March 2010

Privacy Awareness of Web Users

Mihajlo Zeljkovic
Worcester Polytechnic Institute

Major Qualifying Project

CEW - 1001

Worcester Polytechnic Institute

Department of Computer Science

Privacy Awareness of Web Users

Submitted to the Faculty

Of the

Worcester Polytechnic Institute

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

By

Mihajlo Zeljkovic

ABSTRACT

In this project we examine Internet users' Web browsing habits and their level of awareness about privacy on the Web. We created a Web site that describes common ways used online in order to gather information about users without their knowledge. Besides just providing general information to users about this issue, we also show them their own available personal information. Our hope is that this personalized approach will raise their awareness of this important issue.

1. Introduction

Privacy is our ability to selectively reveal information about ourselves. We want to have as much control as possible over what others know about us. Perception of what is considered private varies among people and environments. So, how should the concept of privacy be perceived in the online world?

When browsing the Internet we may have a feeling that we are anonymous. If we do not intentionally reveal some information, nobody can know who we are. This is the impression we believe most Web users have, but the reality is different. One way of identifying users is by using cookies. Cookies are text data a browser uses to store information. Each site can read and set cookies they created. Sites do not have authorization for manipulating cookies created by others.

When we visit a Web site, it is referred to as a first party. Any other site we visit at the same time without our knowledge is known as a third-party site. These sites are used by advertising companies to track users across multiple domains where they read and set cookies. Third-parties gather the information about users from all visited Web sites where they are embedded. This information helps third-parties use behavioral advertising and show content that is more relevant to the users.

Default properties in widely used browsers do not protect people's privacy enough. This project has intention of examining Web users' browsing habits as well as informing users about third-party sites and how they are using information for advertising purposes.

We created a Web site where users can go and be made more aware of what can be inferred about them as individuals. Users are shown a list of sites they visit along all third-party servers found on each. We think that showing users what is pertinent to them will have a stronger impact then simply discussing privacy issues.

In Chapter 2 we describe technologies used and previous work done related to our topic. Chapter 3 describes the overall design of our program and Chapter 4 continues with the exact implementation of it. In Chapter 5 we show how the program works. Chapter 6 presents project results and Chapter 7 derives conclusions from them.

2. Background

As part of the project we developed a Web site whattheyknow.cs.wpi.edu where users can go to, get some of the information that can be inferred about them and provide feedback on their browsing habits. Let us first talk about these overall ideas that we used for building the site: CSS, JavaScript, Perl and MySQL.

2.1 Web Technologies

CSS is used for controlling style and layout of Web pages. In this project we will use its property for displaying links. Every browser has a default property for link to a page that was visited and the one that was not. A page is considered visited if it can be found in the browser's history. Combining this property and JavaScript can give us the list of sites user has visited.

JavaScript is a scripting language used for client side scripting. It can access any element on the client's page and examine its properties. If we create a link for a certain site, we can check its color and determine whether it was visited or not. Doing this for a large number of sites, we can find many that a user has visited.

Perl is a scripting language used for server side scripting. Since it operates on the server side, its content is hidden and cannot be accessed by user. In our project we are using it to lookup a list of third-parties that can be found on each of the visited sites.

MySQL is a database system that consists of a collection of tables that store particular sets of data. We are using it for storing data about participants in our research.

2.2 Third-parties

Diffusion of private information about users as they visit various Web sites through third-party servers has been researched by Balachander Krishnamurthy and Craig Wills. In their paper [1] they analyzed privacy diffusion over an extended period of time and concluded that penetration of the top-10 third-party servers among 1200 popular Web sites increased from 40% in Oct ’05 to 70% in Sep ’08.

In their work about leakage of personally identifiable information via online social networks [2], they discussed about different such leakage occurs and measures to prevent sending data to the third-party servers. One of the measures they suggest is refusing third-party cookies. Users’ view on this option in a browser is one of the questions we address in our project.

2.3 Users’ Awareness

Another protection mechanism is using NAI opt-out cookies [4]. The NAI (Network Advertising Initiative) is a cooperative of online marketing and analytics companies committed to building consumer awareness and establishing responsible business and data management practices and standards. It enables users’ to "opt out" of the behavioral advertising delivered by their member companies. Networks they choose to opt out from will no longer be able to deliver ads tailored to their Web preferences and usage patterns.

Allecia McDonald and Lorrie Cranor tried to get an insight into how much users know about Internet advertising [3]. They interviewed 14 participants in detail in order to better understand their point of view about the issue. Half of them confused cookies with

browser history and none of them was familiar with NAI opt-out cookies. Our project is examining how much Web users' know about these concepts.

2.4 CSS and Browser's History

Browsers maintain history information by default so that previously visited sites can be seen and their link color can be changed from the default when a page containing such a link is shown. Script in [4] uses this CSS property of a link for obtaining the list of visited social networks. We modify this script and use it for obtaining the list of all visited sites.

Quantcast is a Web analytics service that allows users to view audience traffic and demographic data for millions of sites [5]. It has partnerships with large Internet service providers and is able to access their log files and customers' demographic data such as age and gender. This information helps the Quantcast know all sites visited through these Internet service providers and the profile of each visitor.

Combining script from [4] and list of Quantcast top 10k Web sites with percentage of gender in each can give us an estimate of user's gender. This estimate was done in [6]. We use the same formula for determining gender and ask users for a feedback on how accurate it is.

Another Web site that uses similar technique is [7]. It checks for visited sites in different categories, such as search engines, social networks, news, adult entertainment and others. For each site it finds, it checks for the most popular links within it. For example if it determines you went to a news site, it can get a list of some articles you read.

2.5 Techniques for Tracking Users

An alternative to using cookies for tracking Web users includes browser configuration. Research in [8] is using this approach to determine how unique our browser fingerprint is compared to all tested so far. We are using data we have about our participants to do the similar evaluation.

In addition to these techniques for tracking users, there is easy way of retrieving their location. A number of Geo-IP sites available can identify user's location based on IP address. One site like that is used in our project [9].

2.6 Summary

Knowledge we had about previous works done related to our research, techniques for tracking Web users and obtaining the list of sites they visited helped us to come up with the design of our site.

3. Design

We want to show users how much their online privacy is leaked by using sites they are known to visit via the browser history and obtain feedback on their browsing habits. Our intention is to have as many participants as possible. If our program causes inconveniences like having a difficult installation, we think that the number of participants will be significantly lower. This is why we make a site that any user can access with their browser.

3.1 User's Interaction

We inform users about the study and give them some knowledge about online privacy on our home page. If they agree to participate we present data pertinent to them and ask them to fill out a survey.

First, we show them their location and ask for its correctness. After that, we present them their list of visited sites along with the third-party servers that know about them. Based on the sites they visit, we predict their age and gender and ask for a feedback if the prediction was correct. Users are also able to answer our multiple choice questions about their browsing habits and attitudes towards online privacy. They are also able to leave a written comment about their experience.

3.2 Gathering Data

Participation and answering to any of the questions is completely voluntary and users will be informed about it. If they do not want to answer on a particular question for any reason, there it can be left blank. We gather data about users when they decide to participate and update it when they answer the questions.

Besides just collecting data about users' current online behavior, we would also like to educate them about online privacy and possibly influence them to change some of their habits.

4. Implementation

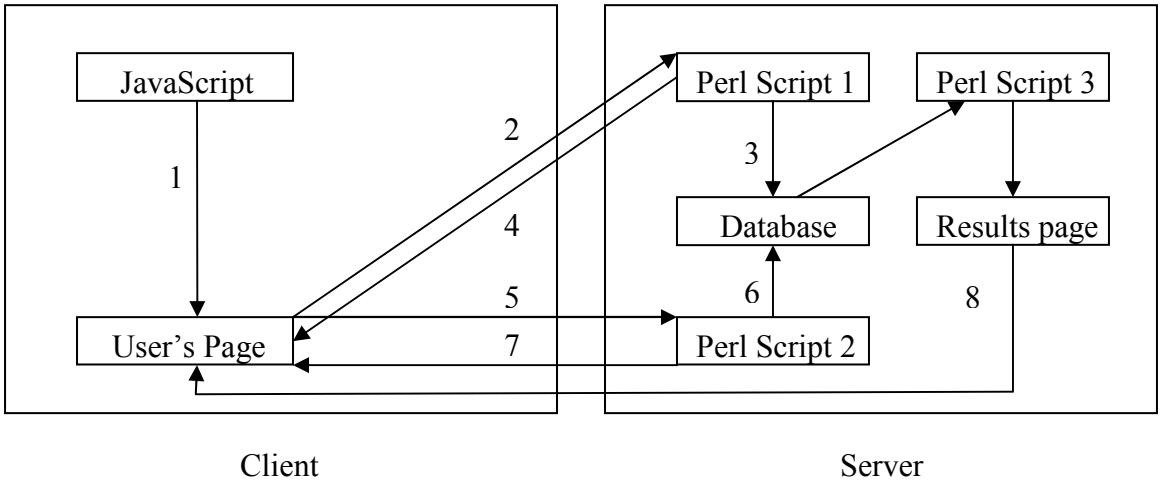

Figure 1 - Site Diagram

4.1 Gathering Data That Can Be Inferred About Users

In Figure 1, we see the diagram of our site whattheyknow.cs.wpi.edu. In step 1 we are using JavaScript for determining list of sites users have visited, their location, browser and list of plug-ins. For obtaining the list of visited sites, we modified and used a JavaScript file that contains functions for doing this [4]. Changes made included different lists of sites that we wanted to test users for. We checked users' history for top 1000 sites on Quantcast list as well as most popular search engines, social networking and adult entertainment sites that were not included in it.

For getting location we used a script from [9] to embed data in user's page, and extracted country, state and city with JavaScript. Retrieving browser and list of plug-ins is done using JavaScript built-in functions. This information is important for determining user's browser uniqueness.

After user agrees to participate in our research in step 2 in Figure 1, list of visited sites and location are sent through post parameters to the Perl script that uses this information to present to users what can be inferred about them.

4.2 Processing Data and Recording New User

For calculating age and gender we are using demographic data from Quantcast. In July '09, we downloaded a file from their site that included percentage of male visitors for each of the top 1000 sites in terms of number of male visitors. For age data we downloaded same kind of data for each of the age groups: 3-12, 13-17, 18-20, 21-24, 25-34, 35-44, 45-54, 55-64 and 65+. In order to calculate age probability we have to include only sites that we have sufficient information about. If we have data for more than two groups missing, we do not take the site into account.

If we have p_1, p_2 … p_n representing percentage of male users from each of the n sites we visited we get:

$$\frac{P(male)}{P(female)} = \frac{\prod_{i=1}^{n} P_i}{\prod_{i=1}^{n} 1 - P_i}$$

$$P(female) = \frac{\prod_{i=1}^{n} 1 - P_i}{\prod_{i=1}^{n} P_i} * P(male)$$

We know that:

$$P(male) + P(female) = 1$$

From here we get:

$$1 - P(male) = \frac{\prod_{i=1}^{n} 1 - P_i}{\prod_{i=1}^{n} P_i} * P(male)$$

$$1 = (1 + \frac{\prod_{i=1}^{n} 1 - P_i}{\prod_{i=1}^{n} P_i}) * P(male)$$

From here we derive the formula:

$$P(male) = \frac{1}{1 + \prod_{i=1}^{n} \frac{1 - p_i}{p_i}}$$

Using similar method we can predict the probability for each age group we have data for. First we calculate:

$$R(x) = \prod_{i=1}^{n} p_{xi} * \frac{p_{xi}}{p_{x-all}}$$

for each of the m groups, where p_{xi} represent percentage of the i-th site audience and $p_{x\text{-}all}$ represents percentage of all internet audience belonging to the group x.

Probability for each group is:

$$P(x) = \frac{R(x)}{\sum_{i=1}^{m} R(i)}$$

In Oct. '09 we ran script that visited home page of each site we have and recorded all currently present third-party servers in a file. Information from this file is used to list all third-parties that know about each site a user has visited.

When users decide to check what is known about them, our Perl script checks if this user already has a record in the MySQL database. If the user is new, their list of sites, location, browser and list of plug-ins are stored in the database in step 3.

4.3 Presenting Information to Users and Getting Feedback

In step 4, users are shown what can be inferred about them and asked to provide a feedback. If they decide to provide a feedback, their answers is sent to the Perl script in step 5, their record in the database is updated in step 6 and in step 7 they get a confirmation that their data has been received. After that, they can decide to see results of all participants in step 8.

