Temporal and Behavioral Analysis of Web Traffic Data

I. Introduction

Understanding the web traffic patterns of university students is crucial for gaining insights into their online behaviors, optimizing campus network performance, and addressing potential security concerns. This report focuses on analyzing a dataset collected from Indiana University, which includes approximately 235 million HTTP requests over a 22-day period in November 2009.

The original dataset chosen for this study is the Indiana University Clicks dataset, which was collected in 2009 as part of the "Analysis & Modeling of Online Traffic Patterns" project at the **Center for Complex Networks and Systems Research** (CNetS) at Indiana University. The project was led by Fil Menczer, along with a team of collaborators, and produced several influential papers presented at conferences like WSDM and Hypertext. Their goal was to develop realistic models of web and social media browsing, which could improve network design, traffic forecasting, site classification, and search ranking algorithms.

The primary aim of this study is to perform a detailed temporal and behavioral analysis of students' web activities. By examining temporal trends, peak usage periods, daily and weekly behavioral patterns, and anomalies in web activity can be identified. These patterns are important for improving user experience, optimizing network resource allocation, and enhancing security measures. Additionally, analyzing source-destination relationships will shed light on common navigation paths, the popularity of various websites, and potential security risks, providing a comprehensive understanding of university students' online behaviors.

II. Data

The dataset used in this study was sourced from Indiana University and captures web traffic over a 22-day period in November 2009. This dataset comprises approximately 235 million HTTP requests, offering a detailed snapshot of web activity within the university network, primarily reflecting the behaviors of students.

Dataset Structure

The dataset is organized into 22 files, each representing a day's worth of HTTP requests from November 1 to November 22, 2009. Each file includes the following columns:

- **count (int):** The number of times a specific HTTP request was made.
- **timestamp (int):** The Unix timestamp indicating when the request occurred.
- **from (string):** The source URL from which the request originated.
- **to (string):** The destination URL to which the request was directed.

Key Statistics

The overall structure and key statistics of the dataset are as follows:

- **Total Files:** 22

- **Total Columns per File:** 4 (count, timestamp, from, to)
- **No Missing Values**
- **No Missing Columns**
- **Minimum Number of Clicks per Day:** 1
- **Maximum Number of Clicks in a Single Day:** 309,333 (November 7)
- **Highest Total Number of Clicks in a Single Day:** 12,372,546 (November 3)

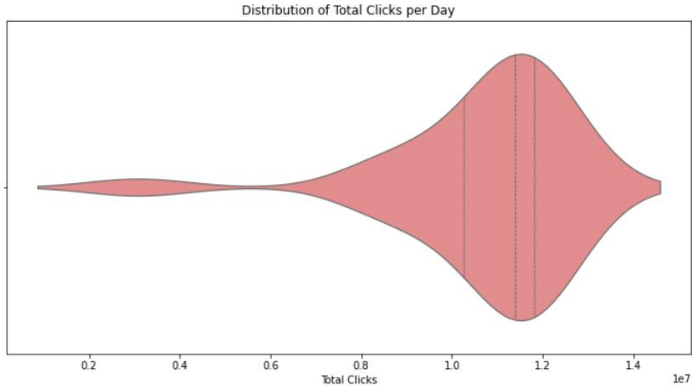


Figure 0. Violin Plot of Total Clicks per Day

This dataset provides a comprehensive snapshot of university students' online behavior over a period of 22 days. The consistent structure and completeness of the dataset enable a wide range of analyses, from identifying temporal trends to exploring source-destination relationships and clustering content. This foundational analysis sets the stage for more advanced investigations into student behaviors, ensuring that the findings are both comprehensive and actionable.

III. Analysis

Claim 1: Social Media and Google Services Dominate Online Activities of Indiana University Students

