

Inference detection technology for Web 2.0

Richard Chow, Philippe Golle, Jessica Staddon

Palo Alto Research Center

{rchow, pgolle, staddon}@parc.com

The current success of Web 2.0 applications is driven by the supply of Web content. Individuals, corporations, governing institutions, diverse organizations have more and more of their data on the web in order to leverage and participate in Web 2.0 technologies. A key aspect of Web 2.0 is empowering even small web sites, the long tail of the Internet, to collaboratively generate and manage their data.

This proliferation of web content represents a data goldmine when armed with the necessary information retrieval tools, and indeed, information has never been easier to find. Search engines allow easy access to the vast amounts of information available on the Web. Online data repositories, newspapers, public records, personal web pages, blogs, etc., make it easy and convenient to look up facts, keep up with events and catch up with people.

On the flip side, information has never been harder to hide, and the potential for privacy compromise is high. With the help of a search engine or web information integration tool, e.g. ZoomInfo, one can easily infer facts, reconstruct events and piece together identities from fragments of information collected from disparate sources. Protecting information requires hiding not only the information itself, but also the myriad of clues that might indirectly lead to it. The failures of individuals to identify those clues are well chronicled. For example, news articles abound about job loss or censure as the result of blogs that were intended to be anonymous (see, for example, the 8/15/2004 Washington Post article about former senate clerk, Jessica Cutler).

It is our position that the very abundance of information and information retrieval tools that together are the source of the problem, are also the solution. Privacy is achieved when the information that is available about an individual is also true of many other individuals. If so, the information is not uniquely identifying. The web can be efficiently mined to estimate how identifying information is, and hence, the privacy risk of releasing the information. Consider for example, a blogger by the name of "Philippe" who makes a reference to his wedding to "Sanae". A simple Google query of "Philippe Sanae wedding" reveals that this blogger's last name is likely to be "Golle" by examining the top hit returned (<http://www.b-e-f.org/newsletter/june2006/june2006.htm>). Similarly, a blogger "Jessica" who makes mention of her cousin "Neilly" and brother "Woozle" can be identifying by reviewing the *only* active site returned by the Google query, "Jessica Woozle Neilly" (http://wiki.hypertwins.org/index.php/Carolina_Friends_School).

These two examples illustrate the potential of the web to provide privacy in Web 2.0 applications. Prior to posting a blog entry, for example, a tool might extract keywords from the entry and pool them with keywords extracted from previous posts to construct Google queries. The hits received in response to the queries can then be examined for references to the blogger. If such references are found, the post might be flagged as a privacy risk.

UNCLASSIFIED//FOR OFFICIAL USE ONLY

(U//FOUO) The [redacted] Family

(U//FOUO) [redacted] is a member of a large and wealthy Saudi family. The family patriarch [redacted] came to the kingdom from Hadramout (South Yemen) sometime around 1930.¹

- In Saudi Arabia [redacted] father became a construction magnate, completing prestigious projects such as the renovation of the holy mosques in Mecca and Medina. As a result, the [redacted] are a highly respected family both within the Saudi royal household and with the public.

(U//FOUO) There is some confusion as to the total number of [redacted] siblings.

- Some cite that he is the youngest of some 20 sons,² while others claim he is the seventh son.³
- The total number of his siblings might be 50,⁴ 52,⁵ or 54.⁶ In an interview [redacted] seemed unsure as well, citing that he had 25 brothers—although he could remember the names of only 20.⁷
- Nearly all of these siblings are half-brothers or half-sisters, as [redacted] father had multiple wives. [redacted] is cited as having only one son.⁸

(U//FOUO) The [redacted] family has denounced [redacted] repeatedly.

- In 1994, the [redacted] family issued a statement expressing its "regret, denunciation and condemnation of all acts that [redacted] may have committed, which we do not condone and we reject."⁹
- After the attacks on the USS on September 11, 2001, the current head of the family [redacted]

Google Web Images Video News Maps Desktop more »

saudi magnate half-brother Search Advanced Search Preferences

Web Results 1 - 10 of about 555 for **saudi magnate half-brother**.

