Write-up Question 1 (14 points): Write 1 or 2 sentences describing what you notice about the distribution for each and how these distributions relate to the true prevalence of "yes" answers in the population.

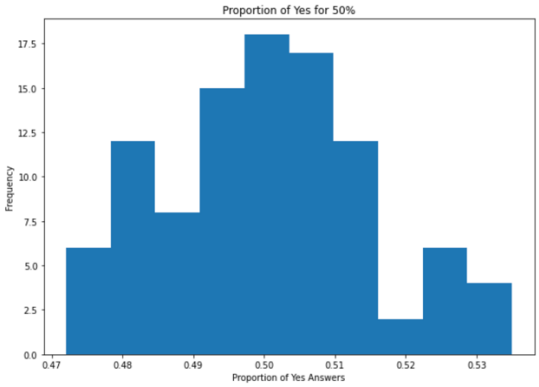


Figure 1

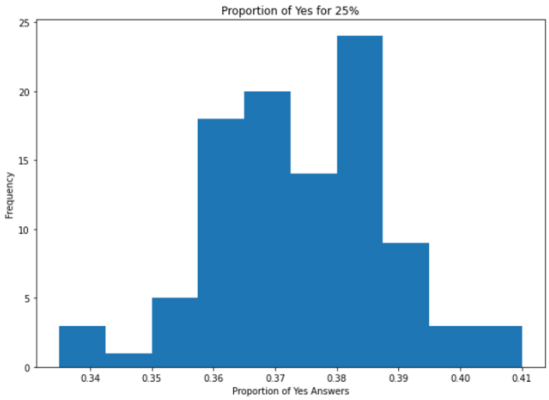


Figure 2

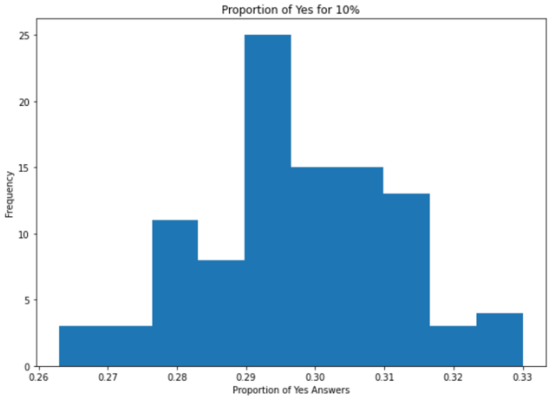


Figure 3

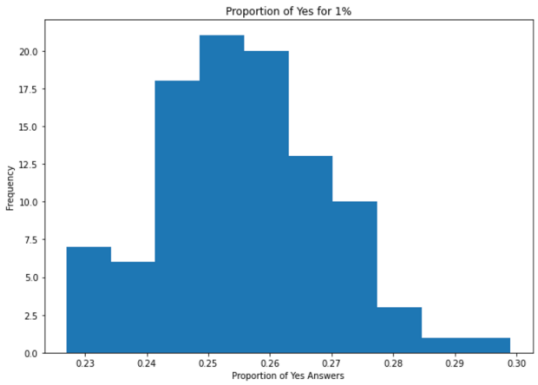


Figure 4

When comparing the various proportions of the “yes” values for each coin flip and corresponding answers, I notice that in Figures 1 and 4, there is a stronger skew towards the right (with the tail being towards the right with a larger frequency of smaller proportions of yes). Figures 2 and 3 seem to be skewed towards the left with a higher frequency of larger “yes” values. Additionally, when considering the percentage of truthful “yes” responses and coin flips that also correspond to a coerced or untruthful yes value, we see the proportions reflecting that. Particularly when comparing the minimum and maximum proportions for each graph and corresponding percent, the proportion of “yes” responses decrease, understandably. Considering the flip to determine whether the “participant” will vote truthfully, there is a 50% chance that a coin will decide their answer and in that case another coin will be flipped with another 50% chance that the answer will be yes. When considering the *other* case when the user then decides their own answer, the 1% graph in particular shows this proportionality as the truth of “yes” is 1% (comparatively very low and close to 0) and, thus, we see how the untruthful “yes” proportions are more apparent at around ¼ of the time.

