions which required them to read information from each visualisation. We then inquired about their reaction to the presented information.

Reactions to the visualisations were fairly uniform across participants. Participants indicated that the extent of the tracking was a revelation to them; for instance: "*I never realised … how many websites were watching [the] website that you're on*" and "*I didn't know that that many websites could all see what you were doing*."

This is an interesting result, as the same information was available to the participants via browser add-ons during the. It indicates that participants more easily appreciate the scale of TPT when it is presented in aggregate, outside the browsing context, than when shown for a single site during the browsing task. This confirms our first hypothesis that, despite the surveillance infrastructure of the Web being detectable by technical means it is largely invisible from a user's point of view.

Participants' responses reflect the general anonymity of parties involved in TPT: "*I have no idea who they are*", "*I never heard of this company, Rubicon*" and "*I probably don't know any of them. I would want to know what the nature of their business is*."

5.2 Exposure to Trackers through Search

As Michael Zimmer argued in his 2006 paper [31], "The Panoptic Gaze of Web Search Engines," search engines are the portal through which much of the Web is accessed. They have unrivalled access to the lists of keywords that describe what individuals and groups are interested in, and which betray their thoughts. However, by acting as a portal to content on the web they also act as a portal to the surveillance infrastructure that permeates it. The goal of this second study was to better understand how search engines expose Web users to surveillance [4].

We obtained search results for 662 categorised search queries, from two search engines (Google and Bing), across 3 English-language markets (United States, United Kingdom, South Africa and India). Using the browser automation framework Selenium and the Firefox web browser, we crawled each set of 10 search results. Using a custom browser add-on we recorded the HTTP Referer headers sent by the browser to create a map of the third parties associated with each visited website.

Based on their network properties and appearance in search results, we categorised the websites that were contacted by the web browser into four groups (Figure 3).

1. *Web sites:* Web domains whose pages appear among search results and are not referred to by other sites. They are thought of as the first party only domains.
2. *Third party only:* Web domains that are referred to by Web sites or other third party domains and never appear among search results nor refer to other domains. Such are, for example, googleanalytics.com or ad services that place ads directly on the Web pages.
3. *Dual role:* Web domains that appear as both first party and third party domains. Example is facebook.com which appears among search results and is referred to by sites that include the Facebook "Likes" widgets.
4. *Ad Exchange Service:* Web domains that appear only as third parties, i.e., do not appear in search results, and refer to other third party domains. They are intermediary third parties that provide a bridge between Web sites and other third parties involved in ad bidding.

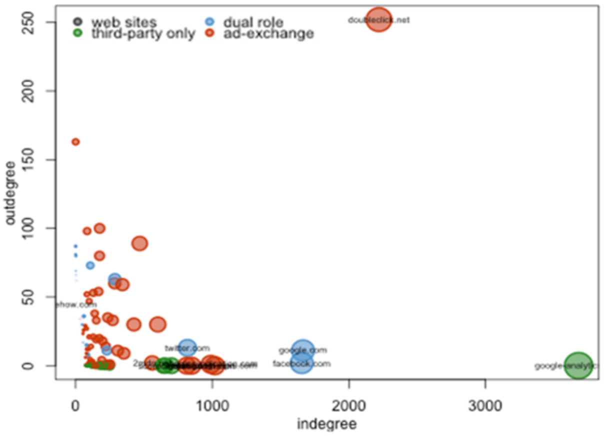
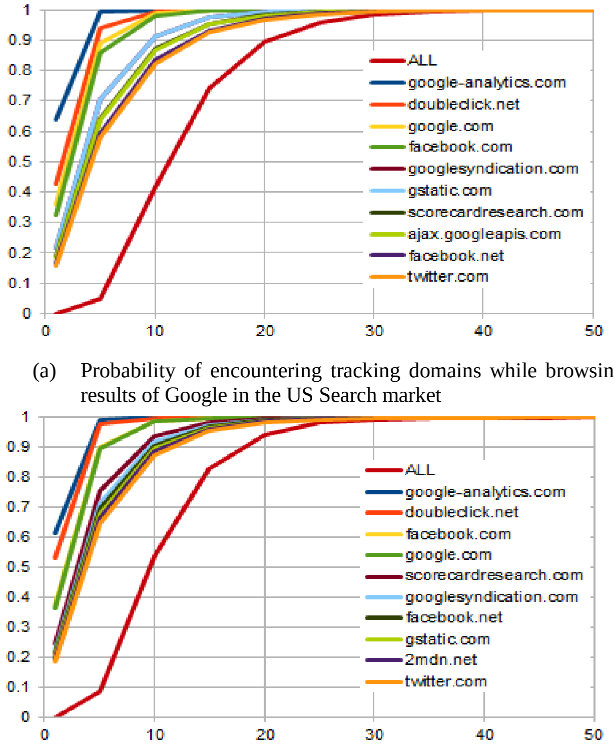