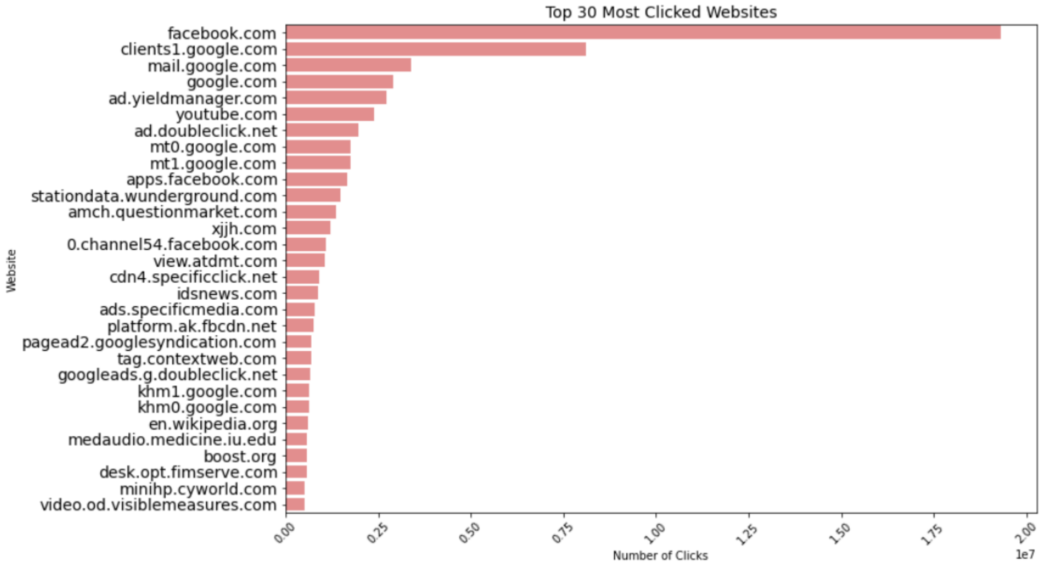


Figure 1. Top 30 Most Visited Domains

The data in Figure 1 reveals that facebook.com received the highest number of clicks, significantly surpassing other websites. This suggests that social media platforms, particularly Facebook, play a central role in university students' online activities. In addition to Facebook, services provided by Google, such as clients1.google.com, mail.google.com, and google.com, also received substantial traffic. This indicates that students rely heavily on a few key platforms for social interaction, information retrieval, and communication.

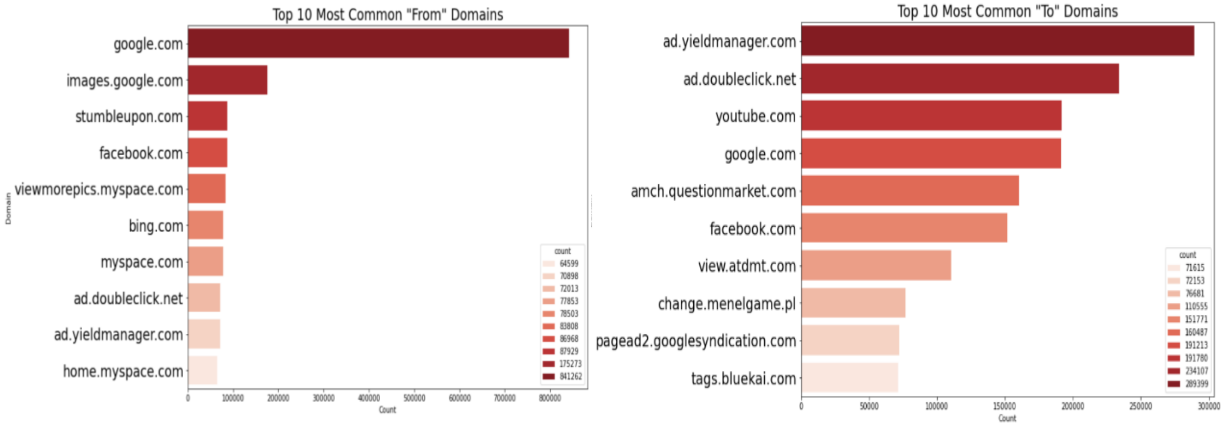


Figure 2. Top 10 Most Common Source-Destination Domains

The first graph in Figure 2 shows that google.com is the most common source domain, followed by images.google.com and stumbleupon.com. This indicates that a significant portion of student web traffic originated from Google's search and image services, as well as from social media and content discovery platforms like StumbleUpon and Facebook. Other notable source domains include myspace.com and bing.com, reflecting students' use of alternative social and search platforms.

Claim 2: There are security risks in web browsing behavior of University of Indiana Students.

The second graph in Figure 2 highlights the most common destination domains, with ad.yieldmanager.com and ad.doubleclick.net topping the list. A simple internet search reveals that these are advertising and tracking services, and the data suggests that a large volume of student traffic was directed towards these advertising and tracking services.

Other prominent destinations include youtube.com, google.com, and facebook.com, once again underscoring the popularity of these platforms not only as source websites, but also destination websites.

However, the significant traffic directed toward market research and tracking services, such as amch.questionmarket.com and view.atdmt.com, poses potential privacy concerns. These services could exploit student data for targeted ads or other tracking purposes, indicating a need for stricter controls and user education to mitigate these vulnerabilities.

Claim 3: Indiana University Students Exhibit Periodic Daily and Weekly Web Traffic Patterns.

