Master's thesis planning report

Preliminary title

Swedes Online: You Are More Tracked Than You Think

Joel Purra, mig@joelpurra.se, joepu444, 070-3521212

v0.2.3 - April 9, 2014

Abstract

How many companies are tracking you online, and how much information does the average Swede leak while using popular .se websites? Many, and a lot - more than you think. Large organizations like A, B and C are able to connect the dots you leave behind during everyday usage, and construct a persona that reflects you from their perspective. Have you told your family, friends or colleagues about your gambling addiction, your sex toy purchases or your alcoholism? Replace with different scary personal information. Even if you didn't tell anyone your deepest secrets, these companies might conclude that they can put labels on you by looking at everything you do online. And now they are selling it as hard facts behind the scenes.

While browsing the web users are both actively and passively being tracked by multiple companies, for the purpose of building a persona for targeted advertising. Sometimes the data collection is visible, as in social network sites and questionnaires, but it's most common in the form of different kinds of external resources which may or may not serve a purpose other than keeping track of your every click. Tracking code is installed on web pages that have adverts as well as those that do not - the spread and reach of tracking across web pages and domains of different kinds increases the quality of the user data collected and inferred, making it more valuable for advertising purposes. With the extent of the use of trackers and other external resources largely unknown and ever evolving, what is already known raises privacy concerns - data considered personal leak without the user's knowledge or explicit permission and end up in privately owned databases for further distribution. Data collection is the new wild west, and you are the new cattle. Or Klondike and gold?

To show the overlap between different sites, front pages of Swedish top sites were downloaded and their resources counted and grouped. In this thesis I show the use of resources internal versus external to the entry domain, which the most common confirmed trackers are, what spread they have and how much the average Swedish internet user can expect to be tracked by visiting some of the most import and popular sites in Sweden.

About The Internet Infrastructure Foundation[1] (.SE)

.SE is also known as Stiftelsen för internetinfrastruktur (IIS).

The Internet Infrastructure Foundation is an independent organization, responsible for the Swedish top level domain, and working for the benefit of the public that promotes the positive development of the internet in Sweden. Their head office is in Stockholm. In 2012 they had 61 employees and a turnover of almost 120 MSEK.[14]

Background and context

Part of .SE's research efforts include continuously analyzing internet infrastructure and usage in Sweden. Yearly reports convey the status of, for example, *Swedes and the internet* and *.se health status*[11] to the public, both in Swedish[2] and English[3]. Information and statistics are also published on a separate portal, in collaboration with other organizations.[4]

The report *.se health status* is based on data collected from around 900 .se domain names deemed of importance to the Swedish society as a whole, as well as random selection of 1% of the registered .se domain names. The research is focused on statistics about usage and security in DNS, IP, web and e-mail; the target audience is IT strategists, executives and directors. Data is analyzed and summarized by Anne-Marie Eklund Löwinder, a world-renown DNS and security expert[5], while the technical aspects and tools are under the supervision of Patrik Wallström, a well known DNSSEC expert and free and open source software advocate[6].

Problem description

> The problem description shall be detailed and include a background and a motivation to why it is important. Expected results shall also be described. The problem description shall be grounded in the literature base and the state-of-practice of the provider of the thesis (company, research group). Plan for adjustment of the problem description along with the progress of the literature studies and pre-study of the provider.

[1] https://www.iis.se/
[2] https://www.iis.se/lar-dig-mer/rapporter/
[3] https://www.iis.se/english/reports/
[4] https://www.iis.se/vad-vi-gor/internetstatistik/
[5] https://www.iis.se/bloggare/anne-marie/
[6] https://www.iis.se/bloggare/pawal/

Background

In everyday web browsing, browsers routinely access a lot of material from other domains or services than the one visited.[3] These external resources vary from content that the user explicitly want to obtain, to implicitly loaded third party services, ads, and non-visible resources with the sole purpose of collecting user data and statistical material.[7] All are downloaded on behalf of the user with no or few limitations, and oftentimes without the user's need, understanding and explicit consent. These external resources can all be seen as browsing habit trackers, whose knowledge and power increase with any additional visits to other domains or services loading the same resources.[12]

While online privacy has been in the spotlight due to recently uncovered mass surveillance operations, the focus has been on national government intelligence agencies collecting information around the globe. They have been able to intercept traffic data and metadata by, among several techniques, covertly hooking into the internet infrastructure. In contrast, external resources are approved by and actively installed by site and service owners, and presented openly to users with basic technical skills and tools. Because these external resources are used on behalf of the service, they are also loaded when end-to-end encryption with HTTPS is enabled for enhanced privacy and security. This encryption gives these private trackers more information than possible with large-scale passive traffic interception, even when there is a security nullifying mixture of encrypted and unencrypted connections.