Figure 3. In-degree vs. out-degree of the nodes of the global network. The size of the nodes is mapped to the total degree.



(a) Probability of encountering tracking domains while browsing results of Google in the US Search market

(b) Probability of encountering tracking domains while browsing the results Bing in the US market

Figure 4. When browsing search results , the user is exposed to tracking domains. We calculate the probability that the user encounters top tracking domains while visiting a random set of search result pages.

Our analysis of the search results in different search markets revealed a consistently high extent of third party tracking. Similarly to previous studies [8], we identified a small number agencies that dominate the user tracking and advertising markets. However, in contrast to Roesner et al. [20] who focus on popular Web sites and tracking classification based on the mechanisms for implementing cookies, we focused on user exposure to tracking in a real usage scenario and broadened the scope of the tracking analysis to the properties of the referrer network itself.

By considering directional links of the referral networks we can easily observe the role of the domains as illustrated in Fig. 7. An in-link to a node is a referral that provides an opportunity for that entity to surveil the user.

In order to estimate the rate at which users are exposed to third parties, we estimate the probability P(T) that a search result exposes the user to a third-party T by calculating the proportion of search results that refer to T. We rank third parties based on P(T) and, for the top 10, determine the likelihood that the user will encounter each of these parties after accessing a number of retrieved search results. We make two simplifying assumptions. First, we assume that any Web page from a given Web site is exposing the user to the same set of trackers. Second, we expect that the user's choice to visit a search result is independent from the previously seen pages. Based on this model we observe the probabilities that a user would have encountered all top ten third parties. We find that after visiting just 30 search results, the probability of getting cookies from all top 10 third party domains is 99.5%. Figure 4 shows the probabilities of encountering the trackers when using Bing and Google search engines in the US Market.

5.3 Exposure to Trackers through Social Media

On today's Web, the role of content discovery is increasingly filled by social media and no longer confined primarily to search engines. Although search still fulfils an important role for purposive information retrieval, social media provides a means for individuals to share content through social networks that form around real-world ties or as a result of shared interests online. Social media is, therefore, a key mechanism through which Web users are brought to online content and hence how they are exposed to the surveillance infrastructure. This additional means of content discovery potentially reveals different information about the user; inferences can now be drawn about which topics a user has a general interest in, which communities they are part of and who they are influenced by.

To investigate the surveillance that users are exposed to through social media we conducted a third study, similar to the study that was conducted on search results. The aim was to obtain a referrer network of the first and third parties involved in URLs shared on the social networking service twitter, and to characterise the way in which users are exposed to surveillance when engaging with the shared URLs.

A dataset of 5.4 million tweets was collected via the twitter API over a period of seven days in January 2013. The stream was filtered to the 54 most popular hashtags in ten topics, obtained from hashtags.org on 05/01/2013: U.S. Politics, TV/Entertainment, Music, General, Business, Tech, Education, Environment, Social Change and Astrology. Our aim was to collect tweets with URLs, focusing on specific topics have been the subject of previous research [26].