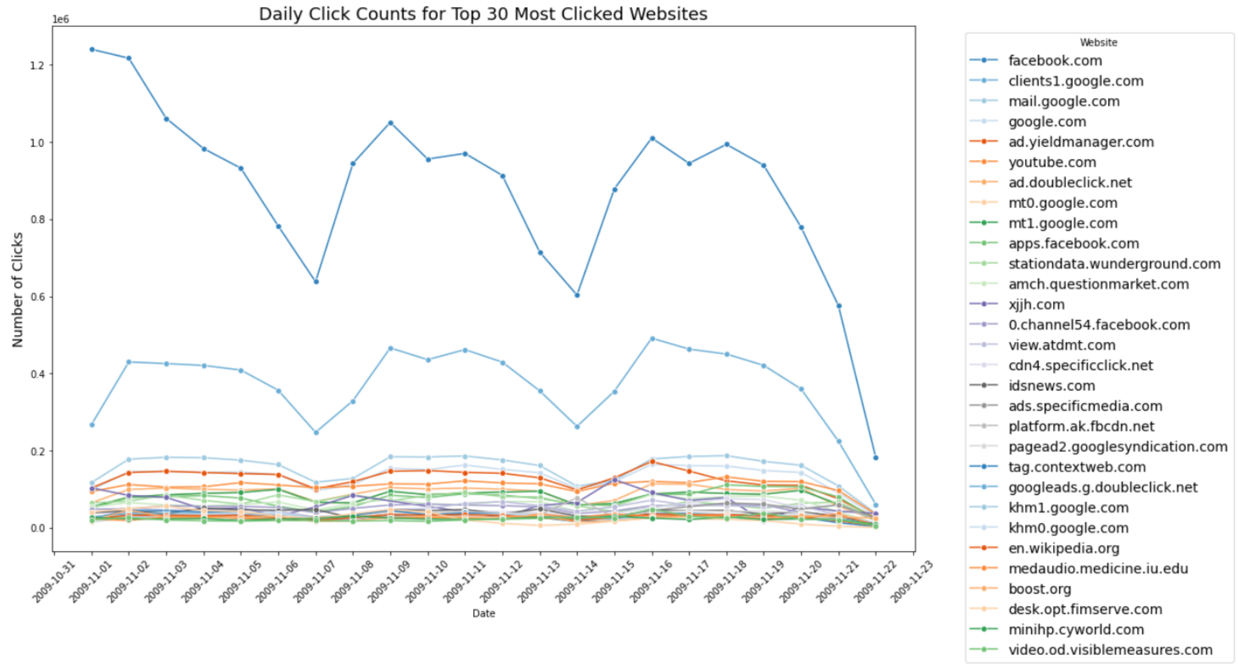


Figure 3. Daily Click Counts for Top 30 Most Visited Domains

A certain periodicity in web traffic can be observed through fluctuations in daily click counts, as shown in Figure 3. Notable peaks and troughs, particularly for facebook.com and clients1.google.com, indicate varying levels of student engagement throughout the day. The stable daily click counts for google.com and mail.google.com suggest consistent student engagement, reflecting these services' essential role in their daily routines. However, a clear pattern of decreased click counts every seven days, starting from November 1, is evident.

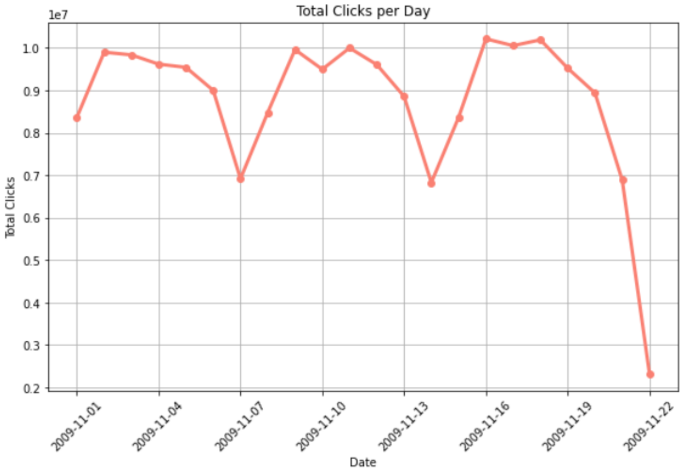


Figure 4. Daily Click Trends

Aggregating click counts for each day into one chart, as shown in Figure 4, confirms that there is indeed a periodic trend. Significant drops in total clicks on November 1, November 7, November 14, and November 21 likely indicate a weekly trend where students' web activity decreases on certain days of the week, possibly due to changes in routine or offline activities during the weekends.

Claim 4: Web traffic depends on the day of the week at Indiana University.