Write-up Question 2 (7 points): 2. In the randomized response experiment you just ran, the coin flip was parameterized at 1/2 probability of answering truthfully. Try biasing the coin so that the probability deciding to *not* answer truthfully is 1/8, and 1/4, and 3/4. Keep the coin flipping process from before (the one with 1/2 probability) for selecting the answer when you are not answering truthfully. Run the simulations above again for each of these three situations, producing a total of 12 additional histograms. Include these in your write-up and comment in a sentence or two about what you observe.

Coin Bias Variation for 50% Truthful Yes:

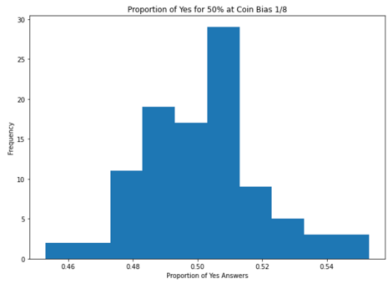


Figure 5

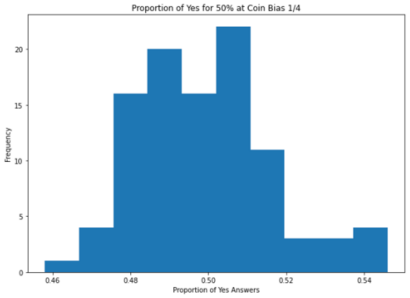


Figure 6

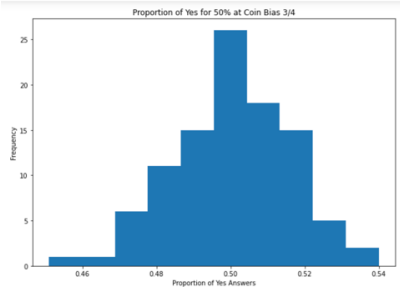


Figure 7

Coin Bias Variation for 25% Truthful Yes:

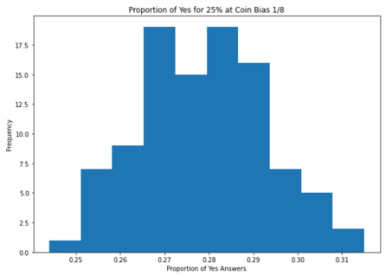


Figure 8

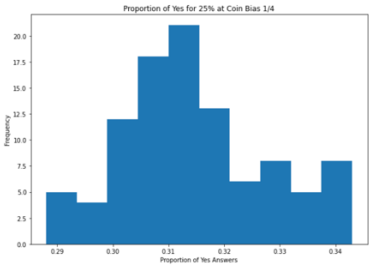


Figure 9

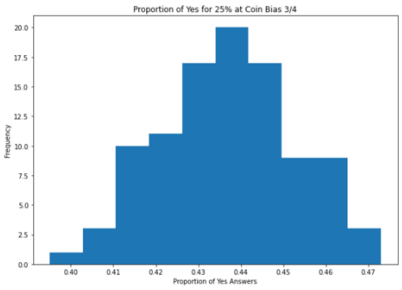


Figure 10

Coin Bias Variation for 10% Truthful Yes:

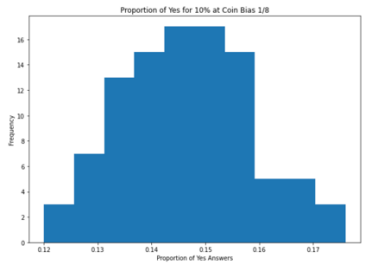


Figure 11

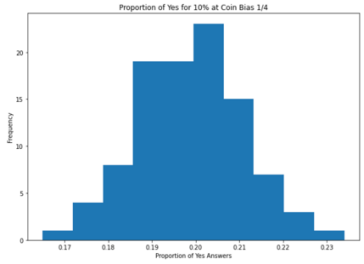


Figure 12

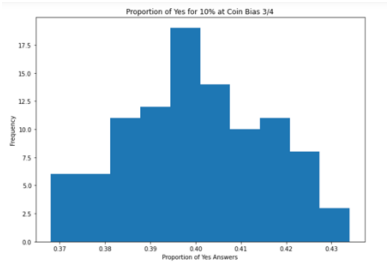


Figure 13

Coin Bias Variation for 1% Truthful Yes:

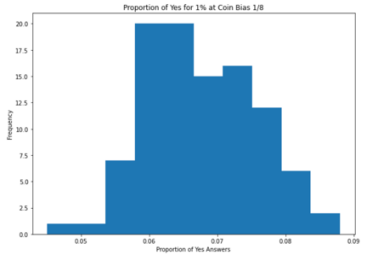


Figure 14

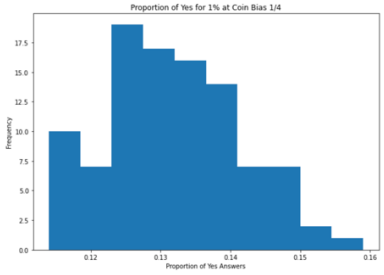


Figure 15

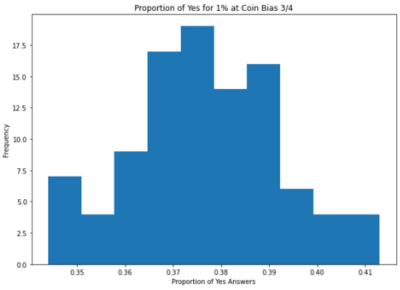


Figure 16

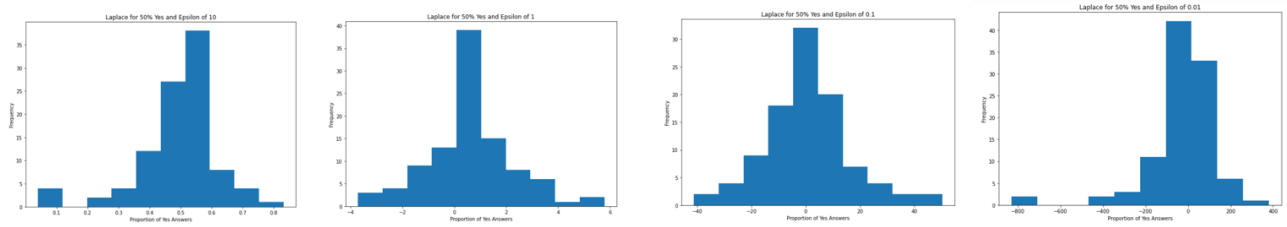
With these graphs, we see what happens to the proportion of "yes" values. This is understandable as, described in Write-up Question 1, how the untruthful values are determined. First, the simulation starts with a flip of either heads or tails, an (ideally) 50% probability of receiving either choice. This means that there is a 50% chance that the simulation will fall into the conditional of an untruthful value that is then determined by another head flip. In this problem, I also assume that the coin flip bias is *only for determining a truthful versus untruthful answer*. This assumption was made based on the instructions statement of "Try biasing the coin so that the probability deciding to *not* answer truthfully is 1/8, and 1/4, and 3/4. Keep the coin flipping process from before (the one with 1/2 probability) for selecting the answer when you are not answering truthfully." I assume that the flip to determine the untruthful "yes" and "no" values are kept at 50%. From this assumption, if the coin flip yields the side that is untruthful according to the bias, the 50% probability of receiving either "yes" or "no" still remains. Thus, we see how decreasing the coin bias as well as the percentage of truthful "yes" values yield relatively drastic decreases in proportion of "yes" values (when comparing ¼ to 1/8 graphs for any given percentage). Additionally, when we compare, say, the ¼ coin bias to the ¾ coin bias, we see an increase in the proportions of the "yes" values for the simulations as the coin is *more* bias to allow the "user" to give a truthful value.

Write-up Question 3 (5 points): The probability of a truthful answer in randomized response is inversely proportional to the amount of "deniability" one has about the answer that one provides to the questioner. Is there a threshold probability at which you would feel comfortable providing information to:

