ANALYSIS ON WEB COOKIES

WEB COOKIE OVERVIEW:

Cookie is a formatted string consisting of the semi-colon and key-value pairs. A simple cookie would appear as follows:

Name=Value; Host=example.com; Path=/account;
Expires=Tue, 1 Dec 2018 10:12:05 UTC; Secure;

Name. The name attribute contains the name given to a cookie sent by a particular server. This uniquely identifies cookies to a particular server.

Value. The value attribute contains the data the cookie is responsible for transmitting between client and server. Value data may be clear text, but is generally encrypted, or obfuscated for security and privacy reasons.

Host. The host attribute identifies the cookie's origin server. This allows a browser to send cookies back to the proper server during subsequent communication. It also distinguishes 1 st and 3 rd party cookies.

Path. The path attribute restricts when a browser sends a cookie back to a host. The value in the path attribute must exist in the URL of the website being requested by the browser.

Expires. The expiration attribute contains a DateTime string announcing when the cookie should be invalidated. The value in the expires attribute distinguishes session and persistent cookies. Note, Firefox converts the DateTime string into an expiry time in seconds.

Secure. The secure attribute is a flag that specifies whether a cookie is transmitted securely via SSL and HTTPS.

HttpOnly. The HttpOnly attribute is a flag that specifies whether a cookie can be accessed programmatically client side.

isDomain. The isDomain attribute is a flag that specifies if a cookie should be sent to a host for URL requests to any subdomain within a site. Note, isDomain is determined by the host attribute and is specific to Firefox.

ANALYSIS ON WEB COOKIES

COOKIE CHARACTERISTICS:

Due to page request timeouts, our Crawler successfully visited 95,220 (95,311) websites. Note, the set of websites that caused a timeout is likely an artifact of our crawl. This is because website availability is an on/off function with respect to time. Therefore, the availability of the 100K websites we visited is biased by the time our Crawler made a request to a specific site. This speaks to the highly dynamic nature of website availability. Even among the top 100K Alexa websites, downtime is not uncommon.

Throughout the rest of this section, we adopt the nomenclature XX% (YY%) to discuss the results from our two crawl campaigns simultaneously. The first percentage represents the crawl conducted in April of 2015 and the second percentage (i.e., the percentage in parenthesis) reflects the results from the crawl conducted in Novem- ber of 2013.

Cookie Attributes:

There are several attributes, Some of them mentioned below

1. The Name Attribute
2. The Host attribute
3. The Expiration Attribute
4. The Secure Attribute
5. The HttpOnly Attribute
6. The isDomain Attribute
7. The Path Attribute

Security and Privacy Concerns:

In total, 80.6% (79.73%) of the cookies harvested have maximal permissions e.g., a root level path, isDomain set, and secure not set. Note, for the second crawl we also add HttpOnly not set to the definition which changes the percentage from 80.6% to 76.1%. The issue with maximal permission cookies is twofold.

ANALYSIS ON WEB COOKIES

Cookie setting Behaviour:

The 1 st and 3 rd party cookies are set on the browser while visiting a web site defines that website's cookie setting behavior. In the following two subsections we stratify our cookie collection by genre and then again by category to further highlight cookie setting behavior.

I. Setting Behavior by Genre:

The majority of cookies a site sets come from hosts labeled with Technology/internet, Shopping, News/Media, Business/Economy, and Entertainment genres.

II. Cookie Category setting behavior by Genre:

Targeting/Advertising cookies are highest in all genres followed by performance, with only a few Strictly Necessary and Functionality cookies. Note, Government/Legal is the one of the few genres where Performance cookies are set almost as much as Targeting/Advertising.

USER INFORMATION LEAKAGE:

A simple definition of user information leakage in the context of web browsing is a user's unintentional transmission of information about their browsing behavior.

As a user U browses the Internet, various entities gather information about U 0 s behavior. This information comes in two forms: intra-domain and inter-domain. Intra-domain information pertains to what U does within a domain and consists of what intra-domain links are clicked, what U searches for, and other page-related activities. Inter-domain information comes from a user visiting pages of different domains.