Exploring Claremont Colleges Library Resource Usage by Wireless Internet Sessions

Lucius Bynum, Andi Chen, Matthew Guillory, Daniel King
Statistical Linear Models, Fall 2017

I. Introduction

The Honnold-Mudd Library at the Claremont Colleges has a significant amount of data concerning the usage of the library's electronic resources. This data is cumbersome, and the library does not have significant staff available to explore and understand this data. We are interested in helping the library better understand the usage of its resources to inform decisions concerning how it allocates and distributes its resources.

II. Problem Statement

Our analysis attempts to answer the following question:

What are the characteristics of a session of wireless internet use (for library resources) on the Claremont Colleges?

In exploring this question, we characterize: (1) the duration of a session, (2) the campus of the user, (3) the location of the IP address, (4) the domains visited during the session, and (5) the patron type of the user.

III. Data Description

Our analysis looks at the following three types of data sets:

Wireless data: contains information on location, campus, time and device of wireless usage in the library for Fall 2016
Ezproxy data: contains detailed information on sessions of library resource usage for all of 2016, including duration, domain visited, campus information, and date-time information
Patron data: maps user ID to campus and patron type

A. Preprocessing and Cleaning

The data was provided to us from scripts written by Sam Kome, and was for the most part tidy. We had to correct a few issues with some of the fields and deal with faulty or missing values, most of which resulted from strange or incorrect behavior in the data collection system itself. The main preprocessing steps were:

1. Read in separate csv/tsv files and concatenate them into individual data frames for analysis.
2. Recreate session duration column, accounting for overnight sessions that start on one day and end on the next.
3. Threshold session duration to exclude sessions less than ~2 hours or greater than 24 hours. This step is discussed further in the analysis section.
4. Join separate tables into one (joining them by session) in order to explore how session relates to other features such as patron, campus, and online resources.

A. Features of Interest

After joining the ezproxy data, we have the following features:

session: unique session ID
scheme: the transport protocol, e.g. 'http'
subdomain: the host, e.g. 'www'
domain: the name of the service used, e.g. 'lexisnexis'
tld: the final portion of a Uniform Resource Locator, e.g. 'com'
path: file path to the requested resource
query: the requested resource or search query
fragment: miscellaneous leftover information
uuid: a unique hashed user ID
campus: 7C campus abbreviation
ipSubnet: the third field of the IP address
session_start: datetime for when the session began
session_end: datetime for when the session ended
session_duration_mins: duration in minutes of the session

After preprocessing the wireless data and keeping only columns of interest, we have the following features:

start_datetime: datetime for when the session began
stop_datetime: datetime for when the session ended
disconnected_at_closing: boolean representing if datapoint disconnected at normal closing hours for that date
connected_at_opening: boolean representing if datapoint connected at normal opening hours for that date

The patron data contains the following features:

uuid: a unique hashed user ID
campus: 7C campus abbreviation
patronType: user category (student, faculty, etc.)

IV. Analysis

In this report, we perform a question-driven analysis of wireless sessions and resource usage, posing questions relevant to characterizing how and when wireless resources are used and then using visualizations and statistical tests to answer those questions.

A. Session Duration

Session duration is automatically defined by the wireless internet system of the Claremont Colleges. Traditionally, wireless sessions represent continuous use without interruptions of a certain duration (e.g. the amount of time spent online without an interruption greater than 30 minutes). In this section, we explore what sessions look like in this dataset and how sessions can help us understand resource usage at the library.

Question 1: Given that they're automatically defined, how are sessions defined and do they have a sensible definition?

We can get a sense of how the wireless internet system defines a session by plotting all values of session duration. Figure 1 shows a plot of all values of session duration in the data in the order provided (ignoring session durations greater than 24 hours). Note that the data is provided in chronological order by default, thus the x-axis represents connections indexed from January 1, 2016 to December 31, 2016. The missing block at the beginning corresponds to January and February, where the wireless internet system did not provide session end times. The absence of any sessions between 0 and ~2 hours indicates that most likely the system considers two hours to be the minimum duration of a session. We suggest that this apparent minimum at two hours is something that the library should look into.