4.4 Issues During Implementation

Page that contains the results of all participants is manually produced by executing Perl script periodically. In the beginning, we used the script to directly show this information to users, but as the database become larger, it required too much time to process all the data.

For checking user's identity we used IP address in the beginning. In order to have this project exempt from further review by WPI Institutional Review Board, information had to be recorded in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Another way of identifying users we implemented at one point was using cookies. However we decided against using them so we could assure users that no cookies were used in our privacy study. The program's final version is using only list of visited sites, location, browser and plug-ins to determine user's identity.

4.5 Summary

Our project uses JavaScript for gathering list of user's visited sites, location, browser and plug-ins. This information is stored in the database and processed by Perl script. Users are shown what can be inferred about them and asked to provide feedback. Feedback is sent to another Perl script that updates entry in the database and user can see the results of all previous participants. Let us now see user's interaction with the program.

5. System In Action

In order to examine users' browsing habits we created a Web site at whattheyknow.cs.wpi.edu. It looks like Figure 2.

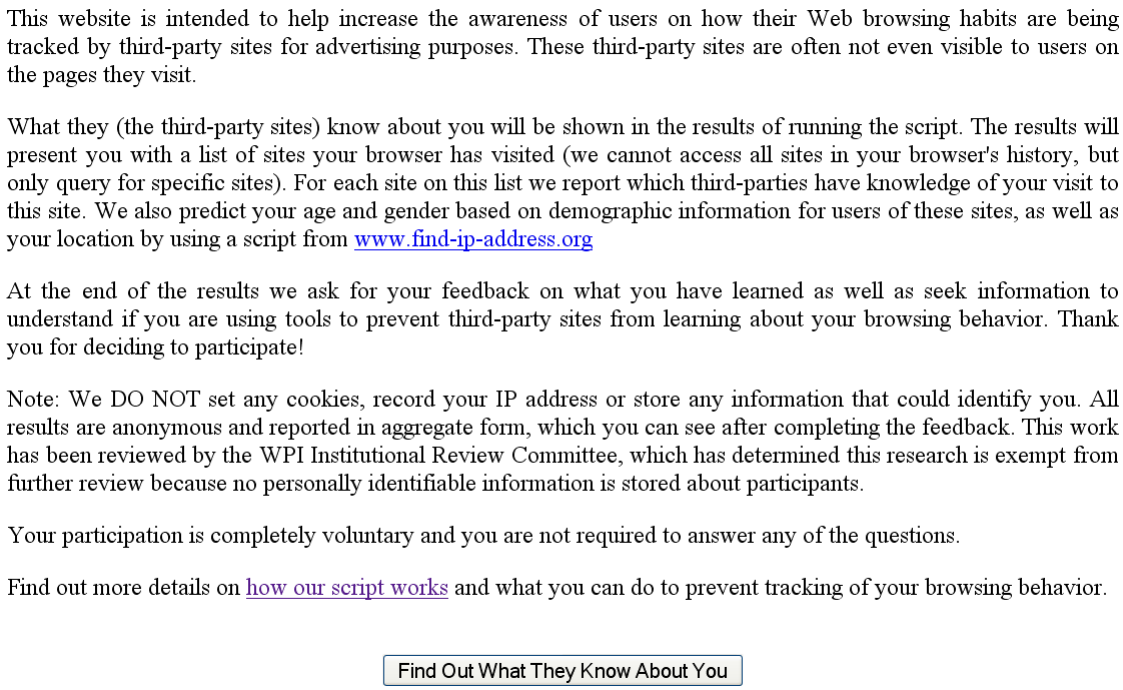

Figure 2 - whattheyknow.cs.wpi.edu Homepage

5.1 Showing What Is Inferred About Users

Users can run the script by pressing "Find Out What They Know" with sample output as shown in Figure 3.

Here is the list of **2 search engines** you have visited along with count and list of third-party tracking sites that know about you (list of site abbreviations).

Visited site	Count	Third-party sites
google.com	1	google
yahoo.com	3	yahoo, yimg, mplex

Here is the list of **1 social network** you have visited along with count and list of third-party tracking sites that know about you (list of site abbreviations).

Visited site	Count	Third-party sites
facebook.com	3	dblclick, atdmt, adv

Here is the list of **3 other popular sites** you have visited along with count and list of third-party tracking sites that know about you (list of site abbreviations).

Visited site	Count	Third-party sites
mail.yahoo.com	1	yimg
maps.google.com	0	
cnn.com	8	dblclick, omniture, scresearch, g-syndication, revsci, aol, dl-rms, vfive

Figure 3 - List of Sites Along With the Third-party Servers

Users get list of visited sites divided into different groups and each site is displayed along with the list of third-parties that have knowledge about it. As shown in Figure 4, we also display demographic data that can be inferred about user and ask for feedback.

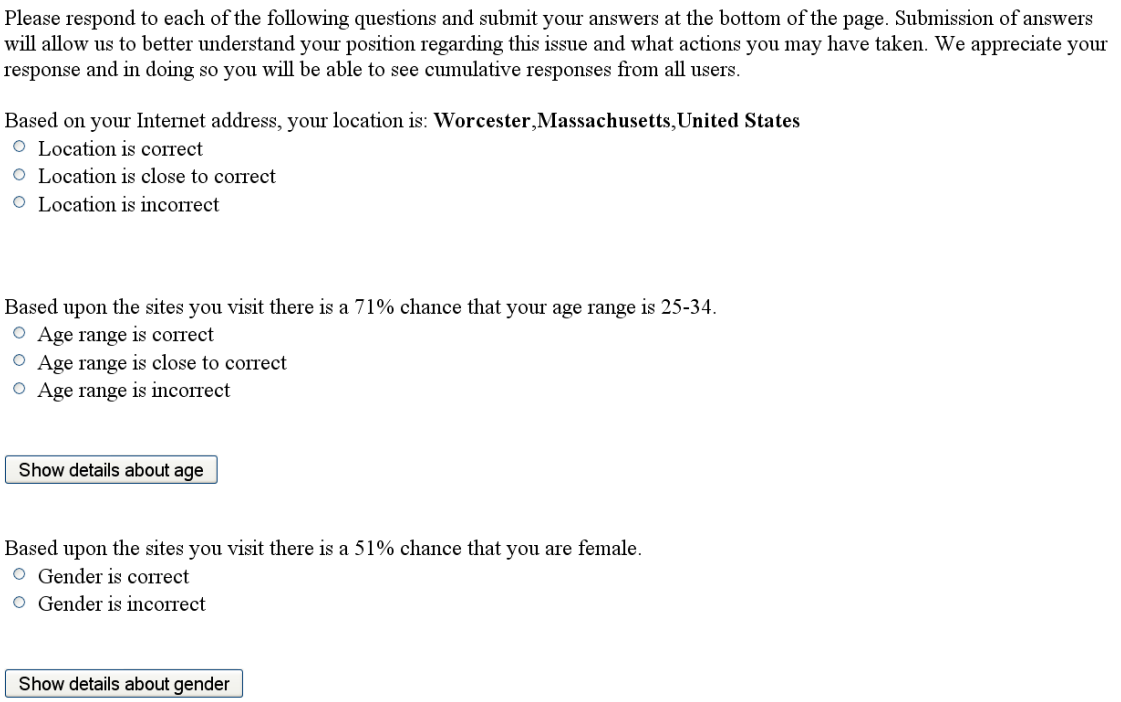

Figure 4 - Our Predictions About Location, Age and Gender

After presenting these data, we want to get feedback from users. As shown in Figure 5, we first ask them about their attitude towards the information that can be inferred about them.

I am concerned that third-party tracking sites have this level of monitoring of my activities.

- ○ Agree
- ○ Not sure
- ○ Disagree

I am concerned that Web sites have this level of information about my location.

- ○ Agree
- ○ Not sure
- ○ Disagree

I am concerned that third-party tracking sites can infer information about demographic information such as age and gender based on the sites I visit.

- ○ Agree
- ○ Not sure
- ○ Disagree

Figure 5 - Questions About User Attitude

Figure 6 shows other questions in the survey that refer to the type of user's location and their browser settings.

Is your current location at:

- ○ Work/School
- ○ Home
- ○ Public

Use of ad blocker tools will prevent some tracking by third-parties identified for your set of visited sites. Do you use any ad blocker tools to prevent the display of advertisements in Web pages?

- ○ Yes
- ○ No
- ○ Don't know

Use of cookie blocking by your browser will prevent some tracking by third-parties identified for your set of visited sites. Do you:

- ○ Allow all cookies (Internet Explorer and Firefox default)
- ○ Allow cookies for only sites I visit (block cookies to third-party sites)
- ○ Allow no cookies (block cookies for all sites)
- ○ Don't know

How often do you delete cookies?

- ○ Often
- ○ Sometimes
- ○ Never
- ○ Don't know

Some third-party sites provide an "opt-out" mechanism to avoid receiving targeted ads. The Network Advertising Initiative (NAI) is a cooperative of such sites. Do you use opt-out cookies for third-party sites?

- ○ Yes
- ○ No
- ○ Don't know

Periodic removal of browser history prevents scripts like this one from detecting what sites you visit, but will not prevent third-party sites from tracking your behavior (they typically use cookies). What browser history settings do you use:

- ○ Use browser default for managing history
- ○ Clear history when browser closes
- ○ Set browser to not remember history
- ○ Don't know

Newer versions of browsers have features to create a "private" browsing session where history is not recorded. These include InPrivate Browsing for Internet Explorer, incognito mode for Chrome and Firefox Private Browsing. Do you use any of these features?

- ○ Yes
- ○ No
- ○ Don't know

Any additional comments about what you learned.

Submit Responses and See Cumulative Results

Figure 6 - Questions About User type of Location and Browser Settings

Answering any question or writing a comment is completely optional.

5.2 Results of All Users

After submitting responses user can see cumulative data from all previous users.

Here is the list of 100 most visited web sites along with percentage of 3749 whattheyknow users who visited them:

Visited site	Percent
google.com	59 %
facebook.com	51 %
youtube.com	42 %
maps.google.com	31 %
cnn.com	24 %
weather.com	24 %
yahoo.com	24 %
ebay.com	23 %
twitter.com	20 %
apple.com	19 %

Number of visited sites for each whattheyknow user:

Max: 372
Median: 8
Mean: 15

Here is the list of 10 most visited third-party sites along with average number of visited sites per user:

Third party	Average # of sites
doubleclick.net(google.com)	7.8
atdmt.com(microsoft.com)	3.9
google-analytics.com(google.com)	3.9
omniture.com	3.8
quantserve.com	3.4
scorecardresearch.com	2.6
advertising.com(aol.com)	2.6
yieldmanager.com(yahoo.com)	2.1
revsci.net	1.7
yimg.com(yahoo.com)	1.6

Figure 7 - Visited Sites and Third Parties' Statistics

In Figure 7 we show 10 most visited sites and 10 third-parties that know the most about users on average. All previous responses to the survey questions are also presented, as we can see in Figures 8 and 9.

Based on your Internet address, your location is:

1797 answers with percentages as follows:

Location is correct	64 %
Location is close to correct	23 %
Location is incorrect	13 %

Based upon the sites you visit is your age range correct?

1804 answers with percentages as follows:

Age range is correct	19 %
Age range is close to correct	23 %
Age range is incorrect	58 %

Based upon the sites you visit is your gender correct?

1793 answers with percentages as follows:

Gender is correct	64 %
Gender is incorrect	36 %

I am concerned that third-party tracking sites have this level of monitoring of my activities.

1771 answers with percentages as follows:

Agree	63 %
Not sure	25 %
Disagree	12 %

I am concerned that Web sites have this level of information about my location.

1778 answers with percentages as follows:

Agree	48 %
Not sure	27 %
Disagree	26 %

I am concerned that third-party tracking sites can infer information about demographic information such as age and gender based on the sites I visit.

1768 answers with percentages as follows:

Agree	54 %
Not sure	25 %
Disagree	21 %

Figure 8 - Results for Questions Pertaining to User Attitude and Correctness of Our Prediction

Is your current location at:

1746 answers with percentages as follows:

Work/School	56 %
Home	40 %
Public	4 %

Use of ad blocker tools will prevent some tracking by third-parties identified for your set of visited sites. Do you use any ad blocker tools to prevent the display of advertisements in Web pages?