Depending on what activities a user performs online, different things can be inferred by trackers on sites where they are installed. For example, a tracker on a news site can draw conclusions about interests from content a user reads (or choses *not* to) by tagging articles with refined keywords and creating an interest graph.[8] The range of taggable interests of course depend on the content of the news site. Private and sensitive information leaked to third party sites during typical interaction with some of the most popular sites in the world include personal identification (full name, date of birth, email, ip address, geolocation) and sensitive information (sexual orientation, religious beliefs, health issues).[12] Social buttons, allowing users to share links with a simple click, are tracking users whether they are registered, logged in or not.[13] They are especially powerful when the user is registered and logged in, combining the full self-provided details of the user with their browsing habits.[5]

Publishers reserve areas of their web pages for displaying different kinds and sizes of advertisements alongside content. Ads chosen for the site may be aligned with the content but it is more valuable the more is known about the visitors. Combining and aggregating information from past visitors means that more information can be assumed about future visitors, on a statistical basis, which will define the general audience of the site. To generate even more revenue per displayed ad, individual users are *targeted* with personalized ads depending on their specific personal data and browsing history.[4] What kind of data can be collected by trackers, and how can they be aggregated both per person and per group of people?

How much does the average user know about external resources being trackers?

Expected results

Previous research show that ads were used on 58% of internet's 500 most popular sites.[7] Google Analytics usage among Swedish top domains was 57% in 2011 and 62% 2012.[10, 11] The assumption is that the number of external resources is at least as big, as they include both ads and Google Analytics. Technical reasons include cloud services hosting sites and services, content delivery networks becoming commonplace[6, 7] for scalable speed improvements and external service providers increasing their quality. Add non-technical reasons.

Sites served over HTTPS are expected to use as many external resources as HTTP, even though some of these external resources might not be served over HTTPS as well.

News sites are expected to allow more trackers than other categories, as their income model include third party advertisements.[7] Commercial sites are expected to have more trackers than government sites.

Direction and scope

Emphasis for the thesis will be on technical analysis, producing aggregate numbers regarding domains and external resources. Social aspects and privacy concerns are considered out of scope.

The thesis will primarily be written from a Swedish perspective. This is in part because .SE has access to the full list of Swedish .se domains, and part because of their previous work with the *.se health status* reports. Focus is to analyze .se domains in the reports, as they have already been deemed important and results can be incorporated in future reports. The main non-technical grouping is also based on the same reports; government, media, banks, larger websites, etcetera.

One assumption is that all external resources can act as trackers, collecting data and tracking users across domains using for example the `Referer` HTTP header[7]. While there are lists of known trackers, used by browser privacy tools, they are not 100% effective.[12, 7] The lists will instead optionally be used to emphasize those external resources as *confirmed* trackers.

Questions

With domain and resource data in place, it will be aggregated to answer the following questions.

Why are these questions important? Why were they chosen?
Group questions by refined category?

- What kinds of resources are there?
- How many resources are internal versus external per domain?

- What is the distribution of different kinds of resources?
- How many external resources are there, considering different levels of uniqueness:
  - Unique URLs?
  - Unique per file URL?
  - Unique per folder URL?
  - Unique per subdomain?
  - Unique per domain?
  - Unique per TLD?
- On how many domains is each external resource is represented?
- How does usage of external resources differ between groups of domains?
- How to mark certain external resources as known trackers?
- What is the usage and distribution of known trackers?
- Are you as tracked using secure HTTPS as insecure HTTP?
- How do the results compare to
  - Historical .se data, if readily available from earlier .SE status checks?
  - Other ccTLDs?
  - Commonly used gTLDs?
  - Recently introduced newTLDs?

Additional questions, which can be considered as bonuses

- Could any external resources actually be considered internal, despite being loaded from external domains?
- How to determine if a resource
  - Crosses Sweden's borders in transit?
  - Is handled by an organization with base or ownership outside of Sweden?
- Which external resources are loaded from Sweden and abroad respectively?
- What user data could potentially be collected, and subsequently inferred?
- To what extent can the average Swedish internet user's browsing habits be correlated across the most commonly visited webpages?

Approach

> The approach is a preliminary description of how the problem will be solved. This section shall also include a description of a method to evaluate that the problem is solved in a satisfactory way.

Based on a list of domains, external resources are listed by downloading of the front page of each domain, and analyzing its HTML content. The URLs of external resources will be extracted, and associated with the domain they were loaded from.

All external resources get some of the relevant data upon each request, even for static resources with no capabilities to dynamically survey the user's browser. While cookies used for tracking have been a concern for many, they are not necessary in order to identify most users upon return, even uniquely on a global level.[2] Cookies will not be considered to be an indicator of tracking, as it can be assumed that a combination of other server and client side techniques can achieve the same goal as a tracking cookie.

