How Much Do We Reveal Through Metadata? An Assessment of Online Privacy

Jonathan Gillett • Joseph Heron • Daniel Smullen • Jeremy Bradbury
Faculty of Science • University of Ontario Institute of Technology • Oshawa, Ontario, Canada
jonathan.gillett@uoit.net, joseph.heron@uoit.net, daniel.smullen@uoit.net, jeremy.bradbury@uoit.ca

1. Motivation

- Revelations about government involvement in mass information collection has garnered a lot of media attention regarding online privacy [1, 2].
- Many people are concerned about the data that is being collected by government and private interests however, most people are unclear about the type and volume and data that can be traced back to them.
- As people go about their ordinary daily browsing activities, personally identifiable information (PII) contained in metadata is pervasive [3].

Research Goal:

To determine differences in perception between the amount of PII revealed through Internet browsing and what our sample population perceives that they reveal.

2. Methodology

Study Entrance

Entrance questionnaire and informed consent process.

Workshop for participants to configure browsers to use our proxy.

Multiple browsing devices can be configured to use the proxy.

Data Capture

Internet traffic is collected during a two hour browsing session.

Data is automatically anonymized, filtering out PII.

Study Exit

Participants complete an exit questionnaire after their session.

Proxy settings are reset to previous configuration.

Access to the proxy will be revoked.

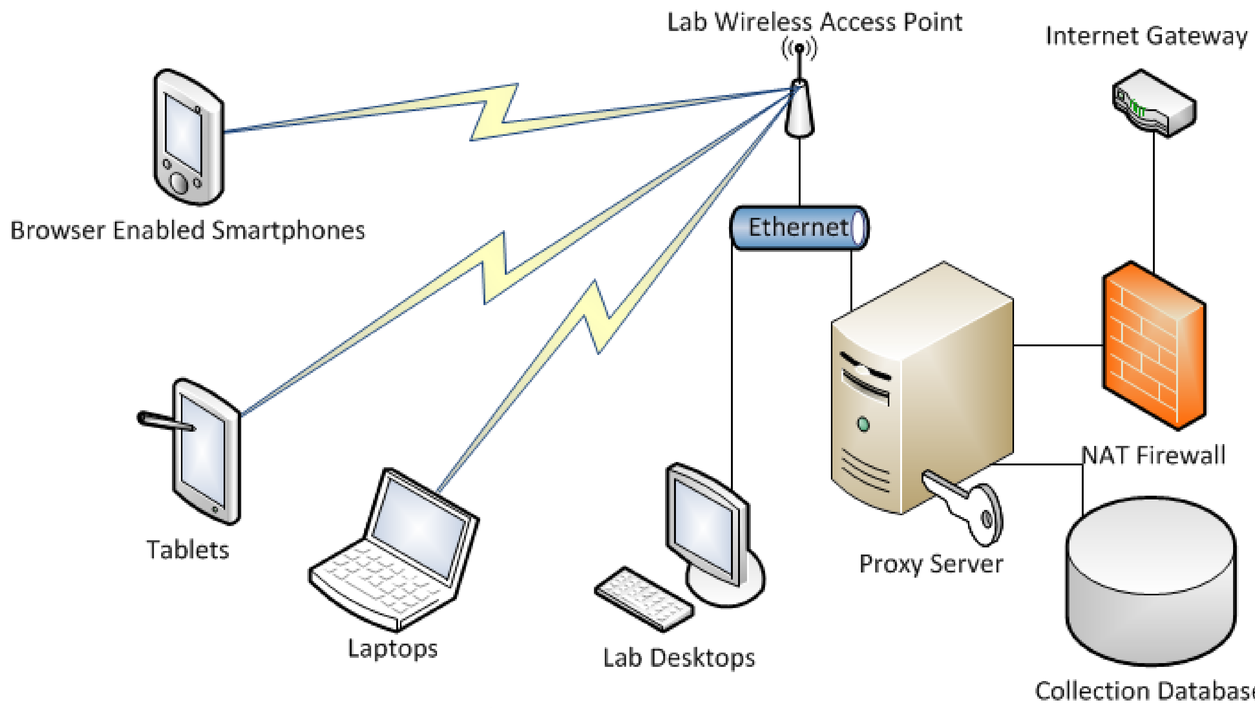
Data Analysis

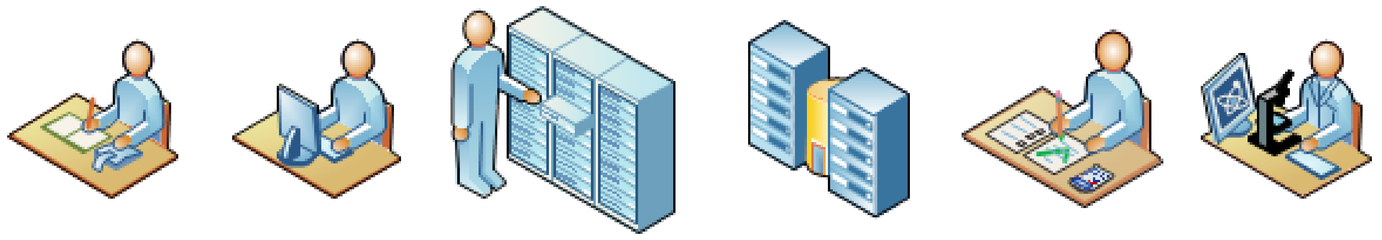
Data is analyzed to determine the amount and type of PII found.

Any PII that is not anonymized automatically is anonymized manually.

Statistical analysis applied to results, and compared to survey results.

3. Data Capture





Informed consent, entrance survey. ➡ **Browser configuration, data collection.** ➡ **Data anonymization.** ➡ **Data storage.** ➡ **Exit survey, configuration reset.** ➡ **Statistical analysis.**

4. Preliminary Results

- Currently undergoing REB evaluation.
- Our sample population - students at UOIT - is likely to have different perceptions to the wider population of Internet users.
  - However, there are a lot of young students between the age of 18-35 who browse the Internet, so this is a good place to start.
- Anonymizing PII data is challenging and can't be done with 100% accuracy using automated processes (regex, text parsing, metadata analysis tools).
- We expect there will likely be a big difference between the amount of PII revealed and the amount which participants think they are revealing – metadata is everywhere!

5. References

[1] G. Greenwald, E. Macaskill, S. Ackerman, " NSA collecting phone records of millions of Verizon customers daily," The Guardian, Vol. 6, no. 5, pp. 13, 2013. URL: http://www.addisonlibrary.org/assets/1/newsletter_pdf/APL_Pol__Group_Articles_July.pdf. [Accessed 10/1/2013]

[2] G. Greenwald, "XKeyscore: NSA tool collects 'nearly everything a user does on the internet'," The Guardian, Vol. 6, no. , pp. , 2013. URL: http://www.addisonlibrary.org/assets/1/newsletter_pdf/APL_Pol__Group_Articles_July.pdf. [Accessed 10/1/2013]

[3] Greschbach, B.; Kreitz, G.; Buchegger, S., "The devil is in the metadata — New privacy challenges in Decentralised Online Social Networks," Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on , vol., no., pp.333,339, 19-23 March 2012 doi: 10.1109/PerComW.2012.6197506 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6197506&isnumber=6197445

Consortium for Software Engineering Research (CSER) Fall 2013 • Markham, Ontario, Canada

Funding provided by: NSERC CRSNG