1769 answers with percentages as follows:

Yes	56 %
No	33 %
Don't know	11 %

Use of cookie blocking by your browser will prevent some tracking by third-parties identified for your set of visited sites. Do you:

1765 answers with percentages as follows:

Allow all cookies (Internet Explorer and Firefox default)	37 %
Allow cookies for only sites I visit (block cookies to third-party sites)	43 %
Allow no cookies (block cookies for all sites)	3 %
Don't know	17 %

How often do you delete cookies?

1773 answers with percentages as follows:

Often	21 %
Sometimes	52 %
Never	21 %
Don't know	6 %

Some third-party sites provide an "opt-out" mechanism to avoid receiving targeted ads. The Network Advertising Initiative (NAI) is a cooperative of such sites. Do you use opt-out cookies for third-party sites?

1750 answers with percentages as follows:

Yes	16 %
No	55 %
Don't know	29 %

Periodic removal of browser history prevents scripts like this one from detecting what sites you visit, but will not prevent third-party sites from tracking your behavior (they typically use cookies). What browser history settings do you use:

1745 answers with percentages as follows:

Use browser default for managing history	68 %
Clear history when browser closes	15 %
Set browser to not remember history	4 %
Don't know	12 %

Newer versions of browsers have features to create a "private" browsing session where history is not recorded. These include InPrivate Browsing for Internet Explorer, incognito mode for Chrome and Firefox Private Browsing. Do you use any of these features?

1753 answers with percentages as follows:

Yes	33 %
No	57 %
Don't know	10 %

Figure 9 - Results for Questions Pertaining to User Type of Location and Browser Settings

6. Results

6.1 Users Statistics

We had 3749 users participating in our project as of Feb ‘10. We received feedback from 1853 (49%) of them. Further results are only for these users, unless noted otherwise.

Table 1 - Users' Location

Location	Percent
United States	87
Massachusetts	26
Worcester	10

As we can see in Table 1, 87% of users are from the United States, 26% are from Massachusetts and 10% are from Worcester. For each of the categories we included users who had their location correct or close to correct. Worcester users are mostly students and employees at WPI who were among the first to test our program.

Table 2 – Users’ Gender

Gender	Overall prediction percent	Actual percent
Male	52	72
Female	48	28

As we can see in Table 2, our users are mostly male. When chances for being male and female are equal, we show female (with 50% chance) and this is the reason for large difference between predicted and actual results. Equal chances occurred 15% of the time.

Table 3 - Users' age groups

Age group	Overall prediction percent	Actual Percent
3-12	15	1
13-17	4	1
18-20	0	0
21-24	0	0
25-34	42	56
35-44	39	42
45-49	0	0
50-54	0	0
55-64	0	0
65+	0	0

Our prediction of age favors groups that are larger than others. As a consequence almost all of our calculations predict either age groups of 25-34 or 35-44. Reason for high percentage for 3-12 is that it is shown (with only 10% chance) when chances for all groups are equal.

6.2 Prediction Feedback

As we can see from Figure 8, location we obtained was correct in 64% of the time. It was close in 23% and incorrect in only 13%. This means that 87% of the time we were really close to their actual location.

As shown in Figure 8, age range prediction was correct in only 19% of the cases. The low result for this could be caused by the formula we used, but it could also be due to the data we have about sites' demographics that might not be precise.

In Figure 8 we can also see that 64% of the time our gender prediction was right. Let us also consider how precise the prediction is depending on our certainty level.

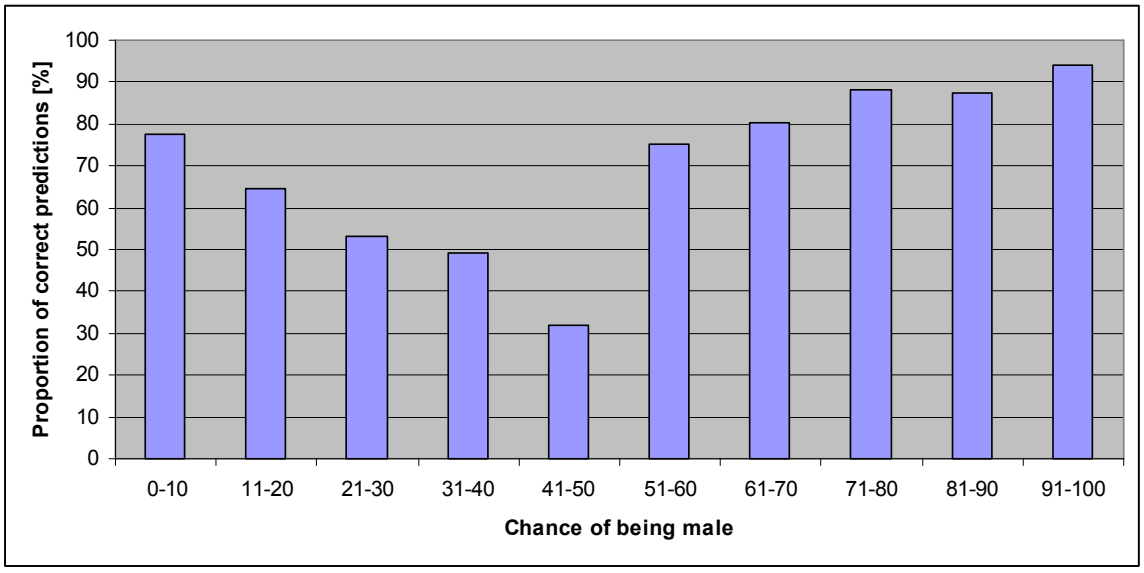

Figure 10 - Correct Predictions for Different Chances of Being Male

As we can see in Figure 10, when the chance with which we predict is close to 0 or 100% predicting female or male respectively, our chances our a lot higher. When we are 90% certain user is a female, we are right 78% of the time and when we are 90% certain a user is male, we are right 94% of the time.

6.3 Visited Sites Analysis

For comparing profile of our participants to the random sample of population we use Quantcast data.

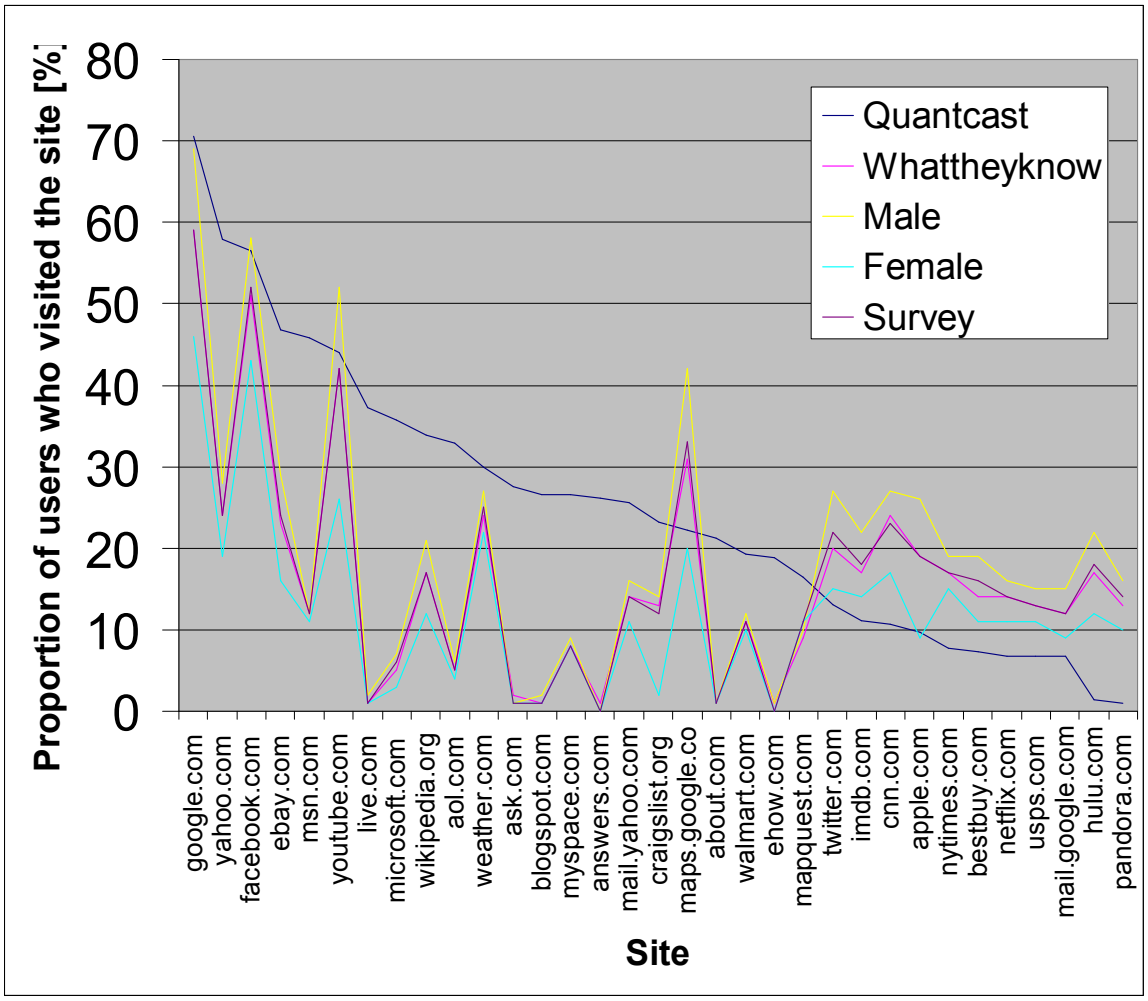

Figure 11 - Visited Sites for Different Groups of Users

First 20 sites on Figure 11 represent top-20 Quantcast Web sites that we checked for. As we can see, most of them were found in users' histories less often than we expected. One of the reasons for this could be that we are only checking for homepages. If a user visits a sub-domain or any other page on that site, we cannot detect it. This type of visit can happen when they use search engines. Some of the sites with great differences compared to Quantcast, such as ask.com, answers.com, about.com and ehow.com often come up when searching for answers to various questions.

On the right side of the graph in Figure 11, we present sites that show up in our top-20 but are ranked lower in Quantcast. The reason for detecting certain site in the history more often than we expect is due to the fact that our sample is not random and has stronger preference for some of them.

Visited sites do not have large differences among different groups. Males, females and users who completed the survey have similar preference when it comes to top-20 Quantcast sites. Our top-20 sites seem to have more male visitors. One of the reasons for this could be that males are dominant in the sample and therefore they have more influence over what is in the top-20.

We tried comparing data between different age groups. The only groups we have enough data for are 25-34 and 35-44.

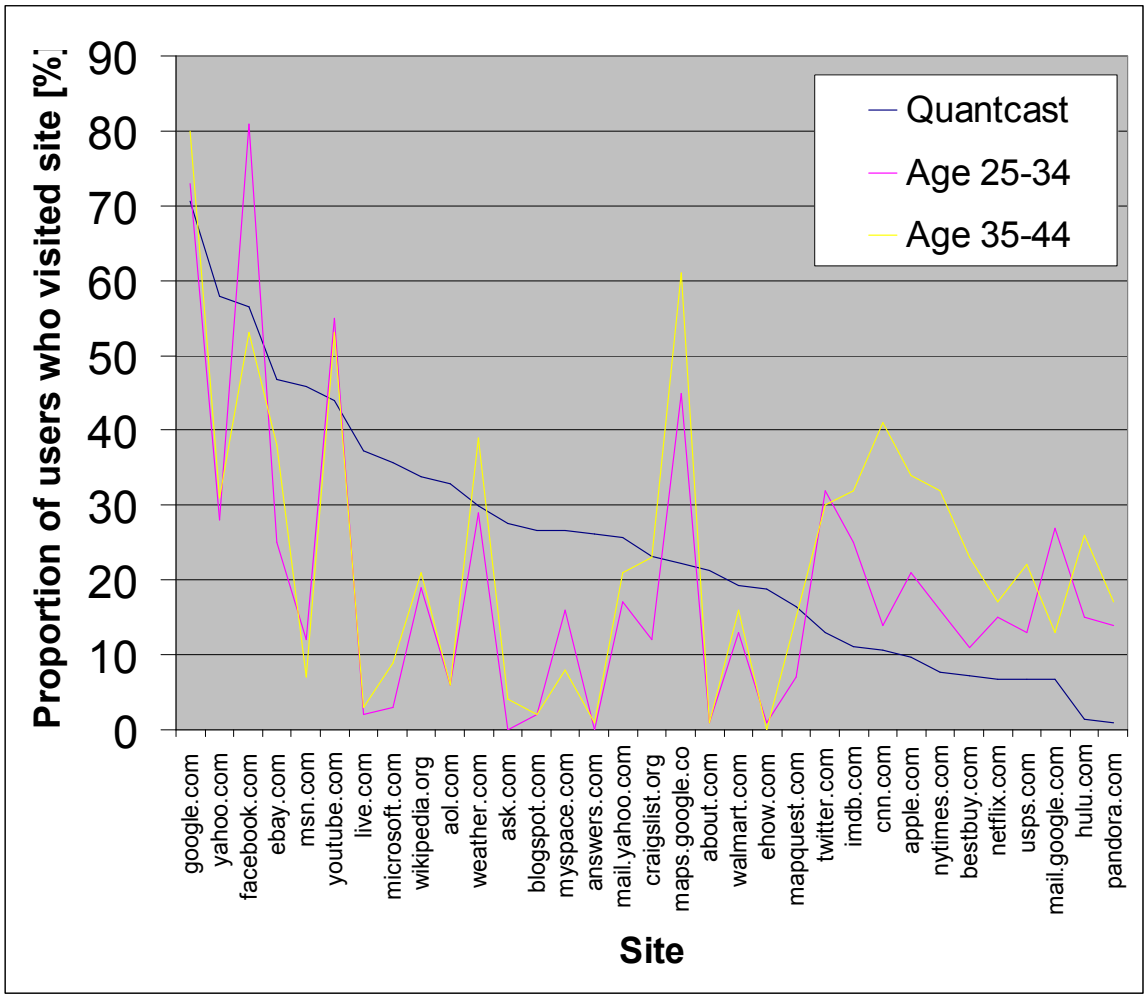

Figure 12 - Visited sites of different age groups

As shown in Figure 12, there is no significant difference between these two populations.