In order to facilitate repeatable and improvable analysis, tools will be developed to perform the collection and aggregation steps automatically. .SE already has a set of tools that run monthly; integration and interoperability will smooth the process and continuous usage.

Potential problems

- Due to the dynamic nature of modern web pages, a static HTML analysis might not be enough. How can pages with dynamic script loading be analyzed?
- Script aggregation and concatenation could give misleading numbers if only analyzed per URL. Is it possible to detect which known scripts are actually running?
- Can Google Tag Manager[7] scripts, which is script aggregation with asynchronous loading directed specifically to marketers, be analyzed to show each included service?
- Can collected data served by different services differ depending on which tool is used to fetch the data?
- Many of the external resources will be overlapping, and downloading them multiple times can be avoided by caching the file the first time in a run. Would keeping a local cache of recently requested URLs affect the results?
- Automated downloading of webpages, especially downloading several in short succession, can be seen by site and service owners as disruptive by using system resources and skewing statistical data. Traversing different

[7] http://www.google.com/tagmanager/

pages on a single website can also be detected by looking at for example navigational patterns.[15, 9] By only downloading the domain root page and associated resources this tool might not fall into that category of detection. Will automated collection done for this report be detected and hindered?

Literature base

The literature base describes the planned literature study and gives examples of different directions of a good theoretical grounding of the work.

Some research has been done surrounding ad networks, trackers and their spread on globally popular sites, as well as what kind of private data users can expect to more or less inadvertently share in the course of normal internet usage. Those papers show both some of the problems and solutions in trying to analyze external resources. The Association for Computing (ACM)[8] group SIGCOMM[9] has a yearly Internet Measurement Conference (IMC)[10], where some papers of interest have been presented. The Passive and Active Measurements (PAM) Conference[11] might also have interesting papers, as well as for example ACM's archives. As for individuals, one of the most connected researchers in this field is Balachander Krishnamurthy[12], who has worked with several groups looking at privacy in both online social networks (OSNs) and general websites.

.SE themselves have written papers analyzing the technical state of services connected to .se domains. While they haven't concentrated on exploring the web services connected to these domains, they do offer some groundwork in terms of selecting and grouping Swedish domains as well as looking at Google Analytics coverage.[10, 11] .SE's Internet Fund[13] has also funded work on discussing and defining online privacy, aimed at those working with or developing systems that handle personal data, often with some kind of internet connection.[1]

Media have made reports regarding mass surveillance, especially by the United States intelligence agency National Security Agency (NSA)[14], but so far few papers seem to have been written. There are also reports on what data private companies are collecting, in part by their online efforts, and how they are packaging it for resale. While media reports aren't academic papers, they provide an up to date source of information needed in explaning parts of the thesis subject.

[8] http://acm.org/
[9] http://sigcomm.org/
[10] http://sigcomm.org/events/imc-conference
[11] http://pam2014.cs.unm.edu/
[12] http://www2.research.att.com/~bala/papers/
[13] http://www.internetfonden.se/
[14] http://www.nsa.gov/

Time plan

The time plan describes the activities and milestones of the work with the resolution of a week. Dates for planned final seminar are included. For degree projects on advanced level (e.g. Master level) dates for half-time checkpoint are also included. For these degree projects the expected results for the half-time checkpoint are also explicitly described. This plan is updated in cooperation with the tutor.

Completed milestones

2014-03-07 Initial subject discussion meeting at .SE, with company supervisor Staffan Hagnell.

2014-03-18 First subject draft ready and sent to examiner, company supervisors and other interested parties.

2014-03-31 Subject draft approved by examiner.

Planned milestones

2014-W15 Finalize planning report.

2014-W15 Start software development efforts.

2014-W19 Half time evaluation. Have preliminary results ready, as a progress indicator.

2014-W23 Thesis draft for supervisors to review, then revise according to comments.

2014-W24 Thesis draft for the examiner to review, then revise according to comments.

2014-W25 Thesis draft for the opposition/peer review, then revise according to comments.

2014-W28 Thesis approval for presentation.

Unplanned milestones

Opponent/peer A student in the same field is required for the opponent/peer review.

Presentation The presentation date is required to be during the regular semester period. The fall semester starts 2014-09-01.

Publication Thesis must be submitted to LiU E-Press after the presentation.

Evalutation Writing of an individual evaluation report, which is then discussed with the examiner and supervisors.