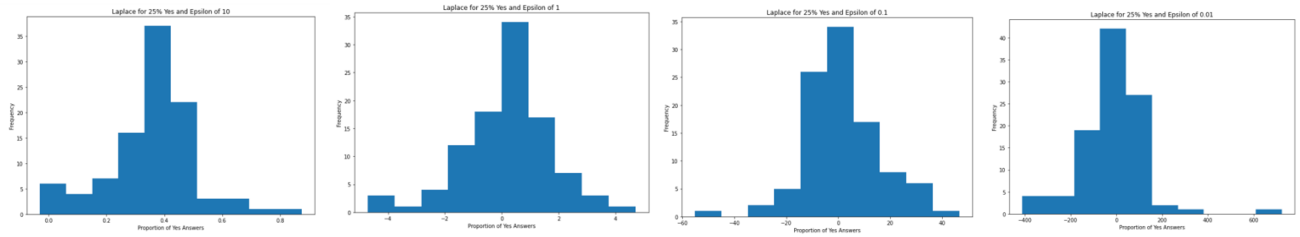
- The government?
  - Probability of 0.9 (90% chance of truth)
    - Assuming the government is acting in an ethical way, I believe it is important for the government to have pieces of information about be that are assumed to be truthful. I would give this a 1.0, but I'm not sure that is an option based on the phrasing of the question. I would hope that information stored about be (such as organ donor status, emergency contact information, social security number, etc.) would be stored correctly in government systems as these are important identifiers to my person.
- A researcher?
  - Probability of 0.9 (90% chance of truth)
    - Assuming my information would not be released to the public, and the researcher has gotten an IRB approval, I would feel comfortable with a higher proportion of my information being assumed to be truthful. Particularly given the measures put in place to protect the research participant, I feel as if my data would be protected in a meaningful way.
- A publicly accessible database?
  - Probability of 0.5 (50% chance of truth)
    - I would be alright with my information available to the public, but I am also assuming that this data has been cleaned as to protect my identity. I also feel like, as a computer science major and avid listener of technology related podcasts, do understand that I as well as most people leave a very large digital footprint, allowing me to be very carful of the things I put both online and locally on my devices (regarding Clouds, search engines, social media, etc.). I obviously wouldn't want my pictures from my 12th birthday taken from my Google drive and plastered on a bunch of billboards, but I am also very well aware that there are laws in place that allow me to maintain possession of my own images even if they are stored on GSuite (please refer to this link).
- Social media?
  - Probability of 0 (0% chance of truth)
    - I really don't like social media, and I think there is a huge lack of transparency regarding the information collected and the means in which it is collected. It's why I

refuse to keep any social media applications on my phone… I feel very strongly that having our information taken from us can quickly lead to user manipulation especially when considering ads, targeted posts, and other ways of psychologically keeping a given user engaged in an unhealthy way in a platform. I definitely would want the as much information kept from social media companies if possible. I recognize that 0 is not a very realistic answer in this sense, but for the purpose of my argument, I'll double down on my answer. I hate social media!

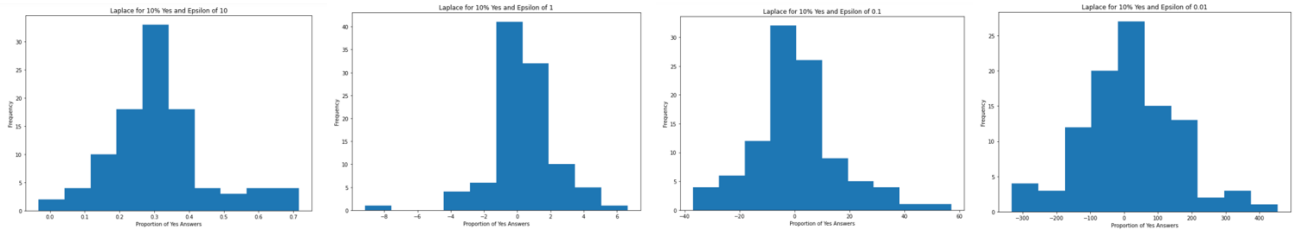
Write-up Question 4 (10 points):
Laplace for 50% Answer "Yes":



