DARPA Memex Project Erodes Internet Privacy

Christopher Furton

Syracuse University

Abstract

In February of 2014, the Defense Advanced Research Projects Agency (DARPA) announced the Memex Project that is currently being used by Federal agencies, law enforcement, and Non-Governmental Organizations (NGOs). The Memex project deploys technology that crawls, indexes, analyzes, extracts, and provides search functionality across the entire Internet including the criminal underground referred to as the Dark Net. Despite the good intention of DARPA, the Memex tool raises several privacy concerns such as scope of use, oversight and transparency, data retention, and information security. With this powerful big data capability, precautions must be taken to protect citizen's privacy rights.

DARPA Memex Project Further Erodes Internet Privacy

What is Memex?

The Defense Advanced Research Projects Agency (DARPA) announced plans to create a project known as Memex on February 9th, 2014 (DARPA, 2014, p. 1). A further look into the Broad Agency Announcement (BAA) shows that DARPA is looking for proposals from industry to "maintain technological superiority in the area of content indexing and web search on the Internet" (DARPA: Information Innovation Office, 2014, p. 4). DARPA identifies a problem with current web search functionality stating that it has limitations on what gets indexed and the richness of available details. For government researchers and law enforcement personnel, current methods used involve manual searching by input of exact information one entry at a time. Further analysis must be done to organize or aggregate beyond a list of links (DARPA: Information Innovation Office, 2014, p. 4).

DARPA plans to solve this problem with the Memex Project by developing technologies that "provide the mechanisms for content discovery, information extraction, information retrieval, user collaboration, and other areas need to address distributed aggregation, analysis, and presentation of web content" (DARPA: Information Innovation Office, 2014, p. 5). To accomplish this, DARPA has divided the work into three technical areas: domain-specific indexing, domain-specific search, and applications. DARPA specifies the need for technology to reach beyond traditional content, specifically naming the Dark Web as a target. The Dark Web refers to the large mass of Internet content not relatively accessible through search engines often requiring special encryption software to access (Chandler, n.d.).

The first technical area DARPA is interested in is domain-specific indexing. This technical area focuses on developing a highly scalable web crawling capability with both content discovery and information extraction. This crawling process will provide automated link discovery including obfuscated links, discovery of deep and dark web content, and hidden services – the function of providing web services such as chat or web page hosting on the Dark Web. Additionally, this capability will include counter-crawling measures such as paywalls or member-only areas, crawler bans, and even human detection. Lastly, this capability must also be able to extract information and include normalization of heterogeneous data, natural language processing for translation, image analysis, extraction of multimedia, and several other functions (DARPA: Information Innovation Office, 2014, p. 7).

The second technical area DARPA is interested in is domain-specific searching. This capability is not the same as current commercial web searching; instead, it will have configurable interfaces into web content indexed by the first technical area. The interfaces, as outlined in the BAA, may include conceptually aggregated results, conceptually connected content, task relevant facets, implicit collaboration for enriched content, explicit collaboration with shared tags, and several other capabilities. Lastly, this technical area will include a query language so that DARPA personnel may modify instructions for the crawlers and information extraction algorithms (DARPA: Information Innovation Office, 2014, p. 7).

The last technical area DARPA is interested in is generically referred to as "applications." This technical area is where system-level concepts of operation and use cases are developed. The utility of the system must be able to evolve over time based off the needs of the Department of Defense and other agencies, and it involves the development of possible new content domains including missing persons, found data, and counterfeit goods. Specified in this technical area is that the integration, testing, and evaluation is to be performed on the open public Internet (DARPA: Information Innovation Office, 2014, pp. 8-9).

Stated Purpose

According to Wired Magazine (2015), Memex's purpose is to uncover patterns and relationships in online data for law enforcement and others who track illegal activity (para. 1). Memex uses automated methods to analyze content in order to uncover hidden relationships between data points. Additionally, it helps researchers determine how much of the dark web's traffic is related to hidden services where content could be indexed. Furthermore, Memex can help investigators understand the turnover of sites, specifically, the relationship between sites when one shuts down and a seemingly unrelated site opens up (Zetter, 2015).

DARPA has gone through great measures to specify that the Memex Project is aimed at indexing "domain-specific" content and gives the example of Human Trafficking – both labor and sex - in the BAA (DARPA: Information Innovation Office, 2014, p. 4). At least one instance of Memex in action has been documented: the New York District Attorney's Office claims that an experimental set of Internet search tools is part of the prosecutor's arsenal that helped secure a sex trafficking conviction. In this instance, Memex was used to scour the Internet looking for advertisements used to lure victims into servitude and to promote their sexual exploitation (Greenemeier, 2015, para. 2-3).

DARPA is an agency under the Department of Defense; however, the stated purpose of Memex has a direct correlation to law enforcement. DARPA has confirmed that in August of 2014, several beta testers were approved to use Memex including two district attorney's offices, a law enforcement group, and a nongovernmental organization (NGO). The next set of tests are expected to begin in early 2015 and include federal and district prosecutors, regional and national law enforcement, and multiple NGOs. Every quarter, DARPA wants to expand user testing until they are comfortable handing the tool over to law enforcement agencies and prosecutors. Eventually, the plan is to have the Memex capability installed locally at law enforcement agencies and to ensure police could access the software from anywhere (Greenemeier, 2015).