Table 4 - Users who visited sites in different categories

Category	Percent
Search engines	66
Social networks	56
Adult entertainment	1

In Table 4 we show percentage of users who visited sites in different categories.

6.4 Third-parties analysis

Table 5 – Top-10 Third-parties Comparison to the List Obtained in March '10 Using a Similar Methodology as Described in [1]

Rank	List obtained in March '10 using a similar methodology as described in [1]	Our research
1	doubleclick.net	doubleclick.net
2	google-analytics.com	atdmt.com
3	2mdn.net	google-analytics.com
4	quantserve.com	omniture.com
5	scorecardresearch.com	quantserve.com
6	atdmt.com	scorecardresearch.com
7	omniture.com	advertising.com
8	googlesyndication.com	yieldmanager.com
9	yieldmanager.com	revsci.net
10	2o7.net	yimg.com

We can compare the list of top-10 third-parties we have to the list obtained in March '10 using a similar methodology as described in [1]. As shown in Table 5, lists are similar. Both lists contain 7 of them: doubleclick.net, google-analytics.com, quantserve.com, scorecardresearch.com, atdmt.com, omniture.com and yieldmanager.com. Top ranked in both is doubleclick.net. Besides atdmt.com and omniture.com that are ranked higher in our list, there are no other relative changes among third-parties contained in both lists.

6.5 Analysis of Users' Attitude towards Online Privacy

From Figure 8 we see that 63% of users are concerned about the level of monitoring third-party sites have, 48% of them are concerned that Web sites have

knowledge of their location and 54% is worried that their age and gender can be predicted based on the sites they visit.

6.6 Users' Type of Location and Browser Settings

From Figure 9 we see that 56% of our users accessed from work, 40% accessed from home and only 4% accessed from public locations. 56% of them use ad blockers. 27% never deletes their cookies and 46% blocks either third-party cookies or all. Only 16% of our users use NAI opt-out mechanism to avoid targeted ads. 19% set their browsers to clear history after closing or not to remember any history at all. Private browsing is used by 33% of our users.

6.7 Comparison of Users' Responses between Male and Female Users

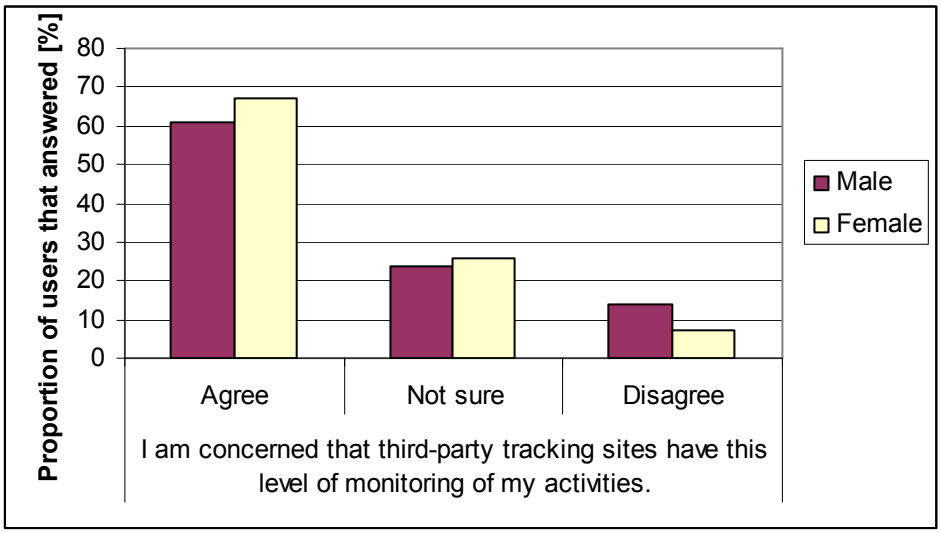

Figure 13 - Male and Female Users Attitude Towards Third-party Monitoring

As shown in Figure 13, less male users are concerned about third-parties tracking. By calculating 95% confidence intervals we can conclude that the difference of male and female users is 6% ± 5%.

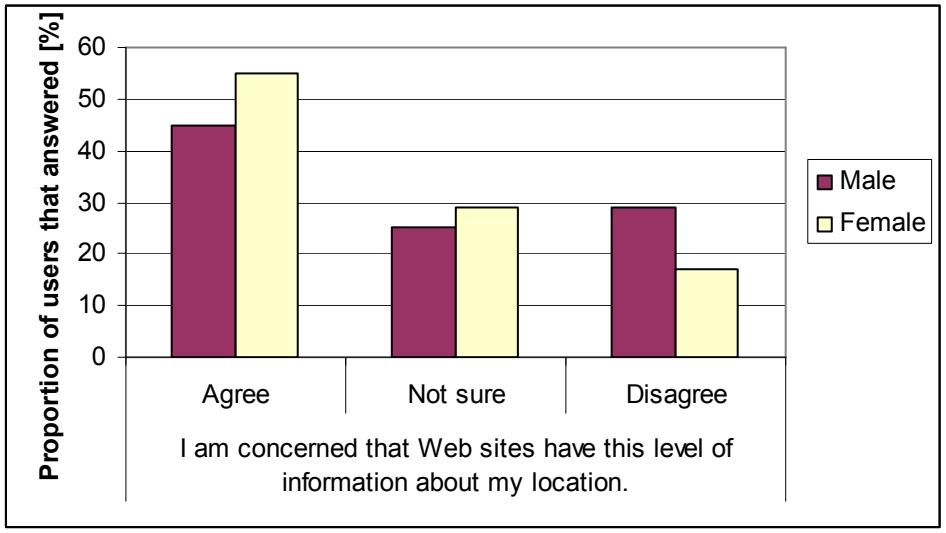

Figure 14 - Male and Female Users Attitude Towards Location Information

As we can see in Figure 14, more female users are concerned that third-parties have information about their location. By calculating 95% confidence intervals we can conclude that the difference of male and female users who are concerned about this issue is 10% ± 5%.

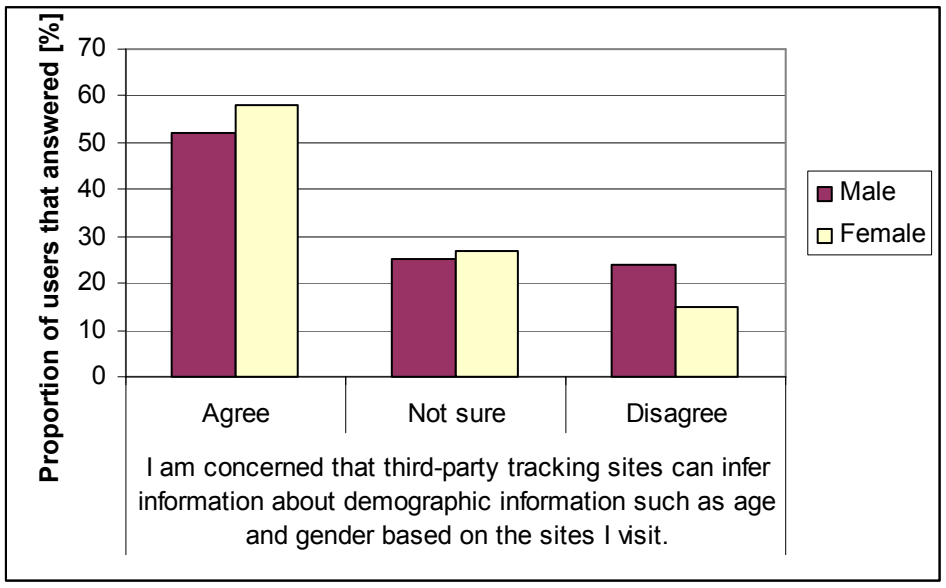

Figure 15 - Male and Female Users Attitude Towards Inferring Their Age and Gender

In Figure 15 we can see that female users are more concerned that third-party sites can infer information about demographics. By calculating 95% confidence intervals we can conclude that the difference of male and female users is 6% ± 5%.

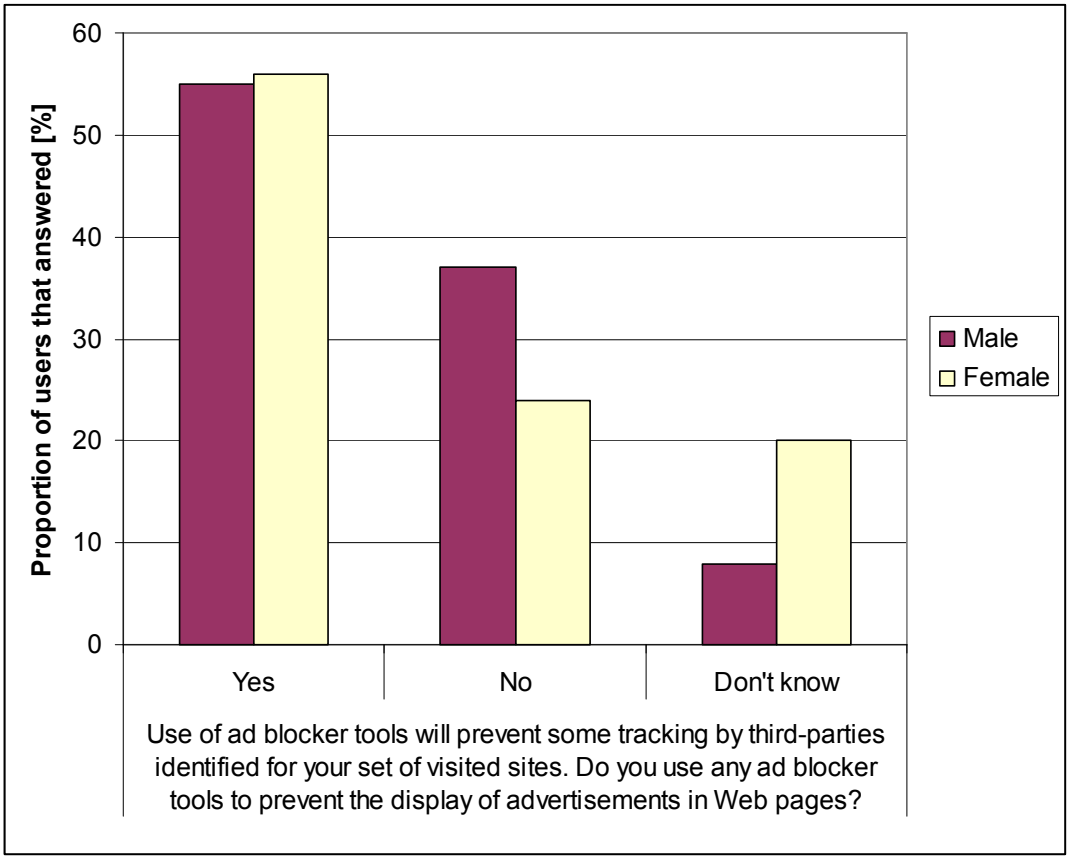

Figure 16 - Male and Female Users Usage of Ad Blockers

In Figure 16 we can see that female users use ad blocker tools more. By calculating 95% confidence intervals we get that the difference is 1% ± 5%. This calculation indicates that the difference is not significant.