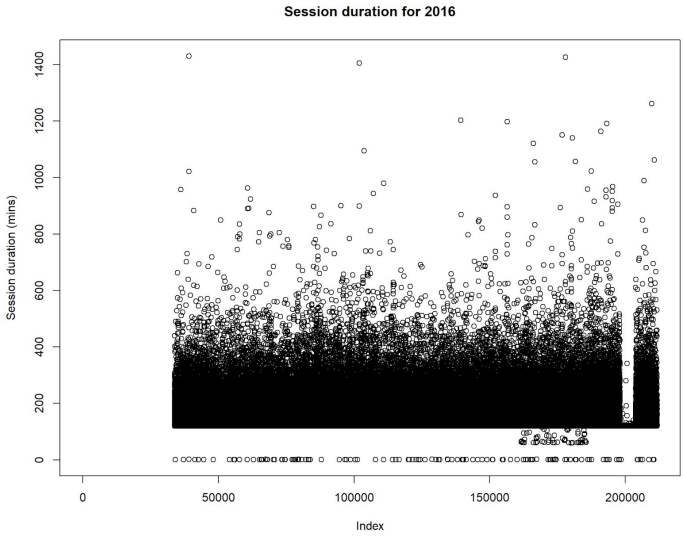

Figure 1: Scatter plot of session duration prior to removing NA and other nonsensical values.

We can also look at the distribution of session durations. Figure 2 shows the net distribution of session durations, corroborating the scatterplot in Figure 1 and showing the lack of values between zero and two hours, as well as where the bulk of the session durations are.

Figure 3 shows the same plot as Figure 1, after removing NA values and removing sessions less than 2 hours and 2 minutes. We now see a much more sensible distribution of session durations. In our analysis throughout the rest of the report, we consider session duration after performing this same thresholding.

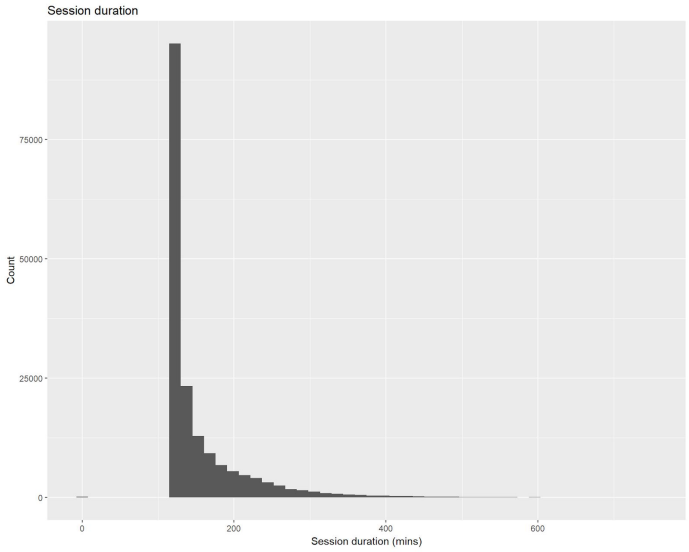

Figure 2: Histogram of session duration prior to removing NA and other nonsensical values.

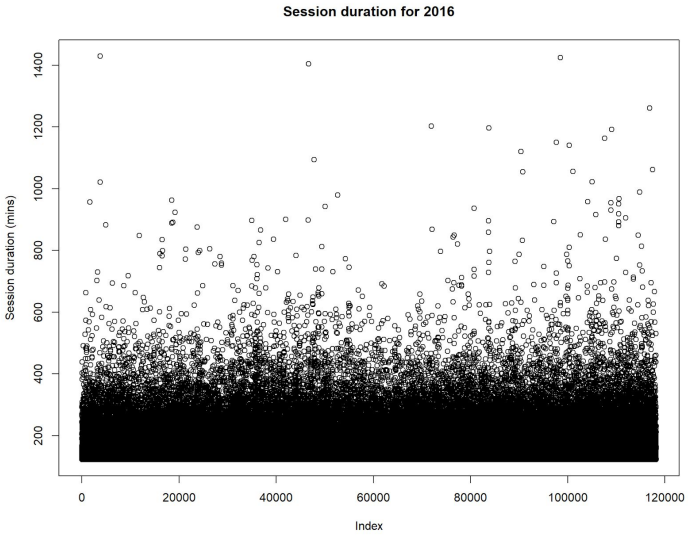

Figure 3: Scatter plot of session duration after removing NA values, and durations greater than a day or less than 2 hours and 2 minutes.