We extracted and expanded the URLs from all the collected tweets that received at least one retweet. Repeated tweets were discarded and only the original ones were kept in the dataset. After applying these filters we arrive at a dataset of 499,916 tweets.

Table 1. Third party domains ranked by % of re-tweeted URLs that refer to the specific domain and by % of users who tweeted URLs that refer to the specific domain.

Third Party Domain	% of URLs	Third Party Domain	% of Users
google-analytics.com	41.5%	google-analytics.com	73.7%
facebook.com	22.2%	doubleclick.net	64.1%
google.com	21.3%	google.com	58.8%
twitter.com	20.7%	gstatic.com	57.4%
gstatic.com	19.2%	googlesyndication.com	57.0%
chartbeat.net	18.8%	googleadservices.com	52.9%
chartbeat.com	18.8%	facebook.com	49.6%
youtube.com	15.6%	googleusercontent.com	48.6%
doubleclick.net	15.3%	youtube.com	45.4%
facebook.net	14.3%	youtube-nocookie.com	39.4%

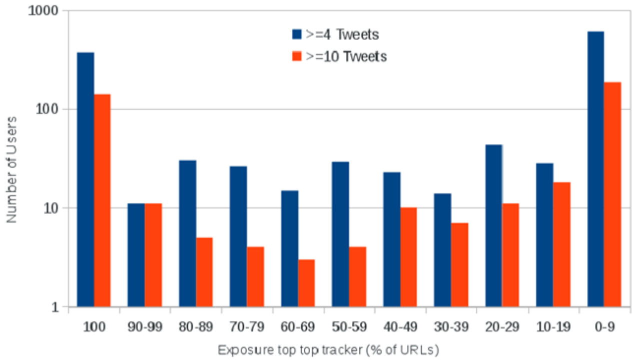


Figure 5. The number of twitter streams in which at least one third party can observe a given percentage of the tweeted URLs.

Again, we used the Firefox web browser and Selenium browser automation framework to visit each of the URLs in the dataset and recorded the HTTP Referer headers to build up a network of domains, indicating which first party domain referred to which third parties. The percentage of URLs that were observable by each, are shown in Table 1. Also in Table 1 are the top third party domains by user coverage. That is to say, the third party domains that were able to observe at least one URL that was tweeted by a given user. For instance, 73.7% of users tweeted at least one URL that was observable by google-analytics.com, and google-analytics.com was able to observe 41.5% of the total URLs that were tweeted.

Subsequently, we were able to analyse what percentage of the URLs in each users' tweets were observable by each third party. Overall results from this analysis are indicated in Figure 5, which shows the number of twitter streams in which the most common tracker could observe the given percentage of URLs. For instance, if 100% of the tweeted URLs were observable by a single tracker, that twitter stream is classified as 100%. If 49% of the tweeted URLs were observable to one third party, and 65% visible to another then that user's Twitter stream is classified as 60-69%.

We repeated this analysis twice—once including all Twitter users with 4 or more tweets in the dataset and once including only those with 10 or more tweets in the dataset. The results show that 100% of the URLs tweeted by many users were observable to the same third party and thus any follower who visited any URLs tweeted by those users was guaranteed to be surveilled by that third party.

5.4 Analysis: Surveillance Hidden in Plain Sight

The results from these three studies reveal three important properties of the web's surveillance architecture that are relevant to this paper:

1. The infrastructure is largely invisible to web users, despite being relatively easy to detect by technical means. Users are surprised when shown how many domains are able to observe their web browsing.
2. Surveillance is very widespread, to the point of being almost unavoidable. We observe that the probability of being observable at least once by all top ten of the most common third parties is 99.5% after visiting just 30 URLs.
3. Surveillance is highly centralized. Despite the huge number of third party domains that were uncovered during the crawling, a handful of domains dominate the network. Doubleclick.com alone was able to observe visits to 15% of all the URLs obtained from twitter, with 64% of all users tweeting at least one URL that is observable by Doubleclick.