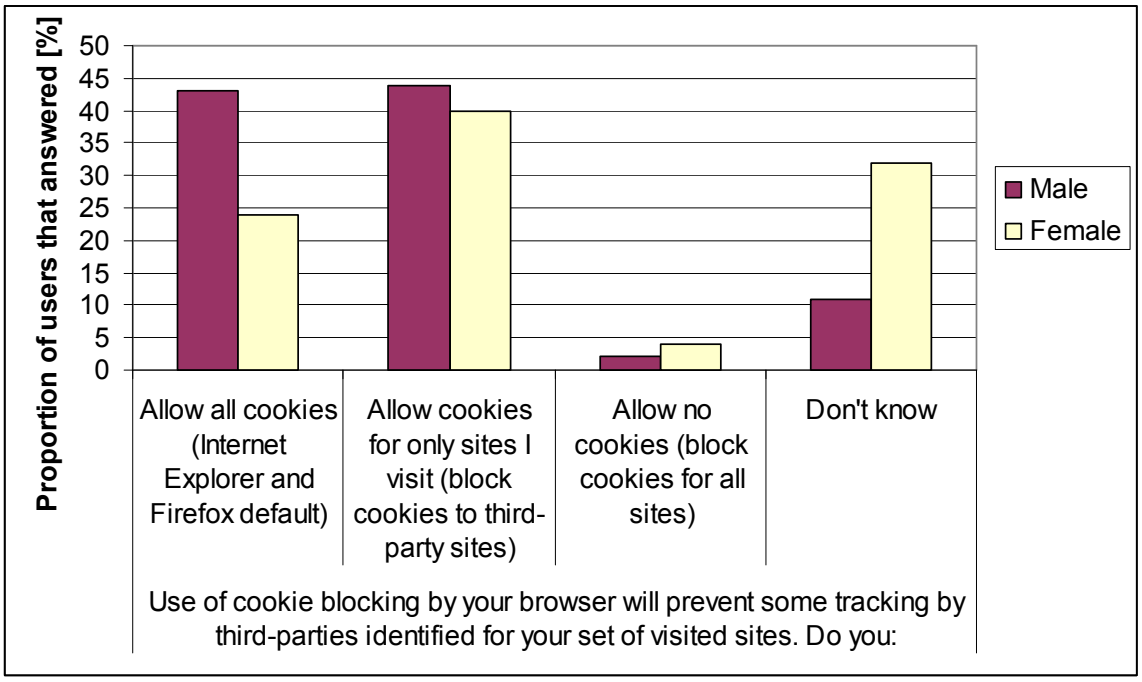

Figure 17 - Male and Female Users Cookies Settings

Figure 17 shows cookies settings of male and female users. If we sum answers for not knowing the browser settings and allowing all cookies we get non-significant difference of 2% more female users who do not block any.

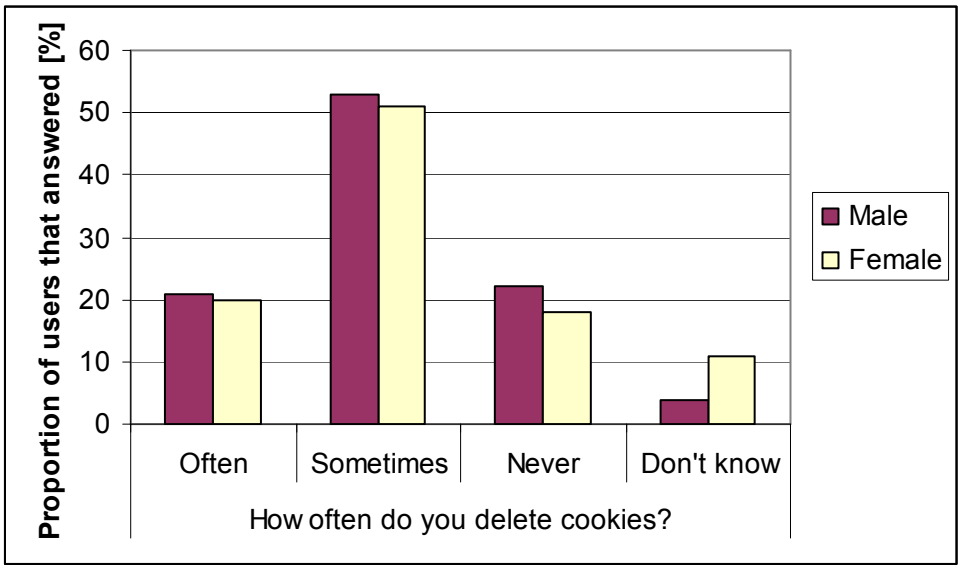

Figure 18 - Male and Female Users Deletion of Cookies

As shown in Figure 18, male and female users have similar habits of deleting cookies. Summing users who have no knowledge about deleting cookies and the ones who never delete them, we get non-significant difference of 3% more female users who do not take any actions concerning cookies.

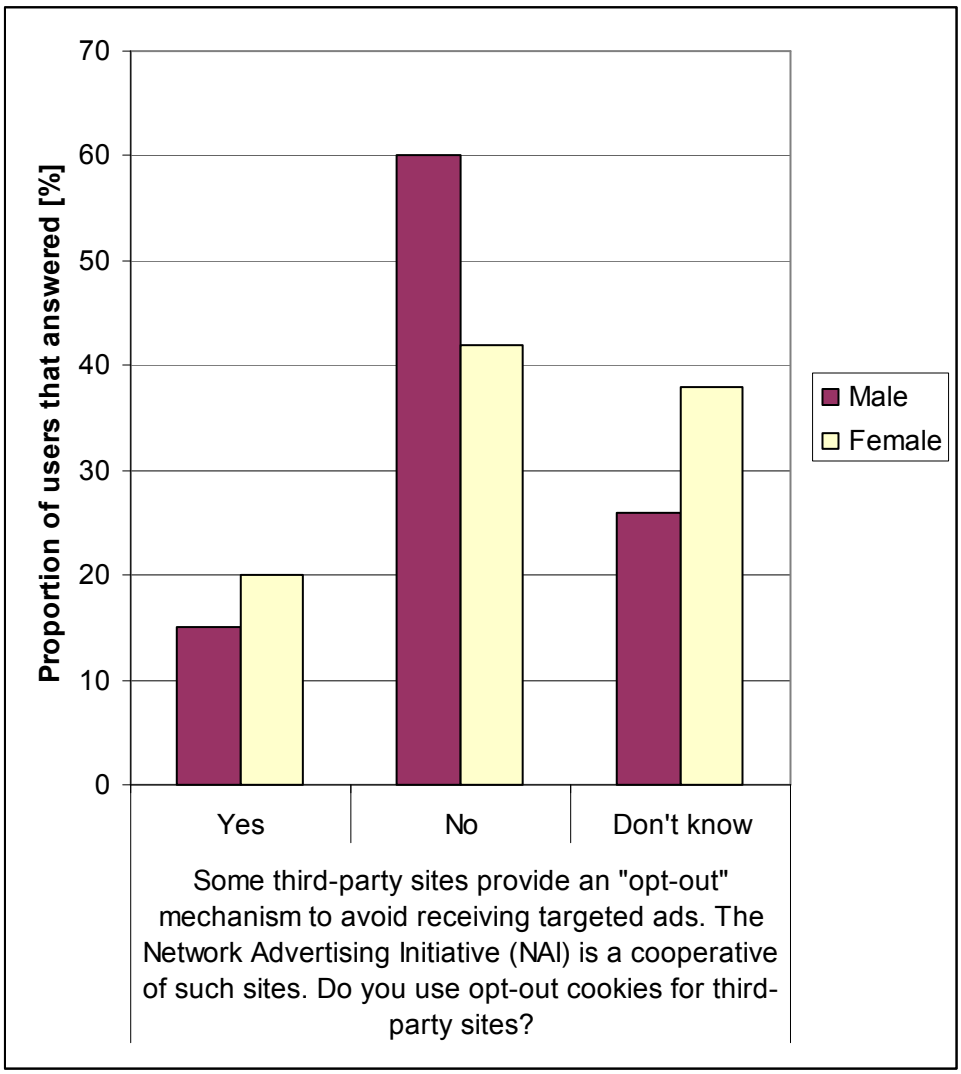

Figure 19 - Male and Female Users Usage of NAI opt-out

In Figure 19 we can see that more female users use NAI opt-out cookies. By calculating 95% confidence intervals, we can conclude that the difference of male and female users who use this mechanism to avoid receiving targeted ads is 5% ± 4%.

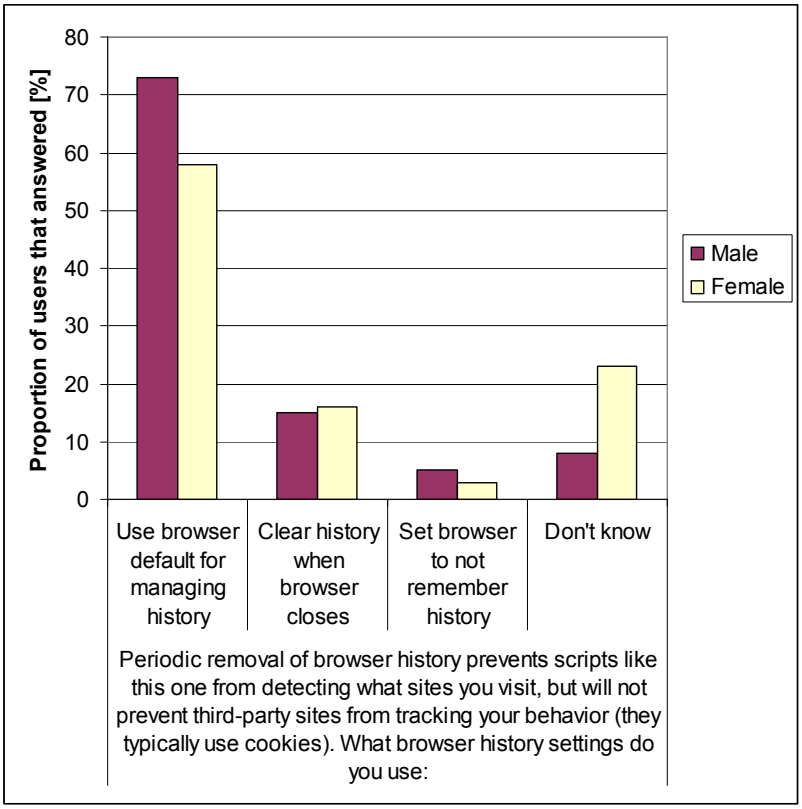

Figure 20 - Male and Female Users Browser History Settings

If we sum users who user browser default for managing history and the ones who do not know from Figure 20, we get the same percent. This means that there is no significant difference between browser history removal settings for male and female users.

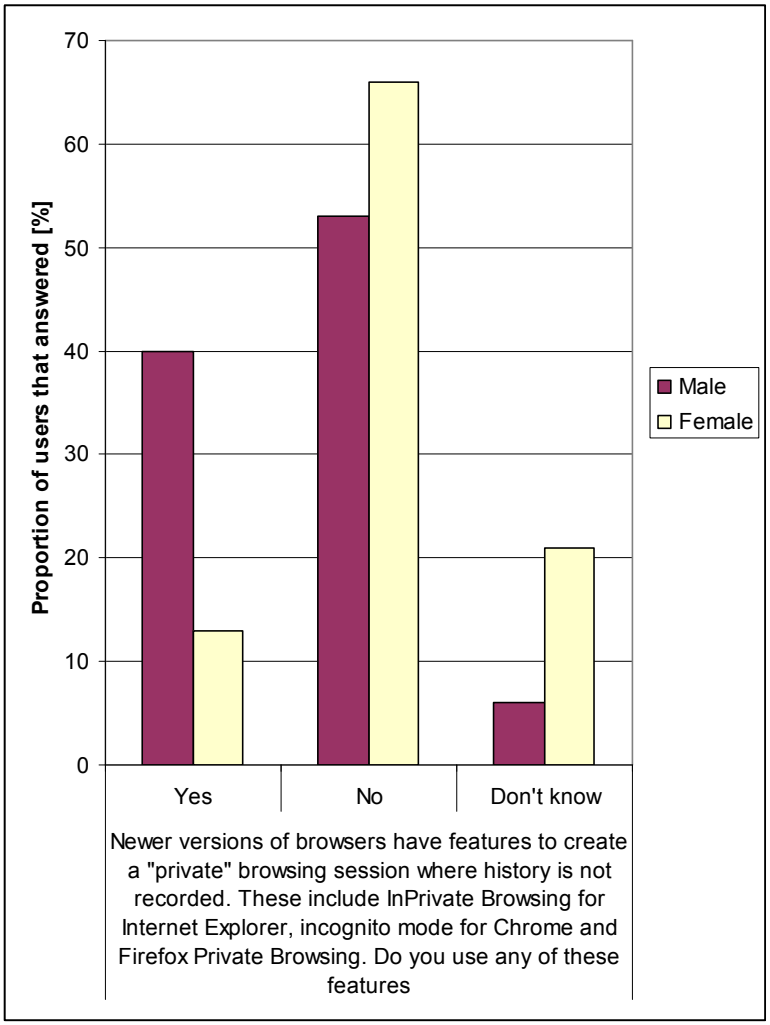

Figure 21 - Male and Female Users Usage of Private Browsing

In Figure 21 we can see that more male users use private browsing option. By calculating 95% confidence intervals we get that the difference of male and female users who use this option is 27% ± 4%.