Bin Laden's half brother I'll pay for defense - U.S. Security ...
 Osama bin Laden's **half-brother** Yeslam Binladin poses at the **Saudi** ... daughters of the late **Saudi** construction **magnate** Mohammed bin Laden, who had 22 wives. ...
[www.msnbc.msn.com/id/8459947/ - 39k - Cached - Similar pages](#)

JURIST - Paper Chase: Bin Laden half-brother offers to pay for ...
 Bin Laden **half-brother** offers to pay for Osama defense ... of the late **Saudi** construction **magnate** Mohammed bin Laden [Wikipedia profile], who had 22 wives. ...
[jurist.law.pitt.edu/paperchase/2005/07/bin-laden-half-brother-offers-to-pay.php - 20k - Cached - Similar pages](#)

People's Daily Online -- Bin Laden's brother to pay for defence
 Yeslam and Osama are among 21 sons and daughters of the late **Saudi** construction **magnate** Mohammed bin Laden, who had 22 wives. Yeslam said he believed his ...
[english.people.com.cn/200507/05/eng20050705_194160.html - 25k - Cached - Similar pages](#)

People's Daily Online -- Bin Laden brother disputes Moore film
 A **half-brother** of Osama bin Laden says he enjoyed most of Michael Moore's ... of the late **Saudi** construction **magnate** Mohammed bin Laden and his 22 wives. ...
[english.people.com.cn/200407/29/eng20040729_151223.html - 17k - Cached - Similar pages](#)

Mideast Dispatch Archive: As bombs rock London, Bin Laden's
 ... of the late **Saudi** construction **magnate** Mohammad Bin Laden, who had 22 wives. ...
 Yeslam said in an interview he believed his **half brother**, thought to be ...
[www.tomgrossmedia.com/mideastdispatches/archives/000390.html - 18k -](#)

Figure 1: A single Google query using keywords extracted from the redacted document on the left reveals the subject of the original document.

There is evidence that the web can form the basis of a privacy tool for a broad array of content. Consider, for example, a redacted biography (<http://www.judicialwatch.org/archive/2005/osama.pdf>) that was released by the FBI. Prior to publication, the biography was redacted to protect the identity of the person whom it describes. All directly identifying information, such as first and last names, was expunged from the biography. The redacted biography contains only keywords that apply to many individuals, such as "half-brother", "Saudi", "magnate" and "Yemen". None of these keywords is particularly identifying by itself, but in aggregate they allow for near-certain identification of Osama Bin Laden. Indeed, as Figure 1 shows, a Google search for the query "Saudi magnate half-brother" returns in the top 10 results, pages that are all related to the Bin Laden family. This inference, as well as potentially many others, should be anticipated and countered in a thorough redaction process.

The need to protect secret information from unwanted inferences extends far beyond the FBI. In addition to intelligence agencies and the military, numerous government agencies, businesses and individuals face the problem of insulating their secrets from the information they disclose publicly. In the litigation industry for example, information protected by client-attorney privilege must be redacted from documents prior to disclosure. In the healthcare industry, it is common practice and mandated by some US state laws, to redact sensitive information (such as HIV status, drug or alcohol abuse and mental health conditions) from medical records prior to releasing them.

In all these instances, the problem is not access control, but inference control. Assuming the existence of mechanisms to control access to a subset of information, the problem is to determine what information can be released publicly without compromising certain secrets; that is, what that subset of information should be. What makes this problem difficult is the quantity and complexity of inferences that arise when published data is combined with, and interpreted against, the backdrop of public knowledge and outside data.

The Web-based approach proposed here is based on the belief that the Web is an excellent proxy for public knowledge, since it encapsulates a large fraction of that knowledge (though certainly not all). Further, the dynamic nature of the Web reflects the dynamic nature of human knowledge and means that the inferences detected today may be different from those drawn yesterday. The likelihood of certain inferences can thus be estimated automatically, at any point in time, by issuing search queries to the Web. In our vision, generic tools would use the Web to detect and alert the user to unwanted inferences.