We caution that the behaviour of many of the third parties is unknown. While companies such as Doubleclick are relatively open about the purpose of their surveillance, others are not. We used the presence of cookies on a domain to determine whether a domain had the potential to track a user across multiple requests but this does not take into account the other stateful or stateless tracking mechanisms that exist.

Whereas a Web user can be reasonably confident in understanding which organisations are providing the online tools that they use, we have seen that Web users are not aware of the scale of surveillance that is possible as they browse the Web and that they have not even heard of many of the organisations that are party to portions of their web browsing history. Furthermore, the surveillance Web directly supplements the dataveillance that is possible through cloud applications like email or productivity tools. Not only do companies like Google have access to the information contained within a person's emails, calendar, documents and search history, they have access to information about their visits to numerous Web sites that rely on services such as Google Analytics or advertising via Doubleclick (a Google subsidiary). Organisations such as Facebook and Twitter can use dataveillance performed via "social widgets" that are embedded on third party websites to supplement the information about a user's social network with data about the websites that they have visited

6. DISCUSSION: ALTERNATIVE WEBS

The findings from the presented three case demonstrate that the surveillance Web, this Grey Web, is the Web. We propose the term "grey" for all its connotations of ambiguous status between black and white markets, greys of shades of grey in an argument, and also grey as hidden, occluded, moving in shadows. As we note, we trace only some aspects of some tracking.

Given Clarke's prescription for dataveillance, this result – the evolution of the web into this grey web - seems to be expected, i.e., inevitable within any networked data system. In the absence of the implicit and explicit controls that he describes, ranging from audit trails to privacy laws, no matter how the Web is re-engineered to hold off the deliberate breaking of our cryptographic codes as done by NSA, this "legal" surveillance, under the guise of "free services" will simply be maintained or recreated—as they are now the "norm", expected and the de facto. Moreover, there is no alternative channel. Citizens cannot readily opt out of this commercial panopticon when increasingly all information from government services to utilities to purchasing of goods takes place over that surveilled channel, by both governments and industries.

Clarke differentiates between "mass dataveillance" and "personal dataveillance". Mass dataveillance is "the systematic use of personal data systems in the investigation or monitoring of the actions or communications of groups of people". Similarly, personal dataveillance is "the systematic use of personal data systems in the investigation or monitoring of the actions or communications of an identified person".

In the Web of 2014 we propose that we are presented with an infrastructure for "mass personal dataveillance", that is, the systematic use of personal data systems to monitor the actions of many identified people. The scope of what is available via the combination of the surveillance Web and its enabling dataveillance apparatus (such as online services) is akin to personal dataveillance, with the purpose to glean information about a specific individual for the purpose of commercial gain, for instance, targeted advertising. The dataveillance is, though, conducted on the same scale as mass surveillance. We are all personally surveilled.

Given the privileged access that governments have to the private dataveillance infrastructure that has been built by the likes of Google, Microsoft or Facebook [15], if we wish to make government surveillance more difficult through engineering measures, we need to tackle this private infrastructure, too.

We note there are a variety of initiatives that technologically describe alternatives to clouds of data shared by large hosts, probed by large services where each transaction is known. Webbox [7], the Locker Project [13], Data.fm [16] are a few of the examples that propose peer-to-peer, controlled sharing networks for exchange of information among groups. These are engineering oriented solutions to a massive, global way of being. They may implicitly rely on a sense of honouring personal privacy and data control as a fundamental component, but, as engineering solutions they, again, *a priori* do not and perhaps on their own cannot embody the explicit/implicit controls that Clarke identifies must precede engineering to ensure that any networked data solution respects auditability and ownership.

Our goal in this presentation has been to offer in quantified terms clear evidence that the Web has moved from a more neutral document/data Web of pre third-party cookies, now being the tracking, surveilling Web at industrial scale.