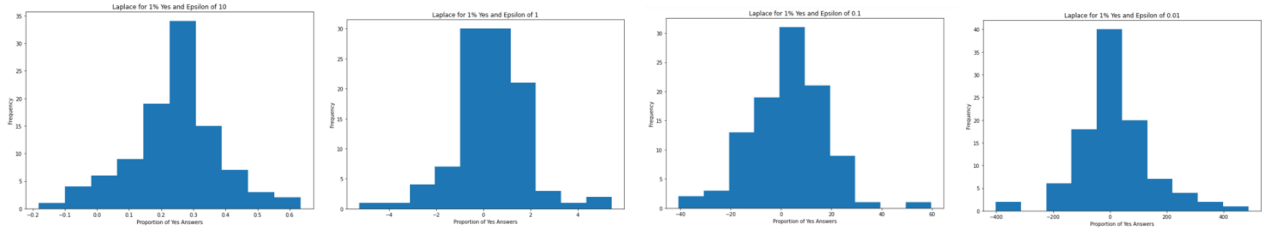
Laplace for 25% Answer "Yes":



Laplace for 10% Answer "Yes":



Laplace for 10% Answer "Yes":



Write-up Question 5 (5 points): Examine the graph you just made. What do you notice about the distribution of answers as epsilon decreases? What happens when epsilon is quite small (e.g., less than 0.01)? Does this pose any problems?

As epsilon decreases, outliers seem to become much more apparent particularly in the frequency of disjointed bars/values from those that are clustered around the peak of the data. Said peak also becomes more prominent as the epsilon values decrease with longer tails in either direction. This could pose an issue of

privacy for the user (referring to the outliers and not the tails) as the outlier values decrease the deniability of a user therefore decreasing the privacy.

Write-up Question 6 (5 points): For each of the points in the graph you just plotted (corresponding to a set of 100 trials), look at the averages. How much do they differ from the true count? What is the standard deviation for each value of epsilon? What does this suggest about repeated queries under the differential privacy framework? How can a data curator/aggregator defend against repeated queries?

When comparing the graphs from Write-up Question 1 and Write-up Question 4, there seems to be a stronger peak in the data in the Laplace graphs. In the graphs from Write-Up Question 1, the values seem more evenly distributed around the "peak" with less of a steep drop off in frequencies of proportions as the x-values increase or decrease away from said peak compared to that of the Laplace inclusive graphs from Write-up Question 4.

```
Mean Square Errors:

50% Truth:
Epsilon = 10
0.017084829859231092
Epsilon = 1
2.658894100090946
Epsilon = 0.1
237.49302167803356
Epsilon = 0.01
28744.99993512124

25% Truth
Epsilon = 10
0.023920675334046407
Epsilon = 1
2.48716420963953
Epsilon = 0.1
221.49213759542522
Epsilon = 0.01
19103.203104288405

10% Truth
Epsilon = 10
0.020236736312412473
Epsilon = 1
3.8097035627908467
Epsilon = 0.1
254.7402781645438
Epsilon = 0.01
19468.26189071885

1% Truth
Epsilon = 10
0.01959735272196923
Epsilon = 1
2.210104337260981
Epsilon = 0.1
218.3979669355029
Epsilon = 0.01
16454.16436365016

Standard Deviations:

50% Truth:
Epsilon = 10
0.131606765957714 87
Epsilon = 1
1.6273592199806999
Epsilon = 0.1
15.40381430739658
Epsilon = 0.01
166.94711049557617

25% Truth
Epsilon = 10
0.153238180785668
Epsilon = 1
1.5633601317611943
Epsilon = 0.1
14.840887592614271
Epsilon = 0.01
137.84393004571388

10% Truth
Epsilon = 10
0.14122916920030018
Epsilon = 1
1.9465900152207356
Epsilon = 0.1
15.942567802569265
Epsilon = 0.01
138.60007779211097

1% Truth
Epsilon = 10
0.14120304122485025
Epsilon = 1
1.4838583172705724
Epsilon = 0.1
14.585379869226017
Epsilon = 0.01
127.52540635305687
```