Question 2: When do people connect and for how long (which days of the week, times of day)?

To explore when people connect and how long they connect for, we can look at mean session duration across days of the week and times of day. Figure 4 shows mean session duration for each day of the week. We see that the longest sessions occur on Saturdays, with mean duration of about 2.5 hours. We also see that the day with the shortest mean session duration is Thursday, though mean duration is similar across all days of the week.

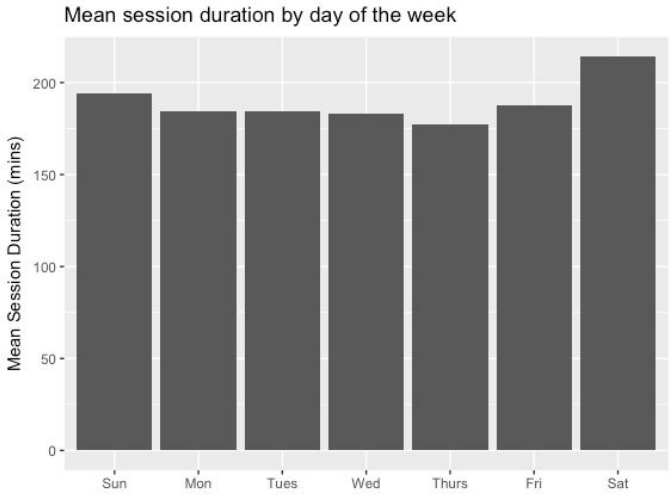

Figure 4: Bar chart of mean session duration for each day of the week. We see that on average the longest sessions occur on Saturdays.

We can also look at mean session duration by time of day. Figure 5 shows mean session duration for each hour of the day (where we take the starting time of the session as the hour). This graph indicates that the longest sessions start around 10:00 AM, 3:00 PM, and 9:00 PM. These appear to be the modes for when long sessions begin at the library.

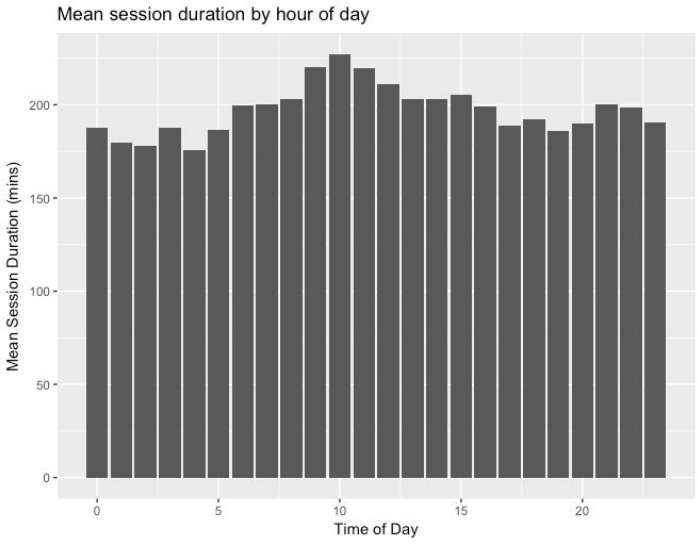

Figure 5: Bar chart of mean session duration for each hour of the day.

Question 3: When are more people staying until closing time or arriving right at opening time?

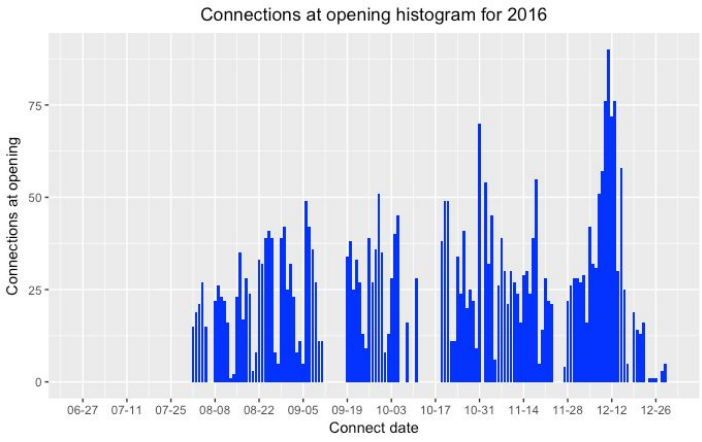

Figure 6: Histogram of the connections at opening over Fall 2016.