Anomalies detected in web activity, such as the significant drop in clicks on Saturdays, were identified using statistical methods. Specifically, an Analysis of Variance (ANOVA) test was used to determine whether the day of the week significantly influences web traffic.

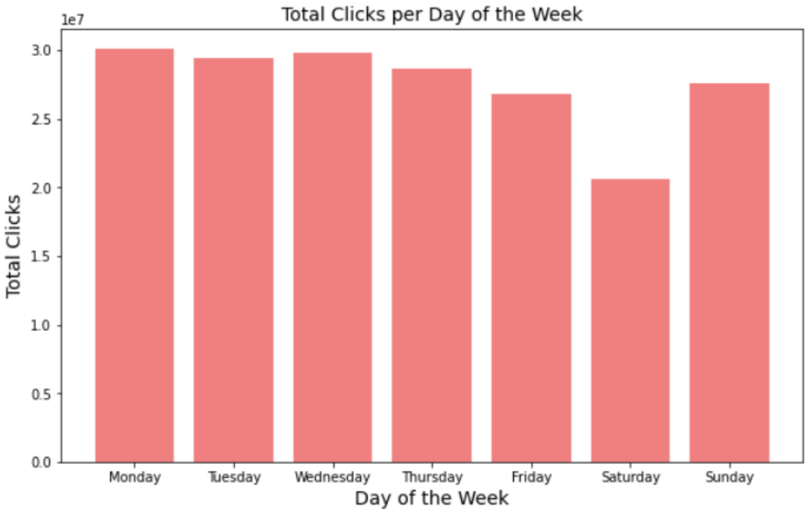


Figure 5. Click Count by Day of Week

Figure 5 reveals that Saturdays have the lowest average click counts, which is further substantiated by the observation that November 1, November 7, November 14, and November 21 in 2009 were all Saturdays. This suggests that students generally engaged less in online activities on Saturdays, likely due to offline activities or changes in routine.

F-Statistic	P-Value
3.457	0.024

Figure 6. ANOVA Significance of Day of the Week

An ANOVA test (Figure 6) was used to test the significance of the day of the week on click counts and corroborate the findings from Figure 5 by comparing a model with the day of the week as a factor and one that does not. ANOVA compares the variance between the group means (in this case, the days of the week) to the variance within the groups. If the between-group variance is significantly larger than the within-group variance, the test will reject the null hypothesis, indicating that the day of the week does indeed have a significant effect on web traffic.

The test was set up as follows:

- **Null Hypothesis:** There is no difference in the mean click counts across the different days of the week. In other words, the day of the week has no significant effect on web traffic.
- **Alternative Hypothesis:** At least one day of the week has a different mean click count compared to the others, implying that the day of the week does significantly influence web traffic.

The result indicates that at a significance level of 0.05, an F-statistic of 3.457 and a P-Value of 0.024, the null hypothesis is rejected and the alternate hypothesis is accepted, suggesting that there is a statistically significant relationship between the day of the week and click counts. This implies that the variations in web traffic are influenced by the day, with weekends, especially Saturdays, showing lower engagement compared to weekdays.

Claim 5: Indiana University students are active on the internet regardless of the time of day.