6.8 Comparison of Users' Responses between Different Age Groups

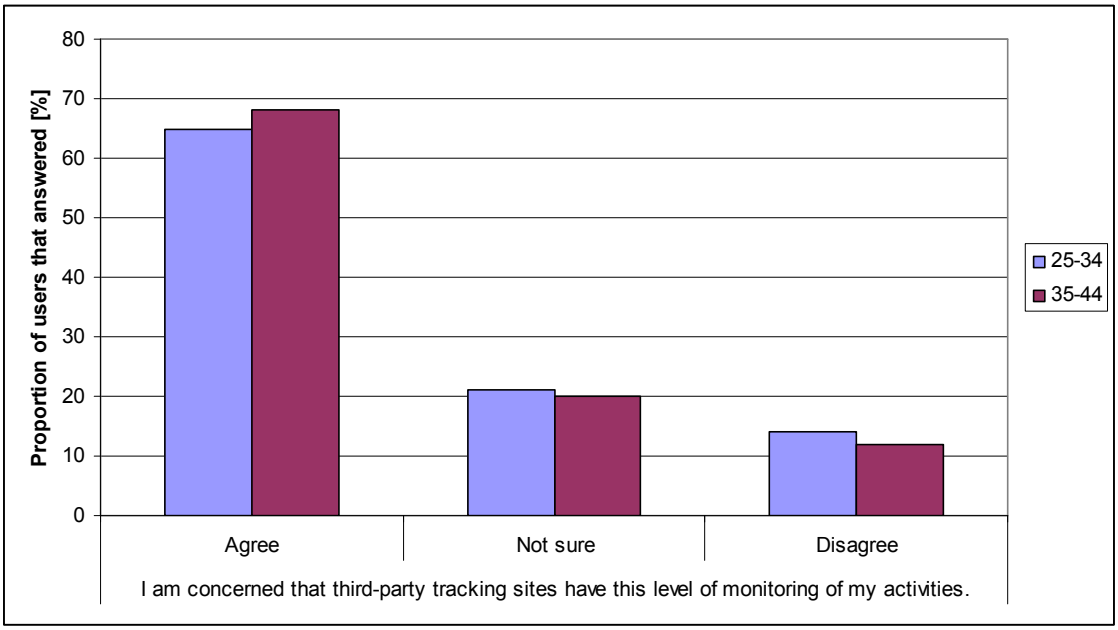

Figure 22 - Users of age 25-34 and 35-44 Attitude Towards Third-party Monitoring

As shown in Figure 22, there is no significant difference between users in age groups 25-34 and 35-44.

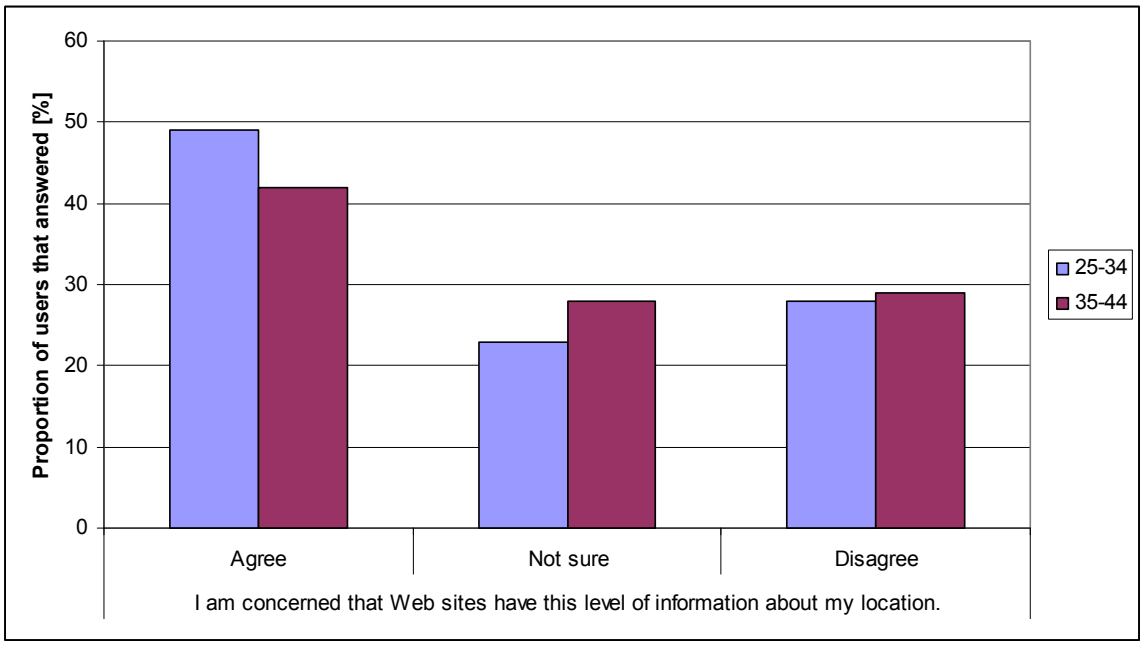

Figure 23 - Users of age 25-34 and 35-44 Attitude Towards Location Information

As we can see in Figure 23, more users of age 25-34 are concerned that their location can be inferred. By calculating 95% confidence intervals we get that the difference users age 25-34 and 35-44 is 6.5% ± 10.5%. Therefore, the difference is not significant.

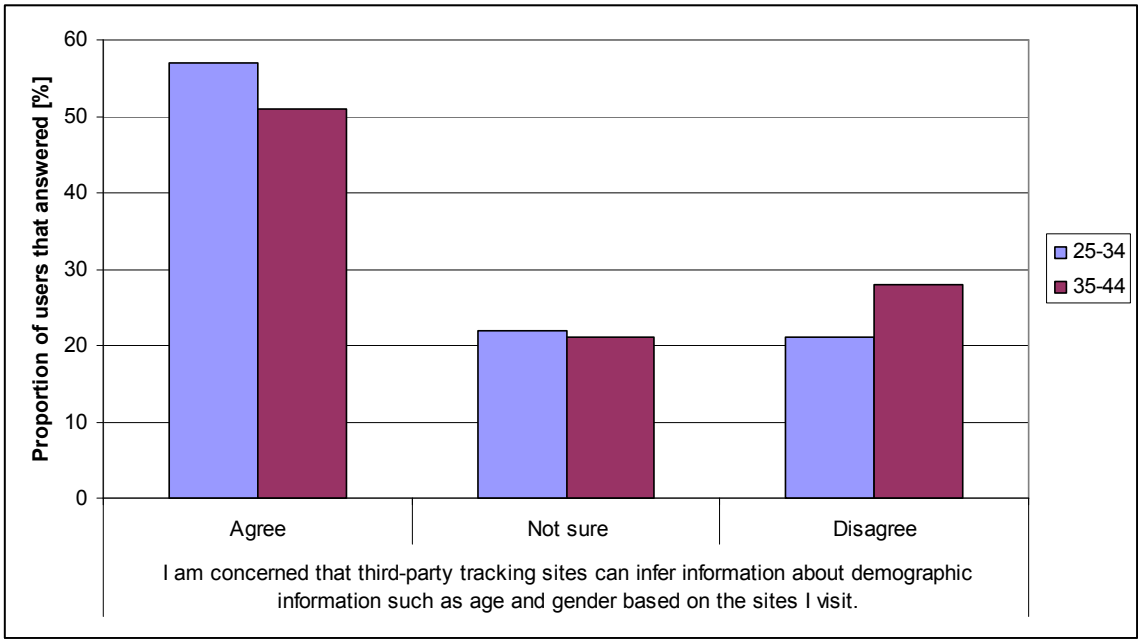

Figure 24 - Users of age 25-34 and 35-44 Attitude Towards Inferring Their Age and Gender

In Figure 24 we see that users of age 25-34 are more concerned that their demographic information can be inferred. By calculating 95% confidence intervals we get that the difference of users age 25-34 and 34-45 is 6% ± 11% . Therefore, the difference is not significant.

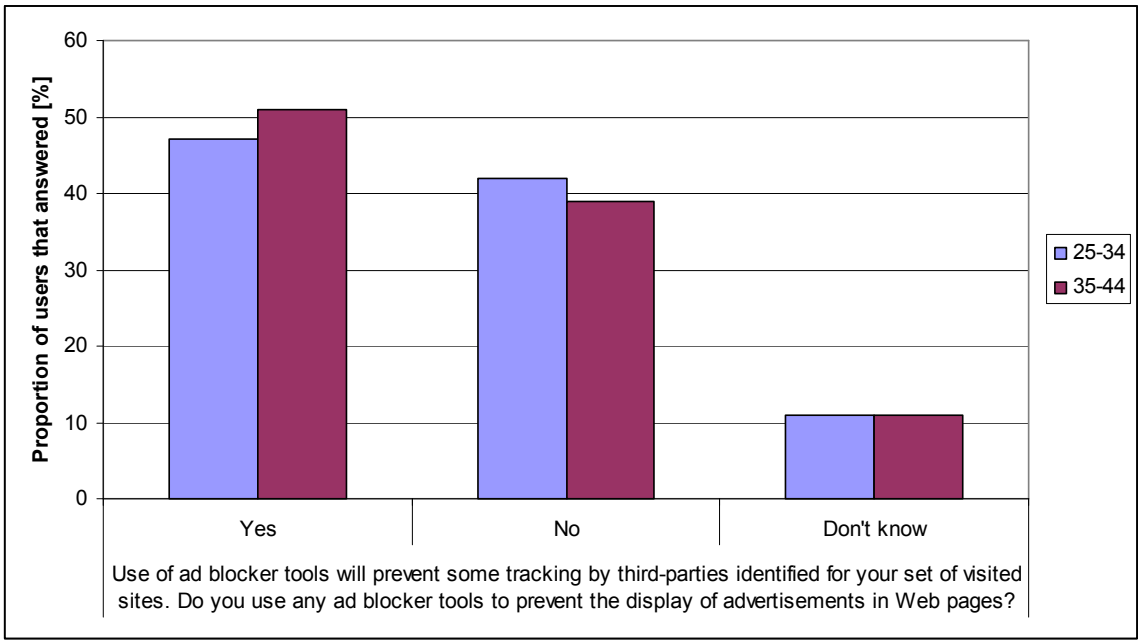

Figure 25 - Users of age 25-34 and 35-44 Usage of Ad Blockers

In Figure 25 we see that users of age 25-34 use ad blockers more, but the difference between age groups is not significant.