Figure 6 shows that the number of connections at opening remains relatively constant throughout, with a peak around finals. The connections at the start of the semester are relatively small compared to the rest of the data, and towards the end of the semester the histogram becomes more dense. There are breaks in mid-October and towards the end of November for Fall Break and Thanksgiving Break respectively. There is also an unknown break in mid-September. Also, comparing to disconnections at closing in Figure 7, there aren't overall many connections at opening for the semester, with a mean of 27.6 and median of 27.

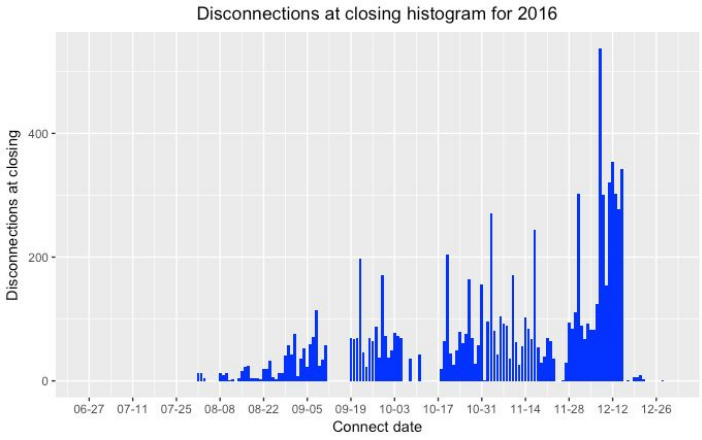

Figure 7: Histogram of the disconnections at closing over Fall 2016.

Figure 7 displays the same breaks as Figure 6. Unlike Figure 6, Figure 7 shows that the number of disconnections at closing increases throughout the semester. As finals week begins, the number of disconnections at closing seems to increase exponentially. Not surprisingly, this indicates that longer hours during finals might be useful to students. In addition, the number of disconnections at closing throughout the semester is relatively high compared to the number of connections at opening. There is a mean of 90 and median of 67 disconnections at closing. This suggests that the library may consider opening and closing later, because there are significantly more disconnections at closing than connections at opening.

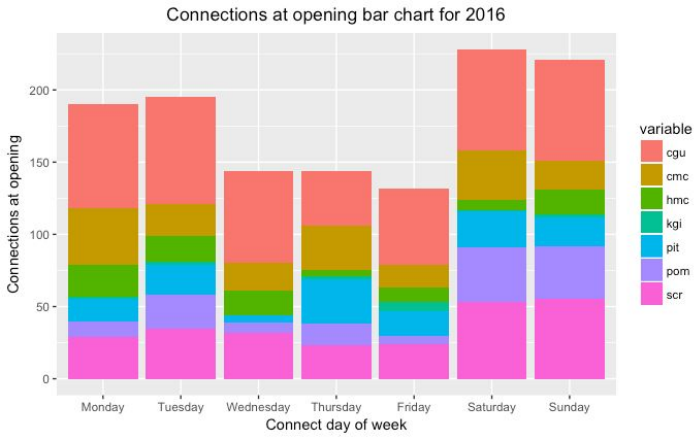

Figure 8: Stacked bar chart of the sum of the connections at opening for each day of the week. Sum for each campus makes up stack for each bar.

Figure 8 shows how the number of connections at opening varies over the week. The weekends have the highest amount of connections overall, and students are least likely to connect at opening on Wednesday, Thursday, and Friday. Also, a much higher proportion of Scripps and Pomona students visit the library on the weekends.