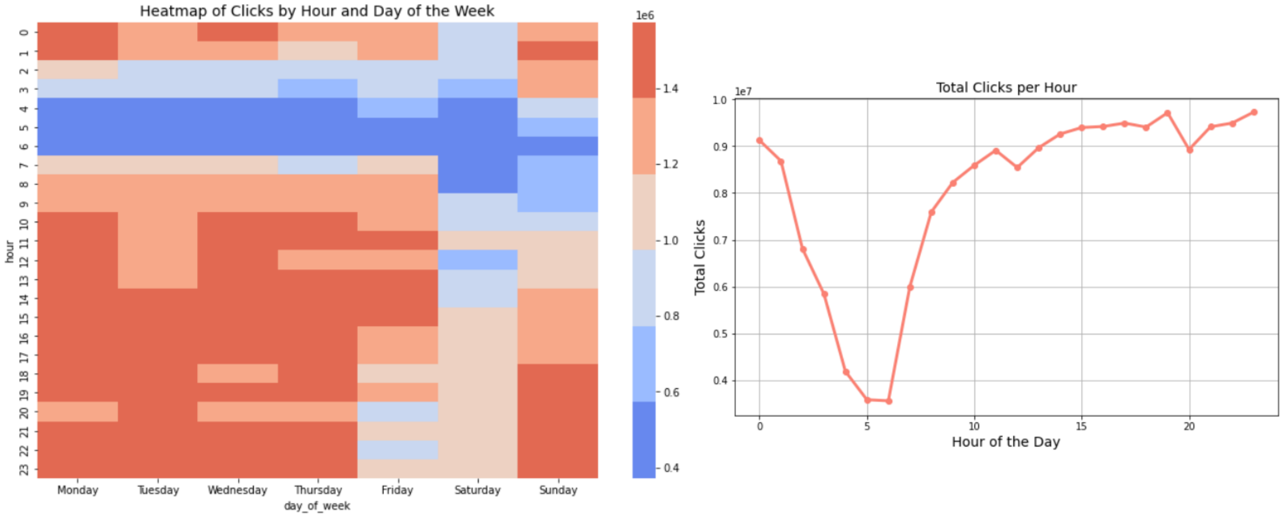


Figure 7. Hourly Click trends

The heatmap in Figure 7 reveals that the lowest click counts occur during the early morning hours between 4 AM and 6 AM, while the highest click counts are prevalent during the late evening and early morning hours, particularly on weekends. This pattern suggests that students tend to engage in online activities more during leisure hours, with weekend activity peaking later in the day. The line graph of total clicks per hour further supports these findings, showing a clear daily cycle where web traffic peaks around midnight, drops to the lowest point around 4-5 AM, and then gradually increases throughout the day.

Again, an ANOVA test is conducted, comparing a model with and without the hour of the day as a factor. The test was set up as follows:

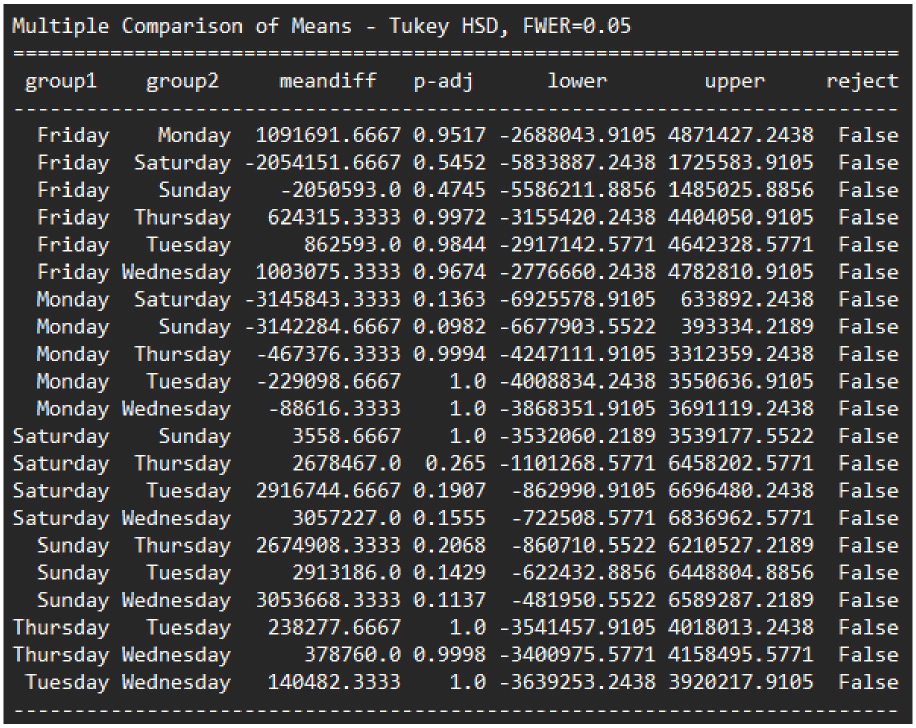
- **Null Hypothesis:** There is no difference in mean click counts across different hours of the day.
- **Alternate Hypothesis:** At least one hour has a different mean click count compared to the others.

F-Statistic	P-Value
26.770	2.332e-115

Figure 8. ANOVA Significance of Hour of Day

The results in Figure 8 indicate that the test yielded an F-Statistic of 26.770 and an extremely small P-Value of 2.332e-115, indicating that at a significance level of 0.05, the null hypothesis is rejected and the alternate hypothesis is accepted, leading to a conclusion that the hour of the day has a highly statistically significant effect on click counts. These results confirm the presence of distinct temporal patterns in student web usage, with specific hours showing consistently higher or lower activity.

To further understand the differences in web traffic between days of the week, a Tukey HSD (Honestly Significant Difference) test was conducted to compare the means of different groups to identify significant differences.