Nomenclature

.SE	The Internet Infrastructure Foundation. An independent organization for the benefit of the public that promotes the positive development of the internet in Sweden. .SE is responsible for the .se top level domain.
.se	The country code top level domain name for Sweden.
CDN	Content delivery network
Content	Information and data that is presented to the user. Includes text, images, video and sound.
Content delivery network	(CDN) The speed at which data can be delivered is dependant on distance between the user and the server. To reduce latency and download times, a content delivery network places multiple servers with the same content in strategic locations, both geographic and network toplolgy wise, closer to groups of users. For example, a CDN could deploy servers in Europe, the US and Australia, and reduce loading speed by setting up the system to automatically use the closest location.
Domain name	A domain name is a human-readable way to navigate to a service on the internet: example.com. Fully qualified domain names (FQDN) have at least two parts - the top level domain name (TLD) and the second-level domain name - but oftentimes more depending on TLD rules and organizational units. Domains are also used, for example, as logical entities in regards to security and privacy scopes on the web, often implemented as same-origin policies. As an example, HTTP cookies are bound to domain that set them.

External resource	A resource downloaded from a domain other than the page that requested it was served from.
External service	A third party service that delives some kind of resource to the user's browser. The service itself can vary from showing additional information and content, to ads and hidden trackers. External services include file hosting services, CDNs, advertisting networks, statistics and analytics collectors, and third party content.
Resource	An entity external to the HTML page that requested it. Types of resources include images, video, audio, CSS, javascript and flash animations.
Third-party content	Content served by another organization than the organization serving the explicitly requested web page. Also see external resource.
Third-party service	A service provided by an organization other than the explicitly requested service. Also see external service.
Tracker	An resource external to the visited page, which upon access receives information about the user's system and the page that requested it. Basic information in the HTTP request to the resource URL includes user agent (browser vendor, type and version down to the patch level, operating system, sometimes hardware type) referer (the full URL of page that requested the resource), an etag (unique string identifying the data from a previous request to the same resource URL) and cookies (previously set by the tracker).
Web browser	Or browser. Software a user utilizes to retrieve, present and traverse information from the web.
Web service	A function performed on the internet, and in this document specifically web sites with a specific purpose directed towards human users. This includes search engines, social networks, online messaging and email as well as content sites such as news sites and blogs.
Web site	A collection of web pages under the same organization or topic. Often all web pages on a domain is considered a site, but a single domain can also contain multiple sites.

Bibliography

[1] Markus Bylund. *Personlig integritet på nätet*. FORES, 1 edition, 2013.

[2] Peter Eckersley. How unique is your web browser? Technical report, Electronig Frontier Foundation, 2009.

[3] Anja Feldmann, Nils Kammenhuber, Olaf Maennel, Bruce Maggs, Roberto De Prisco, and Ravi Sundaram. A methodology for estimating interdomain web traffic demand. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, pages 322–335, New York, NY, USA, 2004. ACM.

[4] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Best paper – follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 141–148, New York, NY, USA, 2013. ACM.

[5] Georgios Kontaxis, Michalis Polychronakis, Angelos D. Keromytis, and Evangelos P. Markatos. Privacy-preserving social plugins. In *Proceedings of the 21st USENIX Conference on Security Symposium*, Security'12, pages 30–30, Berkeley, CA, USA, 2012. USENIX Association.

[6] Balachander Krishnamurthy and Craig E. Wills. Analyzing factors that influence end-to-end web performance. In *Proceedings of the 9th International World Wide Web Conference on Computer Networks : The International Journal of Computer and Telecommunications Netowrking*, pages 17–32, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.

[7] Balachander Krishnamurthy and Craig E. Wills. Cat and mouse: Content delivery tradeoffs in web access. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 337–346, New York, NY, USA, 2006. ACM.

[8] Saurabh Kumar and Mayank Kulkarni. Graph based techniques for user personalization of news streams. In *Proceedings of the 6th ACM India Computing Convention*, Compute '13, pages 12:1–12:7, New York, NY, USA, 2013. ACM.

[9] Anália G. Lourenço and Orlando O. Belo. Catching web crawlers in the act. In *Proceedings of the 6th International Conference on Web Engineering*, ICWE '06, pages 265–272, New York, NY, USA, 2006. ACM.

[10] Anne-Marie Eklund Löwinder and Patrik Wallström. Health status 2011. Technical report, .SE The Internet Infrastructure Foundation, 2011.

[11] Anne-Marie Eklund Löwinder and Patrik Wallström. Health status 2012. Technical report, .SE The Internet Infrastructure Foundation, 2012.

[12] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pages 279–284, New York, NY, USA, 2013. ACM.

[13] Arnold Roosendaal. Facebook tracks and traces everyone: Like this! Research Paper 03/2011, Tilburg Law School Legal Studies, November 30, 2010.

[14] .SE. Annual report, 2013.

[15] Pang-Ning Tan and Vipin Kumar. Discovery of web robot sessions based on their navigational patterns. *Data Min. Knowl. Discov.*, 6(1):9–35, January 2002.