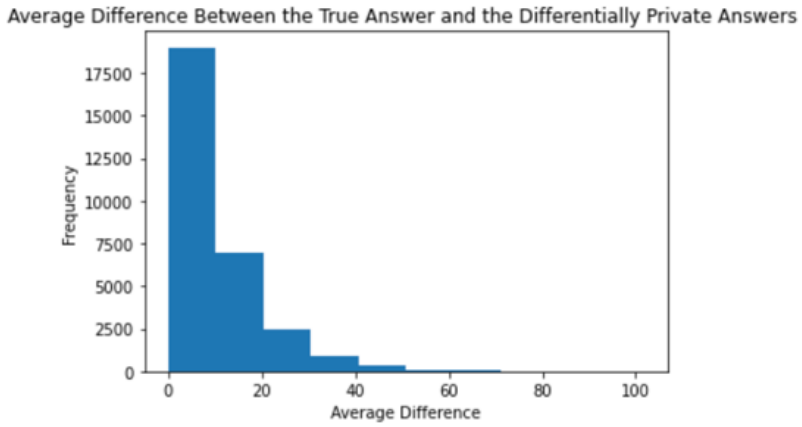
The standard deviations and mean square errors are depicted to the left. The standard deviation values show that, in all cases of varying percentages of truth, the lower the epsilon value, the higher the standard deviation. This is clearly seen in the graphs from Write-Up Question 4, as we see how the higher the epsilon values, the more outliers/the more "spread out" the graphs get. Since a low standard deviation indicates a cluster of the data closer to the mean and a higher standard deviation means that data is more spread out, the graphs are representative of this statistic, especially when comparing the varying epsilon values grouped across the varying percentages of truthful "yes" answers given.

This suggests that, regarding repeated queries under the differential privacy framework, that the higher the epsilon value, the less privacy the test subjects possess. With more repeated queries, the more privacy is lost for the user. This would mean that the privacy budget would be less meaningful and the more risk a user/participant faces when considering their privacy protection in a given experiment. With fewer repeated queries, the more protection the user/participant of the research experiment maintains.

Write-up Question 7 (7 points): First, briefly explain how you compute the sensitivity for the mean of a real valued list. Reflecting on this approach, if you drop data outside of the range, how would that affect the sensitivity analysis? Could this inadvertently reveal information meant to remain private?

I compare the real data without the inclusion of epsilon/Laplace manipulations to the list that *does* include epsilon/Laplace manipulation and scaling to the values. This can certainly lead to the release of data in a breached manner, as the *real value* list is being used to compare the Laplace scaled lists to. If I were to drop information outside of a given range, this would alter the sensitivity analysis by potentially revealing closer values in the lists as outliers would ideally be removed and the privacy of the user would, in turn, me more susceptible to identification when comparing the real and Laplace scaled value lists.

Write-up Question 8 (19 points):



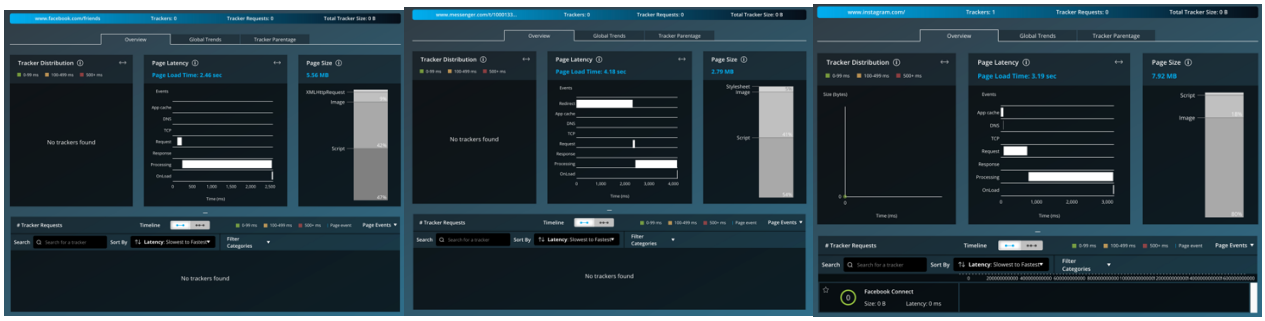
Write-up Question 9 (4 points): Look at magnitude and direction of the error in the plot. What does this suggest about choosing reasonable cutoffs when handling differentially private data? How might one go about minimizing this sort of error? In interpreting the graph, it may be useful to also plot a histogram of the age.

The skew of the plot is relatively heavy toward the right. This suggests that using cutoffs for the data would increase the deniability and privacy of the user as, when the difference between the true answer and the differentially private answer is *smaller*, then this means the true answers are close to the differentially private answers, indicating a higher level of deniability for the user and a stronger protection of the user's privacy. Using bounding and cutoffs especially in the realm of the right tailing data would lead to a stronger protection of the user's privacy through minimizing the average difference. If the average difference is minimized hence leading to closer similarities between the true and differentially private data, then the data would be considered more protected as the users would have higher levels of deniability hence a higher level of privacy protection.

