1. Project Description

A. Description of Project

Most websites track their users. A popular privacy tool, Ghostery, reports eighteen trackers on CNN.com, five trackers on science.com, and one tracker on unomaha.edu. A tracker is code from a third-party company, typically an advertising agency [1] or a data marketer [2], that silently tracks our path around the internet. The tracking code records which websites we visit, in what geographic location, and during what time of day. Unsurprisingly, as everyday internet users learn the extent of this tracking, they have significant concerns [3].

To block a third-party tracker, we must know the domain that is serving the tracking code. For example, once we know that Google Analytics uses http://www.google-analytics.com as its tracking server, we can instruct the browser to block any connections to that host. The current best solution involves a static blocklist of known trackers; it is updated manually by a group of volunteers [4], and has a success rate of 60-80% [5]. While the current blocklist method works, it has the drawback of stale data: as soon as a blocklist is released, trackers can re-tune their methods.

In this project, we are interested in automated methods to detect third-party trackers. In order to achieve this, we will first catalog the third-party requests that are being made from popular websites. Second, we will train a machine-learning classifier to determine whether a third-party domain is performing tracking.

From this study, we expect to create the following new knowledge: (a) determine which features are suggestive of a tracking domain, (b) create extraction methods for those features, and (c) compare the efficacy of automated tracking detection to static tracking methods. To accelerate our project, we will extend the work of Stanford researchers Bay et al. in evaluating machine-learning techniques to identify web trackers [9].

B. Activities/Methodology

Phase 1. Build data collection architecture
In the first phase, we will build our data-collection framework. It takes the URL of a website, loads the page, and then reports (a) any new cookies set by HTTP headers or JavaScript, and (b) a list of all resources that were requested along with the response headers. In terms of implementation, PhantomJS [10] will provide a web browser that can be controlled using HTTP on a port.

In this phase, the bulk of the work is in writing the software that controls PhantomJS. This software will submit URLs to PhantomJS, fetch the results, and then store the results in a database for later processing.

Phase 2: Create feature extractors
In the second phase, we will write the feature extractors. Feature extractors are small functions that take raw data as an input, and then return features as binary or scalar values. Listing 1 shows our initially-planned set of features. As we explore common websites and tracker domains, this list may grow or shrink depending on our observations.

Phase 3: Process candidate websites
In the third phase, we will use the infrastructure that was created in Phase 1 to crawl the target websites. We are targeting the most highly-trafficked websites, as reported by Alexa [6].
Phase 4: Generate model for machine learning
In the fourth phase, we will select those features that provide the highest accuracy for classification. The third-party domains will be split 80% for training and 20% for verification. For training and validation, we will employ Microsoft's list of known trackers [7]. For the classifier, we will use the Elastic Net algorithm [8]. The Elastic Net algorithm was used by Bau et al and employing it will allow us to make meaningful comparisons between our results.

Phase 5: Evaluate accuracy of model
In the fifth phase, we will verify our model against the 20% reserved set to determine its precision as a function of the false-positive rate.

Listing 1: Initial set of features

<table>
<thead>
<tr>
<th>JavaScript analysis</th>
<th>Cookie</th>
<th>HTTP Headers</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Evidence of font profiling</td>
<td>• Request site</td>
<td>• ETAG</td>
</tr>
<tr>
<td>• Evidence of system profiling</td>
<td>• Cookie's domain</td>
<td>• 301 Redirect</td>
</tr>
<tr>
<td>• Presence of obfuscation techniques</td>
<td>• Cookie's TTL</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td>DNS</td>
<td>Resources</td>
</tr>
<tr>
<td>• Extension</td>
<td>• CNAME record for all first-party subdomains.</td>
<td>• HTTP response headers</td>
</tr>
<tr>
<td>• Mime</td>
<td></td>
<td>(including cookie, expire, and ETAG)</td>
</tr>
<tr>
<td>• Resolution</td>
<td></td>
<td>• Resource URL</td>
</tr>
<tr>
<td>• URL Parameters</td>
<td></td>
<td>• MD5 of response body</td>
</tr>
<tr>
<td>• HTTP Cookie</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. Project Timeline

<table>
<thead>
<tr>
<th>Weeks 1 - 4</th>
<th>Phase 1. Build data collection architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weeks 5 - 7</td>
<td>Phase 2: Create feature extractors</td>
</tr>
<tr>
<td>Weeks 8 - 10</td>
<td>Phase 3: Process candidate websites</td>
</tr>
<tr>
<td>Weeks 11 - 13</td>
<td>Phase 4: Generate model for machine learning</td>
</tr>
<tr>
<td>Weeks 14 - 16</td>
<td>Phase 5: Evaluate accuracy of model</td>
</tr>
</tbody>
</table>

D. Student/Faculty Mentor Roles

As researcher, Eric Edens will implement and deploy the tools required to collect and analyze the target websites. As mentor, Dr. William Mahoney will provide guidance in malware analysis and language analysis. Dr. Mahoney and Eric Edens will check-in weekly to monitor the progress of the project and to analyze data.
2. Budget Justification

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Stipend</td>
<td>$2,000</td>
</tr>
<tr>
<td>Research Supplies</td>
<td>$41</td>
</tr>
<tr>
<td><strong>Student Total</strong></td>
<td><strong>$2,041</strong></td>
</tr>
</tbody>
</table>

**Student Stipend**
We estimate that the project will require 16 weeks of work, with each week requiring 14 hours of work. In total, this is 224 hours of work, which equates to a wage of approximately $8.93 / hour. As there are no external dependencies, the project can begin as soon as we are awarded the grant.

**Materials and Supplies**
We have allotted $41 for research supplies. Two lab notebooks and two USB drives will be used for data collection and archival. In addition, we have identified important reference materials; these are not included in the budget, as they are available through the UNO library.

**Research Supplies**
- Two Lab Notebooks, VELA Series-A Computation Laboratory Notebook, $27
- Two USB Drives, Kingston Digital DataTraveler, $14

**Reference Materials**
- *Learning From Data* (Abu-Mostafa, Magdon-Ismail), 2012
- *Artificial Intelligence* (Russell, Norvig), 2009
- *Pattern Recognition and Machine Learning* (Bishop), 2007

3. References

October 14, 2013

Fuse Panel;

Please carefully review this proposal from Mr. Eric Edens, as the research is in an interesting and “hot” area of Information Assurance. All of the recent news concerning the NSA tracking individuals through the use of their cell phone “meta data” has cast a light on how much privacy Americans (don’t) have. A disturbing trend on the Internet, which is not receiving much attention, is the tracking of individuals via monitoring their site visits. Tracking systems collect data that is then often mingled with information you might voluntarily share while, for instance, looking for airfares or while shopping. This mixture makes it possible for tracking systems to determine what sites you visit and what your personal interests are. Recently this technology has extended into the browsing done by cell phones as well. Frequently, Android applications will now ask for permission to access the cell’s unique identity number, which is then collected along with the usage history and thus links all of the information to an individual.

But all of this is background. Eric is a good student in my Programming Languages class and I would like for him to have this award so that he can conduct research into the tracking realm. It is an area that I am aware of but currently know little about, and a result of his research will be more information that we can utilize in classes such as our Introduction to IA general education class.

Sincerely,

William R. Mahoney, PhD
Associate Professor
School of Interdisciplinary Informatics