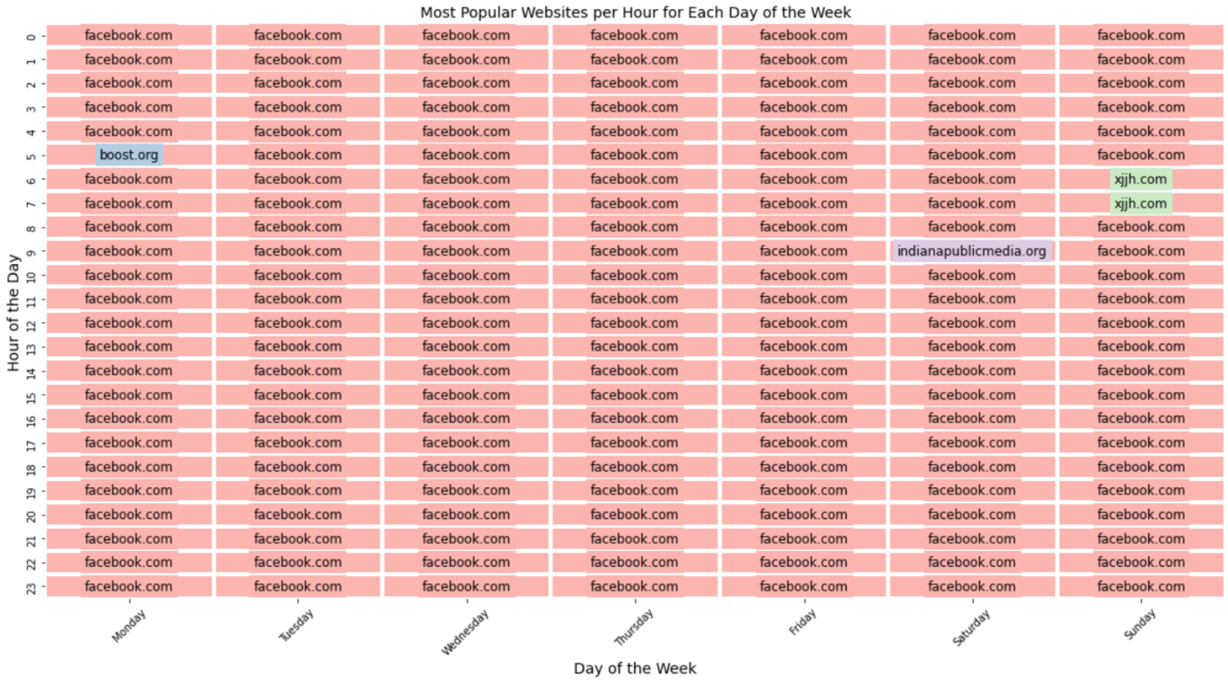
```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
==========================================================================
 group1    group2     meandiff   p-adj     lower         upper      reject
--------------------------------------------------------------------------
  Friday    Monday  1091691.6667 0.9517 -2688043.9105 4871427.2438   False
  Friday  Saturday -2054151.6667 0.5452 -5833887.2438 1725583.9105   False
  Friday    Sunday    -2050593.0 0.4745 -5586211.8856 1485025.8856   False
  Friday  Thursday   624315.3333 0.9972 -3155420.2438 4404050.9105   False
  Friday   Tuesday      862593.0 0.9844 -2917142.5771 4642328.5771   False
  Friday Wednesday  1003075.3333 0.9674 -2776660.2438 4782810.9105   False
  Monday  Saturday -3145843.3333 0.1363 -6925578.9105  633892.2438   False
  Monday    Sunday -3142284.6667 0.0982 -6677903.5522  393334.2189   False
  Monday  Thursday  -467376.3333 0.9994 -4247111.9105 3312359.2438   False
  Monday   Tuesday  -229098.6667    1.0 -4008834.2438 3550636.9105   False
  Monday Wednesday   -88616.3333    1.0 -3868351.9105 3691119.2438   False
Saturday    Sunday     3558.6667    1.0 -3532060.2189 3539177.5522   False
Saturday  Thursday     2678467.0  0.265 -1101268.5771 6458202.5771   False
Saturday   Tuesday  2916744.6667 0.1907  -862990.9105 6696480.2438   False
Saturday Wednesday     3057227.0 0.1555  -722508.5771 6836962.5771   False
  Sunday  Thursday  2674908.3333 0.2068  -860710.5522 6210527.2189   False
  Sunday   Tuesday     2913186.0 0.1429  -622432.8856 6448804.8856   False
  Sunday Wednesday  3053668.3333 0.1137  -481950.5522 6589287.2189   False
Thursday   Tuesday   238277.6667    1.0 -3541457.9105 4018013.2438   False
Thursday Wednesday      378760.0 0.9998 -3400975.5771 4158495.5771   False
 Tuesday Wednesday   140482.3333    1.0 -3639253.2438 3920217.9105   False
--------------------------------------------------------------------------
```

Figure 9. Tukey HSD (Honestly Significant Difference)

The results show no significant differences in click counts between any pairs of days, as all p-values are above 0.05. For instance, the comparison between Friday and Monday yields a mean difference of 1,091,691.67 clicks with a p-value of 0.952, indicating no statistically significant difference. Similarly, comparisons between other days, such as Friday vs. Saturday and Monday vs. Sunday, also show non-significant results. These findings suggest that while there are observable daily patterns in web traffic, the overall differences between specific days are small.