Hypothetical Privacy Abuse

The Dark Web provides an avenue for cyber criminals, human traffickers, child pornographers, and other criminals to conduct business (Bradbury, 2014). Additionally, it also provides an avenue for planning and coordination of terrorism activities posing a threat to National Security (Sachan, 2012). It is only logical that the United States Government would be interested in gaining better insight and surveillance into the Dark Web for both national security and law enforcement reasons; however, with that comes the need to protect citizen's privacy. Memex, as an international "Big Data" tool, provides the capability for content discovery, index, search, aggregation, and extraction on a very large scale introducing a swarm of information privacy concerns.

Scope of Use

One major privacy concern involves the breadth of data collected by Memex's web crawling indexers. Since the web crawlers are designed to ignore the content owner's explicit crawler prohibitions – often done by robots.txt files – as well as penetrating paywalls and membership areas, it is highly feasible that content intended to be private will get vacuumed into the Memex system. Therefore, Memex can be seen as another attempt by the United States Government to increase the size of the figurative information haystack in order to find more needles. Despite the usage of "domain-specific" collection techniques, it is inevitable that non-domain data will get introduced into Memex.

Another privacy concern involves the scope of users being allowed access to Memex-siphoned data. As mentioned earlier in this paper, DARPA has already given access to one private Non-Governmental Organization (NGO) which introduces concerns, as noted by Mueller (2010) in regards to private organizations and content regulation, about governance arrangements that may "avoid substantive and due process rights" (p. 213). With DARPA planning to extend access to many more NGOs, citizens should be concerned about who has access to this potentially massive trove of well-organized and readily accessible private information.

Transparency and Oversight

By design, Memex provides for dynamic on-demand content domain generation that enables Federal agencies, local law enforcement, and other organizations to select surveillance zones without judicial oversight. The content domain example given by DARPA is for human trafficking; however, as noted earlier in this paper, web crawlers have the ability to take commands from controllers to add or modify content domains. This functionality is powerful as it potentially allows investigators and private organizations access to personal information collected without oversight and may compromise citizen's due process protections. Although currently unknown, it seems feasible that a law-abiding citizen, who posts a request online for a consensual intimate meeting, may get her information indexed under Memex's human trafficking content domain. This functionality without oversight and transparency should invoke concerns from privacy advocates.

Data Safeguards and Retention

As with any "big data" system, details over how data is collected, stored, and transmitted is a significant concern. Hackers, identity thieves, and foreign governments may have the potential and desire to target Memex's web crawlers or central repositories. These systems must be protected from possible exploitation or hijacking considering the amount of sensitive information indexed. The data collected, stored, and transmitted – as well as the webcrawlers themselves – must have safeguards protecting them from compromise by malicious actors intent on invading citizen's privacy or committing crimes.

Along with concerns over safeguarding data is the duration of time that collected information will be retained. In a related situation where governmental authorities built databases of citizen's vehicle location information via Automated License Plate Recognition (ALPR) technology, privacy advocates argued over concerns of data retention (Lynch, 2013). Advocates have already filed lawsuits attempting to make usage of this technology more transparent (Electronic Frontier Foundation, 2013). Similarly, privacy advocates will likely be concerned with Memex's data retention policies.

Conclusion and the Author's Personal Reflections

Based off currently published public information, DARPA's Memex Program is a powerful tool built with the best of intentions: to protect national security and to combat cybercrimes. As a technical feat, its ability to crawl the vast scope of the Internet, evade paywalls and member-only areas, perform in-depth analytical functions, and extract information is impressive. The breadth of deployment of this tool to include law enforcement, district attorneys, and NGOs introduces the possibility of abuse. Big Data tools that collect citizen's information will inherently encroach on privacy rights and Constitutional protections. Memex invokes discussion over Internet privacy and whether any and all information posted online is considered public and open for collection or analysis. Regardless, usage of a tool like Memex should require strict usage rules that outline scope, additional oversight and transparency, and must be properly secured. Additionally, information collected by Memex about citizens should be accessible under current freedom of information act laws. Memex is not the first – and will not be the last – information tool that challenges citizen’s privacy rights.

References

Bradbury, D. (2014). Unveiling the dark web. *Network Security*, 14-17.

Chandler, N. (n.d.). *How the deep web works*. Retrieved from How Stuff Works: Computer: http://computer.howstuffworks.com/internet/basics/how-the-deep-web-works.htm

DARPA. (2014, 02 09). *DARPA:News Events*. Retrieved from Memex aims to create a new paradigm for domain-specific search.: http://www.darpa.mil/newsevents/releases/2014/02/09.aspx

DARPA: Information Innovation Office. (2014). *Broad Agency Announcement: Memex.* Retrieved from http://go.usa.gov/BBc5

Electronic Frontier Foundation. (2013, 05 06). EFF and ACLU Sue LA Law-Enforcement Agencies Over License-Plate Reader Records. *EFF Press Release*.

Greenemeier, L. (2015, 02 08). Human traffickers caught on hidden Internet. *Scientific American*. Retrieved from http://www.scientificamerican.com/article/human-traffickers-caught-on-hidden-internet/

Zetter, K. (2015, 02 15). Darpa is developing a search engine for the dark web. *Wired*. Retrieved from http://www.wired.com/2015/02/darpa-memex-dark-web/

Lynch, J. (2013, 05 06). Automated License Plate Readers Threaten Our Privacy. *EFF DeepLinks*.

Mueller, M. (2010). *Networks and States: The Global Politics of Internet Governance.* Massachusettes Institute of Technology.

Sachan, A. (2012). Countering terrorism through Dark Web analysis. *IEEE*.