How to move forward from this point we suggest is a grand challenge for Web Science. We may need to think boldly about blank slates and starting fresh; of moving out of the monoculture of the surveillance Web into a multicosm of multiple Webs. At this point, the appropriate place for Web re-imagining is in the realm of research rather than industry or government, as it has been evidenced by the degree to which trying to backfill the current Web with new policy is failing. For instance, in Europe, there is the 1995 Data Protection Directive and the ePrivacy Directive [29], the latter specifically targeting stateful tracking mechanisms by requiring user consent. We see this manifest in the UK as a banner that shows up on Web sites simply stating that a page uses cookies. One can either use the page or not—a false option surely if one must access the service.

More recently, a 2014 deadline for the proposed Data Protection Regulations was dropped by the European Council, in favour of less urgent "timely adoption" at the request of the UK, to provide more time for the government to 'consider the implications of the regulations for businesses'. Given the blurred distinction between state and private dataveillance, it is not clear why the surveillance Web is not covered by legislation that governs surveillance, such as the UK's Regulation of Investigatory Powers Act (RIPA). That said, exercising existing regulations has proven to be difficult, with American companies questioning whether European courts have jurisdiction to try cases brought under European laws [30].

While current businesses and policy makers are enmeshed in current technologies and interests of the surveillance Web, we need Web Scientists to reimagine technology, law, economics and social policy re-delivering a new variant, pre "betrayal."

Fundamentally, the observed phenomena is, in many ways, the repeat of history. In any technological revolution we, as citizens have had very little say in the technologies that are largely thrust upon us, from agriculture to the industrial revolution to the railway with its robber barons, and oil with it nation state interest. The surveillance Web has evolved as the next channel for a global economy and its interests. In each epoch there have been those to argue that the new technology can also bring new benefits to the dispossessed. Certainly, but at what cost? It will be interesting to see whether the interdisciplinary community of scholars in the new field of Web Science might have any greater effect in mitigating the current control of the Web by the same old few major players we seem to find in any utility/industry and re-create a new status quo.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have contributed methods and results for quantifying and presenting the prevalence of user tracking on the Web. We used this evidence to support our claim that the Web has been transformed from the information media into a complex economic ecosystem based on tracking and surveillance of the user's behavior. We recognize that we are able to graph and represent the tracking that remains traceable trough observing the HTTP traffic and site referral. Thus, the provided view is far from exhaustive of the types of data gathering, analysis, and trading on the Web.

We have suggested that the prevalence of this non-governmental, largely commercial and unregulated tracking, is an evolution in the Web that has moved from its early days as a Web of hyperlinked documents and the means for connecting people, to the elaborate Clarke's dataveillence network. We have suggested that in the absence of the implicit and explicit controls promoted by Clarke, this transition of the Web into mass personal dataveillence was inevitable.

Our future work will focus on promoting the framing of the privacy issues as the Web transformation into the surveillance Web. We present the surveillance Web to the Web Science community to embrace it as a grand challenge and inspiration for devising the Web that we, citizens, want for ourselves.

8. REFERENCES

[1] Campbell, A. 2009. Social Activity Becomes Significant Source of Website Traffic. Small Business Trends. Available: http://smallbiztrends.com/2009/03/social-activity-significant-source-website-traffic.html

[2] Casey, K. 2013. Online Privacy: We Just Don't Care. Information Week. Available: www.informationweek.com/security/risk-management/online-privacy-we-just-dont-care/d/d-id/1110535

[3] Clarke, R. 1988. Information technology and dataveillance. Communications of the ACM. 37, 5 (1988), 498–512.

[4] Gomer, R. Milic-Frayling, N. and schraefel, m.c. Network analysis of third party tracking: user exposure to tracking cookies through search. Web Intelligence 2013 (2013)

[5] Hill, K. 2010. Zuckerberg's right: Young people don't care (as much) about privacy. Forbes.com. Available: http://www.forbes.com/sites/kashmirhill/2010/01/10/zuckerbergs-right-young-people-dont-care-as-much-about-privacy/

[6] Kahneman, D. 2003. A perspective on judgment and choice: mapping bounded rationality. The American psychologist. 58, 9 (Sep. 2003), 697–720.