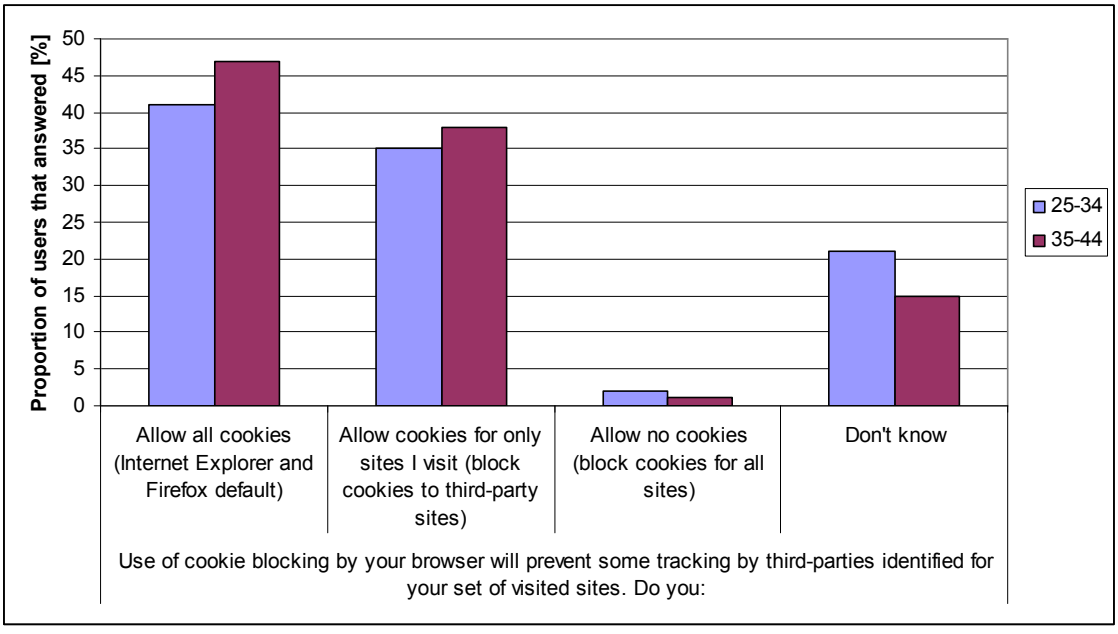

Figure 26 - Users of age 25-34 and 35-44 Cookies Settings

Figure 26 shows cookies settings of users in different age groups. If we sum answers for not knowing the browser settings and allowing all cookies we get the same percent for both age groups.

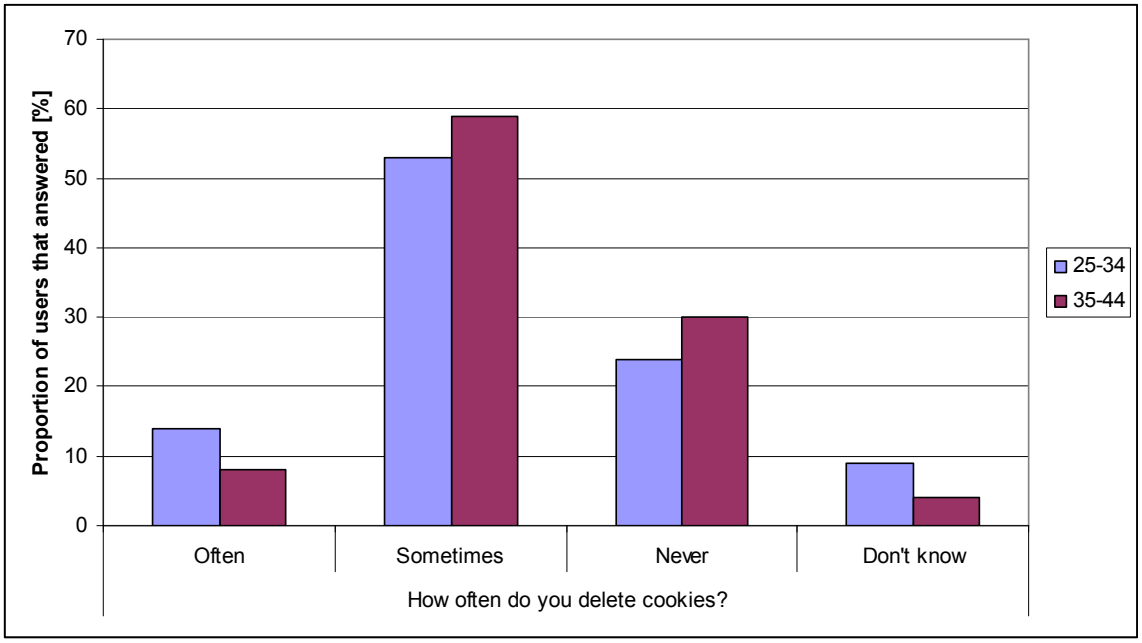

Figure 27 - Users of age 25-34 and 35-44 Deletion of Cookies

As shown in Figure 18, users of both age groups have similar habits of deleting cookies. Summing users who have no knowledge about deleting cookies and the ones who never delete them, we get non-significant difference of 1% more users age 35-44 who do not take any actions concerning cookies. Larger percent of users of age 25-34 appears to delete cookies often. By calculating 95% confidence intervals we get that the difference of age groups is 5.5% ± 6.5%. This calculation indicates that the difference is not significant.

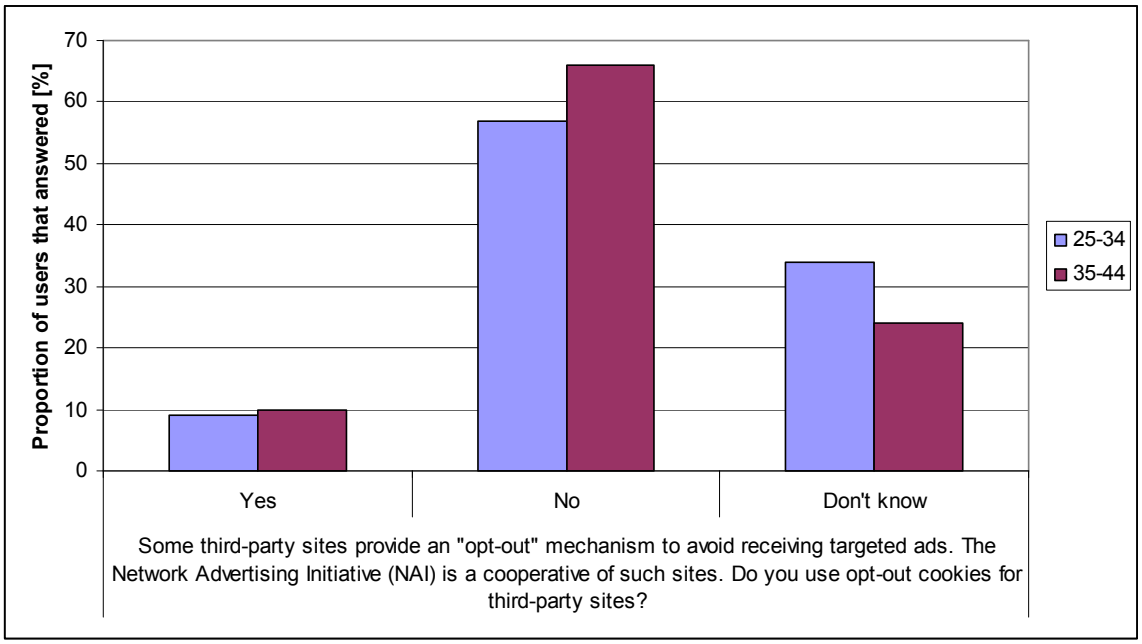

Figure 28 - Users of age 25-34 and 35-44 Usage of NAI opt-out

In Figure 28 we see that both age groups usage of NAI opt-out cookies is similar.

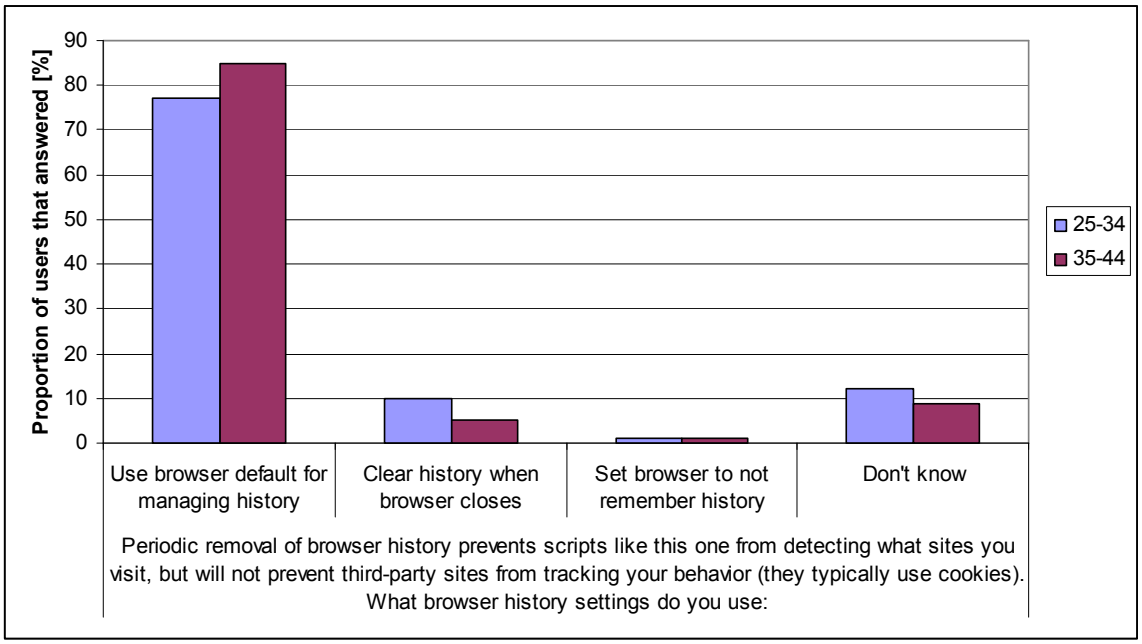

Figure 29 - Users of age 25-34 and 35-44 Browser History Settings

If we sum users who user browser default for managing history and the ones who do not know from Figure 29, we get difference of 5%. By calculating 95% confidence intervals we get that this difference is 4.5% ± 6.5%. This calculation indicates that there is no significant difference between browser history removal settings for age groups 25-34 and 35-44.

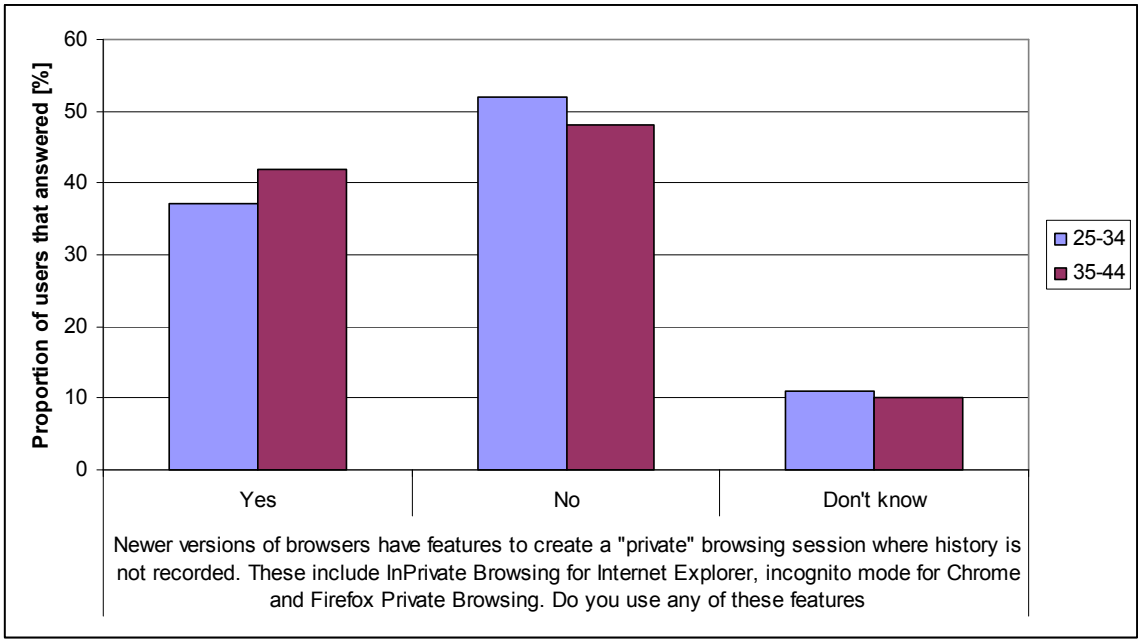

Figure 30 - Users of age 25-34 and 35-44 Usage of Private Browsing

In Figure 30 we see that more users in age group 35-44 use private browsing option. By calculating 95% confidence intervals we get that the difference is 7% ± 8%. This calculation indicates that the difference is not significant.

6.9 Analysis of Location Correctness

When we analyze location correctness for home, work and public users we get the graph in Figure 31.