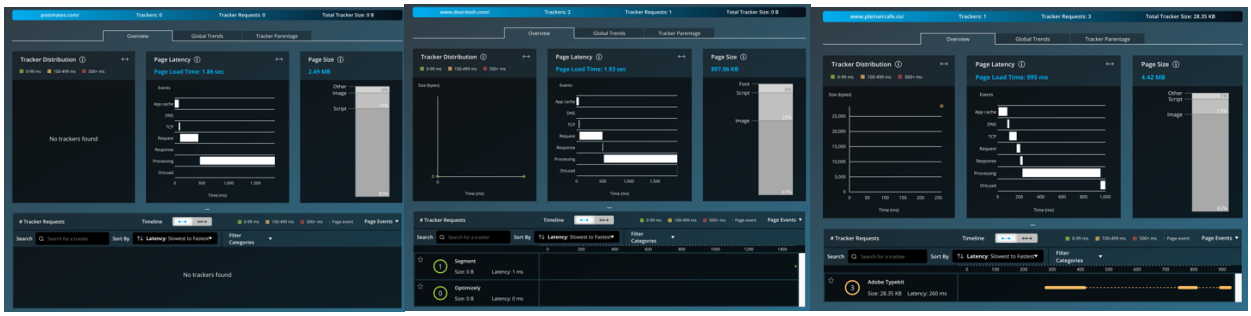
Write-up Question 10 (12 points): On what sites do you seem to encounter the largest number of trackers, and on what sites do you seem to encounter the least? What did you find most surprising from using your chosen tool to see who is tracking you?

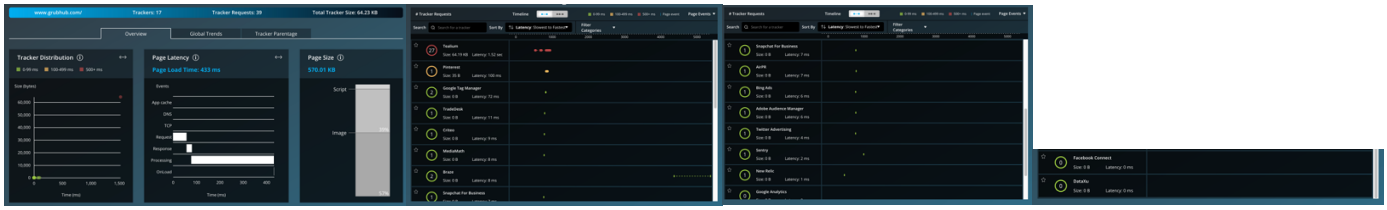
1. On what sites do you seem to encounter the largest number of trackers?
    a. I encountered the largest number of trackers on GrubHub. I truly could not reason through why GrubHub, of not only the Food Ordering websites but also of all of the websites in general had such a high number of trackers detected!
2. On what sites do you seem to encounter the least number of trackers?
    a. I encountered the fewest trackers on the sites relating to Social Media (specifically Facebook and Facebook Messenger). There were also no trackers detected by Ghostery on sites such as Postmates, Google Search, and Jupyter.
3. What did you find most surprising from using your chosen tool to see who is tracking you?
    a. I was most surprised to find that GrubHub had the most trackers detected. I would have expected that websites such as Instagram would have that same level of trackers, but I was also surprised to find that they did not but rather had *none* at all or only one tracker. I would have thought the roles would have been reversed for these websites.