[7] Van Kleek, M. et al. 2012. A decentralized architecture for consolidating personal information ecosystems: The WebBox. Personal Information Management Workshop, part of CSCW2012 (2012).

[8] Krishnamurthy, B. and Wills, C. 2009. Privacy diffusion on the web: A longitudinal perspective. WWW '09 (2009), 541–550.

[9] Kristol, D. and Montulli, L. 1997. RFC 2109: HTTP state management mechanism.

[10] Lambrecht, A. and Tucker, C. 2011. When does Retargeting Work? Timing Information Specificity. SSRN Electronic Journal. (2011).

[11] Langheinrich, M. et al. 1999. Unintrusive customization techniques for Web advertising. Computer Networks. 31, 11-16 (1999), 1259–1272.

[12] Mayer, J.R. and Mitchell, J.C. 2012. Third-Party Web Tracking : Policy and Technology. Proceedings of the 2012 IEEE Symposium on Security and Privacy (2012), 413–427.

[13] Miller, J. and Murtha-Smith, S. 2010. The Locker Project. http://www.lockerproject.org

[14] Mittal, S. 2010. User Privacy and the Evolution of Third-Party Tracking Mechanisms on the World Wide Web. SSRN Electronic Journal. (2010).

[15] National Security Agency 2013. PRISM Overview. Available: http://www.theguardian.com/world/interactive/2013/nov/01/prism-slides-nsa-document

[16] Presbrey, J. 2010. Read Write Linked Data Space. data.fm.

[17] Rao, L. 2010. Twitter Seeing 90 Million Tweets Per Day, 25 Percent Contain Links. Tech Crunch. Available: http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/

[18] Riederer, C. and Erramilli, V. 2011. For sale: your data: by: you. Proceedings of the 10th ACM Workshop on Hot Topics in Networks - HotNets '11. (2011), 1–6.

[19] Rodrigues, T. et al. 2011. On word-of-mouth based discovery of the web. Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference - IMC '11 (2011), 381.

[20] Roesner, F. et al. 2012. Detecting and Defending Against Third-Party Tracking on the Web. Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (2012).

[21] Schonfeld, E. 2010. Facebook Drives 44 Percent of Social Sharing On The Web. Tech Crunch. Available: http://techcrunch.com/2010/02/16/facebook-44-percent-social-sharing/

[22] Shah, R.C. and Kesan, J.P. The Privatization of the Internet's Backbone Network. *Journal of Broadcasting & Electronic Media 51*, 1 (2007), 93–109.

[23] Sweeney, L. et al. 2013. Identifying Participants in the Personal Genome Project by Name. SSRN Electronic Journal. (2013), 1–4.

[24] Tomlin, J. 2000. An entropy approach to unintrusive targeted advertising on the Web. Computer Networks. 33, 1-6 (2000), 767–774.

[25] Toubiana, V. et al. 2010. Adnostic: Privacy Preserving Targeted Advertising. NDSS. (2010).

[26] Weng, J. et al. 2010. Twitterrank: finding topic-sensitive influential twitterers. WSDM. (2010), 261–270.

[27] Wu, S. et al. 2011. Who says what to whom on twitter. WWW '11 (New York, New York, USA, 2011), 705.

[28] Zimmer, M. 2006. THE PANOPTIC GAZE OF WEB SEARCH ENGINES: GOOGLE AS AN INFRASTRUCTURE OF DATAVEILLANCE. National Communication Association Conference (2006).

[29] 2009. DIRECTIVE 2009/136/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.

[30] 2014. Vidal Hall, Hann & Bradshaw V Google Inc., England and Wales High Court, Case No: HQ13X03128.