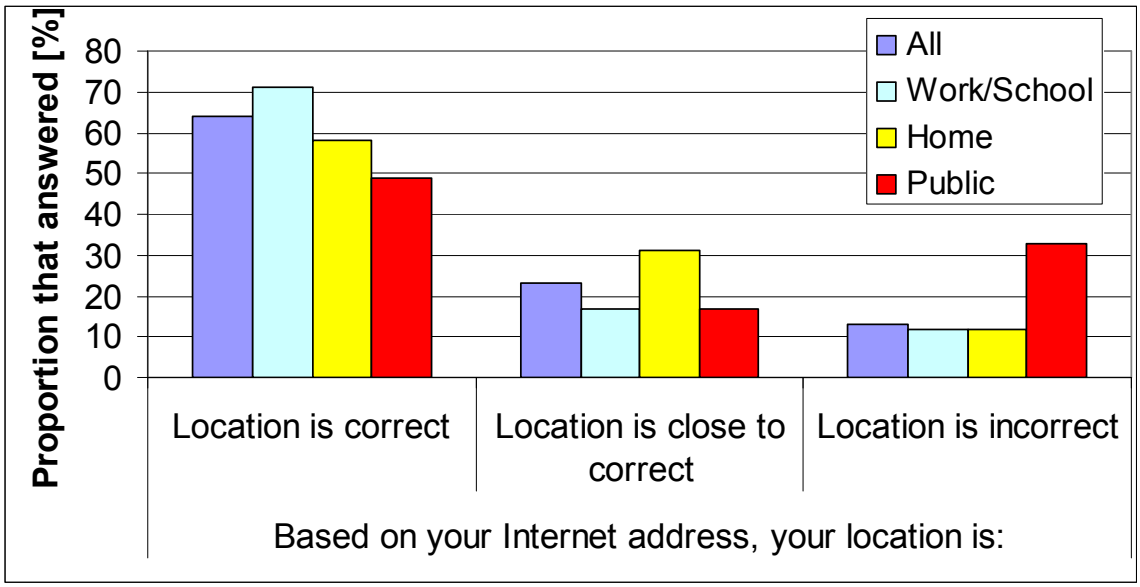

Figure 31 - Location Correctness for Different Access Points

We managed to completely miss location for all users in 13% of the time. This number seems to be a lot larger for users who accessed our program from public computers. By calculating 95% confidence intervals we can conclude that the difference of incorrect location of users who use public computers and the others is 16% ± 12%. This means that location of public computers is harder to track.

In Figure 32 we see the comparison of users whose predicted location was in the United States and the ones whose location was outside of it.

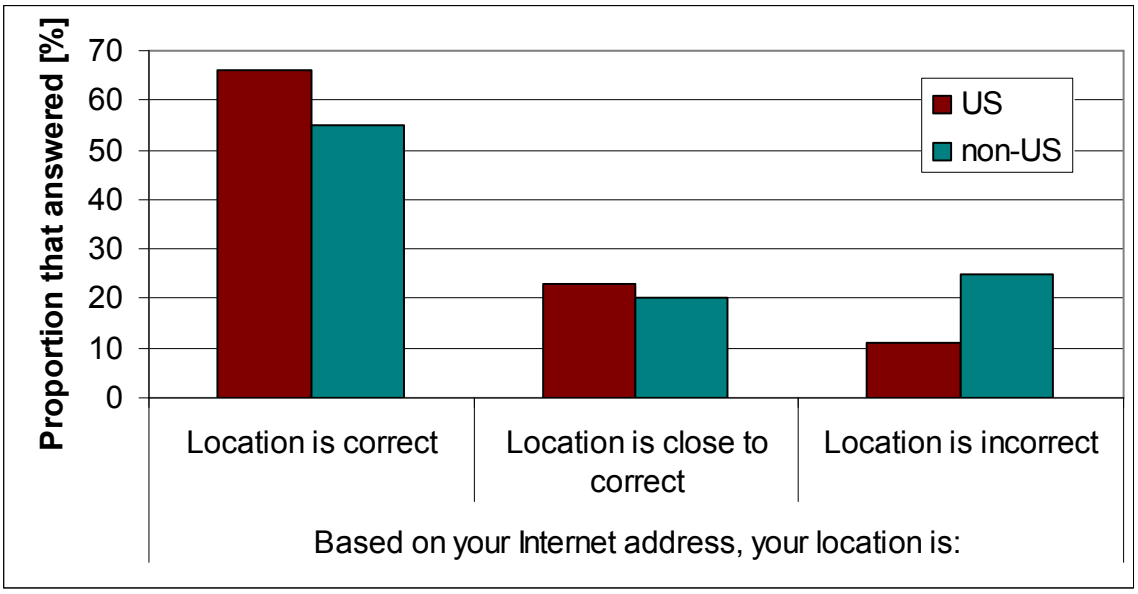

Figure 32 - Location Prediction Comparison of People from United States and Users Outside of it

If we calculate 95% confidence intervals we get that the difference between these two groups of users is 11% ± 4%. Site that we are using seems to have a lot better ways of tracking users within the United States. We tried to do the same analysis for Worcester, Massachusetts and United States and results are shown in Figure 33.

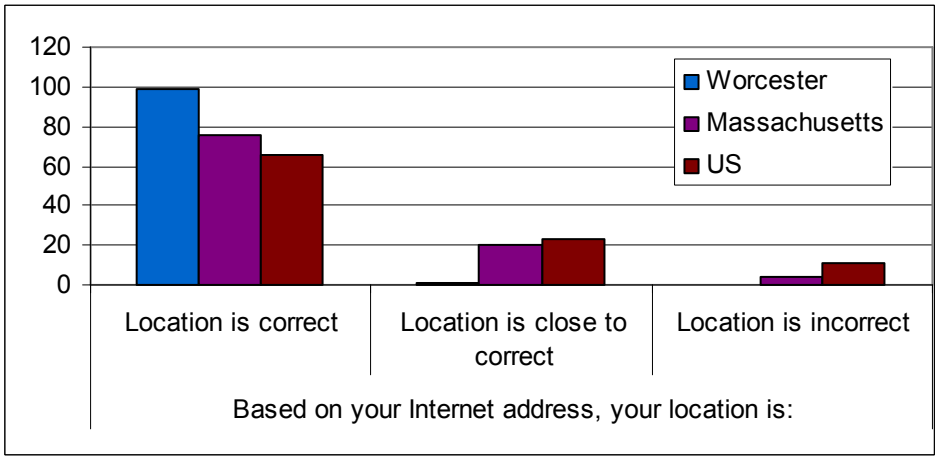

Figure 33 - Location Correctness for Different Regions

We expect that most of the Worcester users are either students or employees at WPI accessing from campus network or really close to it. As a consequence we get 99 percent accuracy. There is no significant difference between predictions for United States and Massachusetts.

6.10 Browser Fingerprinting

Based on the data we have, we can determine how unique browser configurations of users are. As a browser fingerprint we take its type and list of all plug-ins installed. We go through list of users and check if their browser is unique compared to the others. In order to have more accurate data, we want to check only distinct users. There is a small chance that some of our users have multiple entries in the database if they ran the program with different sites in browser history. We expect that users from different locations have a strong chance of being distinct and use them for comparison. Using this approach we get 1057 distinct users and 85% of them have unique browser configurations.

6.11 Summary

We had 3749 users participating in our research. 49% of them gave us feedbacks about their browsing habits. 87% of them were from the United States and 72% were male. Our age range prediction was correct 19% of the time. Gender prediction was correct 64% of the time, but it was more accurate for stronger predictions. We compared our male and female users and found significant difference in behavior. Based on the feedback we received about age, we only had enough data to compare age groups 25-34 and 35-44. We did not find significant difference between behaviors of these two groups. We also compared users from different types of access points and concluded that the

location of users accessing from public location and outside of United States is harder to track. We compared the list of our top-10 third-parties to the list obtained in March '10 using a similar methodology as described in [1] and concluded that the results were close. We also did similar comparison to [8] and determined that 85% of users can be uniquely identified using only their browser type and list of plug-ins.

7. Conclusions

7.1 Future Work

We asked users questions about their attitudes and measures of protection from online tracking. Future work is to better understand users' point of view about online privacy. We plan to perform a more thorough examination that will enable us to do it.

7.2 Summary

From the results we obtained, we can conclude that online users do not take the privacy issues seriously. Based on the feedback received, a lot of users expressed concern about the information that can be inferred about them. However, some of their actions show different. About 80% of our users never delete history, 60% of them do not block any cookies and only 21% delete them often. Only 33% of our users protect their privacy by using private browsing option that some of the modern browsers have. Most browsers keep history and cookies for an extensive period of time by default. Third-parties take an advantage of these settings and gather information about users. Even though users have multiple ways of protecting themselves, they still decide not to do it. Some of them make this decision because of lack of awareness and others just choose convenience over protection.

8. References

[1] B. Krishnamurthy and C. E. Wills. Privacy diffusion on the Web: A longitudinal perspective, World Wide Web Conference, Madrid, Spain, April 2009 http://Web.cs.wpi.edu/~cew/papers/www09.pdf

[2] B. Krishnamurthy and C. E. Wills. On the Leakage of Personally Identifiable Information Via Online Social Networks, Barcelona, Spain, August 2009 http://Web.cs.wpi.edu/~cew/papers/wosn09.pdf

[3] A. M. McDonald and L. F. Cranor. An Empirical Study of How People Perceive Online Behavioral Advertising, November 10. 2009

[4] A. Raskin, How to Detect Social Sites Your Visitors Use, http://www.azarask.in/blog/post/socialhistoryjs/

[5] Quantcast, July '09, Feb '10, http://www.quantcast.com

[6] M. Nolet, Using Your Browser URL History to Estimage Gender, July 13. 2008, http://www.mikeonads.com/2008/07/13/using-your-browser-url-history-estimate-gender/

[7] What the Internet Knows About You, http://whattheinternetknowsaboutyou.com

[8] Panopticlick, http://panopticlick.eff.org/

[9] Find IP Address, www.find-ip-address.org

APPENDIX

How the Site Works

The goal of this website is to help users become more aware of what is known about them by third-party sites, which are sites that are often hidden from view as users visit sites around the Internet.

This website works by using a piece of JavaScript code to check if various popular and specific types of websites are included in the browser's history. For example, the script checks to see if the URL like www.cnn.com/ has been previously visited. Browsers maintain history information by default so that previously visited sites can be seen and their link color can be changed from the default when a page containing such a link is shown. It is not possible to list the contents of the history via such a script, but only to query whether or not specific URLs are contained within the history. The existence of such a script has been previously publicized with a similar site being WhatTheInternetKnowsAboutYou.com.

What's unique about this website is that it not only shows you sites that you have visited, but also shows you the list of third-party sites that track your behavior across these visited sites. This tracking is done in order to build up a profile of your Internet activities so that your interests and other demographics can be inferred thus allowing targeted advertisements to be served. This website shows your inferred demographics of age range and gender based on the set of sites you visit. It also shows your location based on the Internet address of your machine.

Why the List of Sites Shown is an Approximation

The set of visited sites shown to you and the set of third-party sites present on these visited sites is intended for awareness and is likely only an approximation for any user. There a number of reasons that it is only an approximation:

1. The script cannot examine your entire history, but only query for specific sites. Not all websites containing third-party sites are queried by the script.
2. The set of third-party sites present on a visited site may change over time. The third-parties shown in the results was determined in October 2009.
3. Third-party sites typically use cookies to track users. If you remove cookie information or block advertisements then the list of third-parties shown will not be accurate for you.
4. While this script uses browser history to determine some of the sites you visit, third-party tracking sites do not (to our knowledge) use browsing history to track your behavior. If you periodically remove your history then the list of visited sites shown by our script may be smaller than the list of tracked sites.

What Can I Do to Prevent Tracking

Preventing browser history detection is difficult as discussed in http://whattheinternetknowsaboutyou.com/docs/solutions.html

There are some steps you can take to limit, with various degrees of effectiveness, tracking by third-party sites:

1. Block known third-party tracking content using ad blocking software or extensions. Some content is easily recognizable, but other content is not and in some cases is "hidden" as part of the visited site domain.
2. Disallow third-party cookies. Again this helps to limit tracking, but some third-party sites also track some activity via cookies of the visited site.
3. Disallow all cookies. This prevents all tracking via cookies, but unfortunately the availability of cookies is required by some legitimate sites that you visit so you will then have to selectively allow such cookies.

If you have any comments or questions, please contact us at whattheyknow@cs.wpi.edu