Claim 6: Social Media Played a Central Role in Indiana University Students' Lives in 2009.

The data shows that Facebook consistently ranked as the top domain visited at any hour, as illustrated in Figure 10. This dominance indicates that social media was a central part of students' online activities even in 2009. Although other domains, such as boost.org, indianapublicmedia.org, and xjjh.com, appeared sporadically, facebook.com remained the most popular domain by far. This suggests that the appeal of social media platforms, particularly Facebook, was already well-established among university students, attracting significant user engagement at all times.

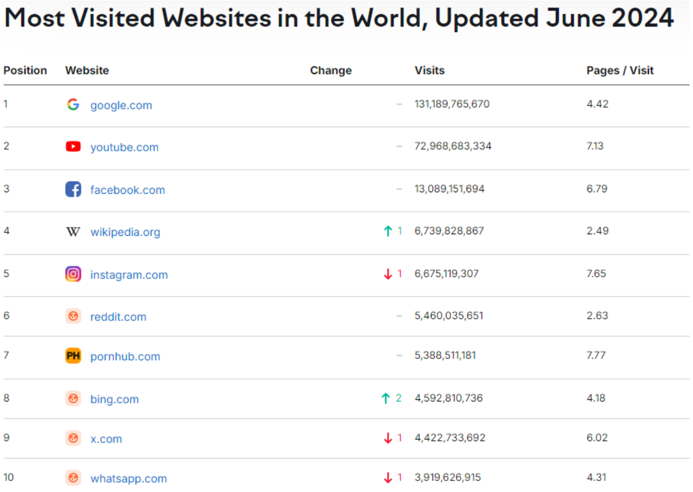


	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
0	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
1	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
2	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
3	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
4	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
5	boost.org	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
6	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	xjjh.com
7	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	xjjh.com
8	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
9	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	indianapublicmedia.org	facebook.com
10	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
11	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
12	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
13	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
14	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
15	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
16	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
17	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
18	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
19	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
20	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
21	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
22	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com
23	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com	facebook.com

Figure 10. Most Popular Websites Per Hour Each Day of the Week

Claim 7: Social media remains central to people’s online activities.

While the analysis focuses on data from November 2009 due to limitations in the data, it is important to consider how the landscape of popular websites has evolved over time. In 2009, platforms like Facebook were already central to university students' online activities. However, the digital landscape has shifted significantly since then. Using Semrush's analytics tool, a comparison of the most visited websites globally in the present day reveals both continuity and change. It is important to note, however, that these are statistics for global users worldwide, and not just students at Indiana University.



Most Visited Websites in the World, Updated June 2024

Position	Website	Change	Visits	Pages / Visit
1	google.com	–	131,189,765,670	4.42
2	youtube.com	–	72,968,683,334	7.13
3	facebook.com	–	13,089,151,694	6.79
4	wikipedia.org	↑ 1	6,739,828,867	2.49
5	instagram.com	↓ 1	6,675,119,307	7.65
6	reddit.com	–	5,460,035,651	2.63
7	pornhub.com	–	5,388,511,181	7.77
8	bing.com	↑ 2	4,592,810,736	4.18
9	x.com	↓ 1	4,422,733,692	6.02
10	whatsapp.com	↓ 1	3,919,626,915	4.31

Figure 11. Most popular websites in the world, June 2024 (Semrush)

Figure 11 shows that Facebook remains one of the top platforms, newer entrants like Instagram and Reddit have become major players, reflecting changes in user preferences and the rise of different forms of social media and information-sharing platforms while still showing a continued dominance of social media in the top 10. Additionally, search engines like Google and content platforms like YouTube continue to dominate, similar to their influence in 2009.