Social Media:

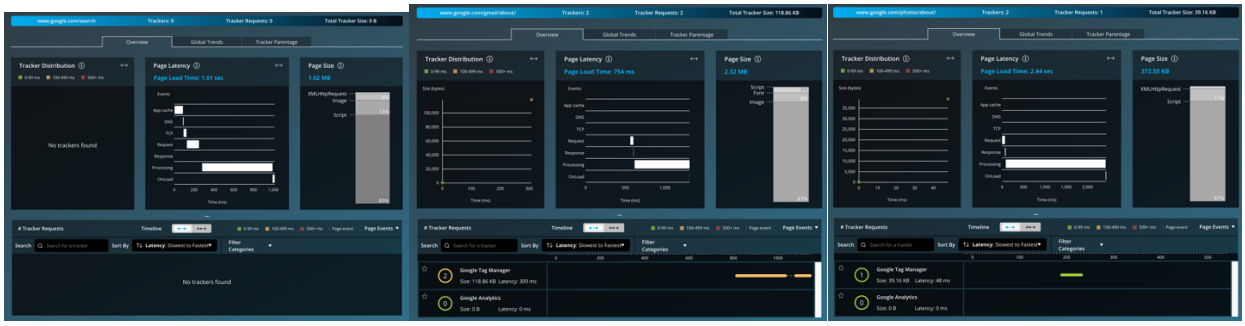


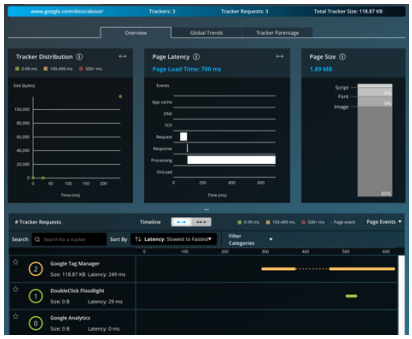
Online Food Ordering:



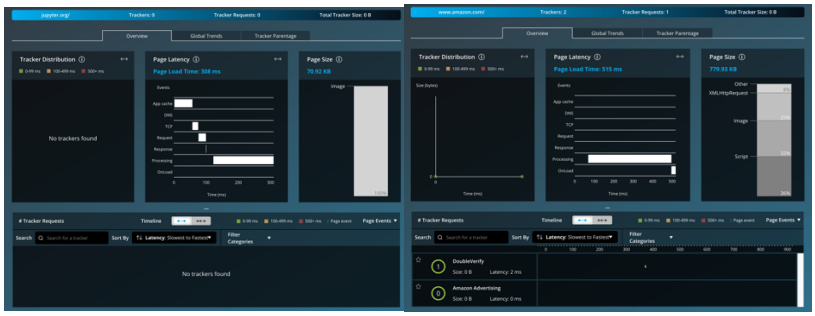


Google Suite:





Other Websites I Frequent:



Write-up Question 11 (12 points): Who might be the intended users of this tool from what you can gather? What information does the tool seem to focus on conveying, and do you think it's the right information to convey to those users? If you were redesigning the tool, what might you consider doing differently?

I was initially under the belief that this platform was built for the average computer user (someone who does not possess a strong knowledge of computers and user privacy) but, as I began using it for this assignment, I realized I did not have a solid grasp on the statistics I was looking at. This is what leads me to believe that this browser and the data collected from it are geared towards people who are more attune with privacy terminology and the field of privacy itself. This would include people who are researching it or work in a field that deals heavily with the privacy of a user rather than the user themselves. I drew this conclusion based on the more specific statistics given that aren't merely "one a scale of one to ten, how bad is this website in terms of user tracking?". Assuming the target audience is researchers and more knowledgeable user privacy industry professionals, I would say this browser does a very good job at conveying relevant information that accurately conveys a level of detail in the statistics that is required for certain aspects of research that may be done in both academic and corporate environments. It is, of course, very dependent on the context and use of the information and what statistics are required for certain projects, but, for the purpose of this assignment, the trackers were very straight forward and simple to understand. I, however, did not know how to interpret or understand the severity of statistics such as DNS, TCP, Response, Processing, OnLoad, etc. Before encountering this question and speaking only as a user, I would consider a more minimal design approach for the platform. If it were intended for more knowledgeable audiences/computer scientists and researchers, I think the design works well for the intended audience. However, if the designers look to broaden the scope of users, I believe they should consider a more user-friendly interface including a simpler layout as well as simpler/more generic statistics that are geared towards an audience that knows a bit less about computer security and privacy. I also spent a solid three minutes trying to minimize the statistics I accidentally opened on the bottom of my browser that I finally got to open in another tab. That functionality (the default of opening and taking up approximately 40% of my viewable page in the browser) as well as not opening in a new tab initially is something that could be improved upon just from my experience.