IV. Conclusion

This report provides a comprehensive analysis of university students' web traffic behavior at Indiana University over a 22-day period in November 2009. Keeping the limitations of the data in mind, the analysis highlights several key findings:

1. **Dominance of social media and Google Services:** Social media platforms such as Facebook, and Google services were central to students' online activities.
2. **Potential Security Risks:** The significant traffic directed toward advertising and tracking services poses privacy concerns, suggesting a need for stricter network controls and user education to protect student data.
3. **Periodic and Temporal Web Traffic Patterns:** There are clear daily and weekly, and hourly patterns in web traffic, with students showing less online activity on Saturdays and more activity during late evening hours, particularly on weekends, and exhibiting a sharp decrease in activity from 4-5 AM every day of the week.
4. **Central Role of social media:** Facebook's dominance as the most visited domain across all hours each day of the week underscores the central role that social media played in the lives of university students in 2009.

V. Limitations

While this analysis offers valuable insights into the web traffic behavior of university students at Indiana University in November 2009, several limitations must be acknowledged:

1. **Data Specificity:** The dataset is specific to Indiana University and only covers a 22-day period in November 2009. This temporal and institutional specificity limits the generalizability of the findings. The observed patterns might not apply to other universities, regions, or time periods, particularly given the rapid evolution of online behavior and technology since 2009.
2. **Temporal Constraints:** The analysis is confined to 22 days, which may not fully capture the seasonal or event-based variations in student web traffic.
3. **Data Limitations:** The dataset does not include demographic data (e.g., undergraduate vs. graduate students) or ways to identify individual usage patterns. Other potentially influential factors such as specific academic disciplines or dormitory locations are also not included.
4. **ANOVA vs Tukey HSD:** The different findings between ANOVA and the Tukey HSD test suggest that while there is a general difference in web traffic across the week (as identified by ANOVA), the differences between specific pairs of days are not large enough to be statistically significant when accounting for multiple comparisons. This is likely due to a difference in what ANOVA and Tukey HSD tests are testing. ANOVA tests whether there is any overall difference among the group means, but it does not tell us which specific groups differ, while Tukey HSD is more conservative in making conclusions and performs multiple pairwise comparisons to determine exactly which groups (days) differ from each other. This indicates that, while the overall trend shows some variation depending on the day, the specific daily differences are subtle and do not consistently reach statistical significance.

VI. Action Items from First Draft

1. **Unclear Topic & Dataset Background:** The report was rewritten to be more specific about what it is analyzing, such as being specific about analyzing web browsing behaviors of students at Indiana University, and being open and clear about the limitations in the dataset. Additionally, the introduction section now includes a brief background on the dataset.
2. **Explanation of Methods for Identifying and Classifying Websites and Services:** The report was updated to include detailed explanations for how different types of websites and services were identified and classified in the dataset, such as in Claim 2, where the analysis focuses on specific categories like advertising and tracking services.
3. **Detection and Handling of Anomalies in Web Activity:** Discussions on how anomalies in web activity were detected, such as using statistical methods like ANOVA, were

included in the updated report. An example is in Claim 4, where significant drops in clicks on Saturdays were identified and analyzed.

4. **Elaboration on Statistical Analyses:** The report was updated to include detailed explanations of the ANOVA and Tukey HSD tests, including the null and alternate hypotheses, the resulting significance, F-statistics and P-values.
5. **Logic Behind Claims:** The claims in the report are now logically connected, with clear explanations for how each claim was derived from the data. For example, Claim 3 connects daily and weekly web traffic patterns to the broader context of student behavior, and Claim 6 highlights the role of social media based on observed traffic patterns.
6. **Inclusion of Comparative Analysis with Recent Data:** The report includes a comparison of the 2009 data with more recent data from 2024. Semrush's analytics tool was used to highlight continuity and change in the popularity of websites over time.
7. **Discussion of Potential Security Risks:** The report discusses potential security risks identified through the analysis in Claim 2, which highlights the risks posed by significant traffic directed toward advertising and tracking services and suggests the need for stricter network controls and user education.
8. **Expanded Limitations Section:** The limitations section has been expanded to discuss limitations due to the specificity of the dataset, temporal constraints, the lack of demographic data and the use of HTTP-only data. Differences in findings between ANOVA and Tukey HSD tests are also explained.

VII. Code, Data & References

1. **Code**
   a. Chen, Z. (n.d.). *web-traffic-analysis*. GitHub. https://github.com/zihengcchen/web-traffic-analysis
2. **Data**
   a. *Analysis & Modeling of Online Traffic Patterns*. CNetS. (n.d.). https://cnets.indiana.edu/groups/nan/webtraffic
   b. Nikolov, D., & Menczer, F. (2020, January 24). *Web clicks from Indiana University*. Zenodo. https://zenodo.org/records/2650234
3. **References**
   a. *Top websites in the world - June 2024 most visited & popular rankings*. Semrush. (n.d.). https://www.semrush.com/website/top/?utm_campaign=most-popular-websites