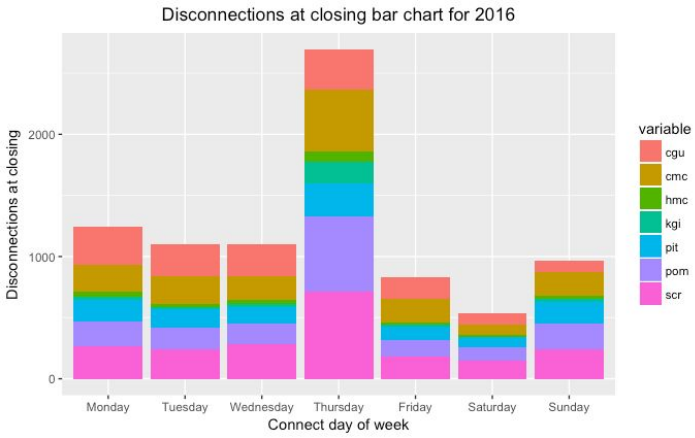

Figure 9: Stacked bar chart of the sum of the disconnections at closing for each day of the week. Sum for each campus makes up stack for each bar.

Thursday has by far the most disconnections at closing. In fact, the total number of disconnections at closing from Scripps on Thursday exceeds the number of total disconnections at closing on Saturday. For Scripps, Claremont-Mckenna, and Pomona, the number of students disconnecting at closing on Thursday is far more than on any other day. For example, the mean of Pomona students disconnecting at closing excluding Thursday is 167 with a standard deviation of 36. However, when we include Thursday, these numbers change to 232 and 173 respectively. We do not have a good explanation for this peak on Thursday, and suggest that this is something that library look into.

B. Sessions by Campus

We can explore wireless sessions by each campus individually to see how different campuses on the 7Cs interact with the library's wireless resources.

Question 1: Which campuses have the longest sessions?

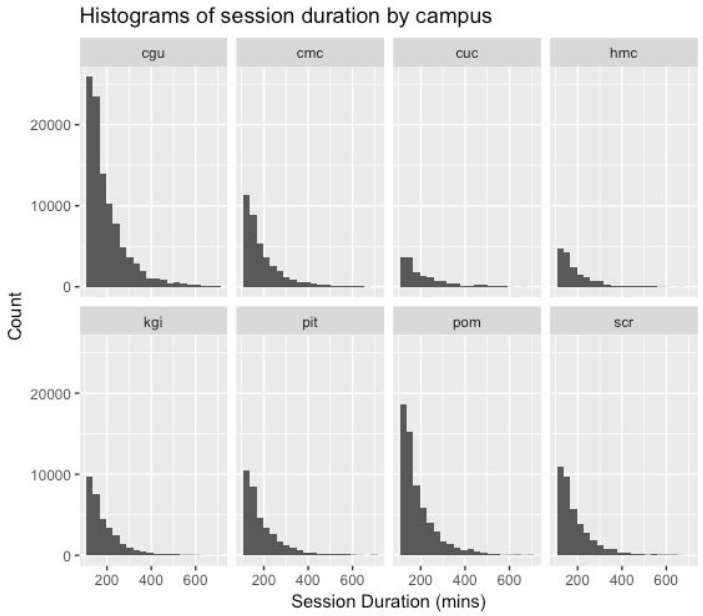

Figure 10: Histogram of session durations for each campus. The histogram for Keck Science Department was removed given its small number of observations.

Figure 10 shows the distribution of session duration by each campus. We can see counts for mean session durations as well as the relative number of sessions for each campus by the size of each histogram. CGU dominates in terms of number of sessions and the CGU distribution has the longest tail showing the highest number of long sessions, followed closely by CMC and Pomona. It is important to note the relative population of each campus. The percentage of the total population that each campus is responsible for is as follows:

Campus	Percentage of Population
CGU	14.3%
CMC	18.4%
HMC	11.5%
KGI	6.1%
PIT	13.0%
POM	24.2%
SCR	13.4%

Given the small size of CGU relative to Pomona for example, CGU's higher overall usage is even more impressive. To explore which of the differences in mean session duration are statistically significant, we can look at

Tukey Honest Significant Difference (HSD) confidence intervals. Those that do not contain zero indicate a significant difference (and provide an estimate of that difference). Figure 11 displays these intervals, showing the difference in mean session duration for each pair of campuses.

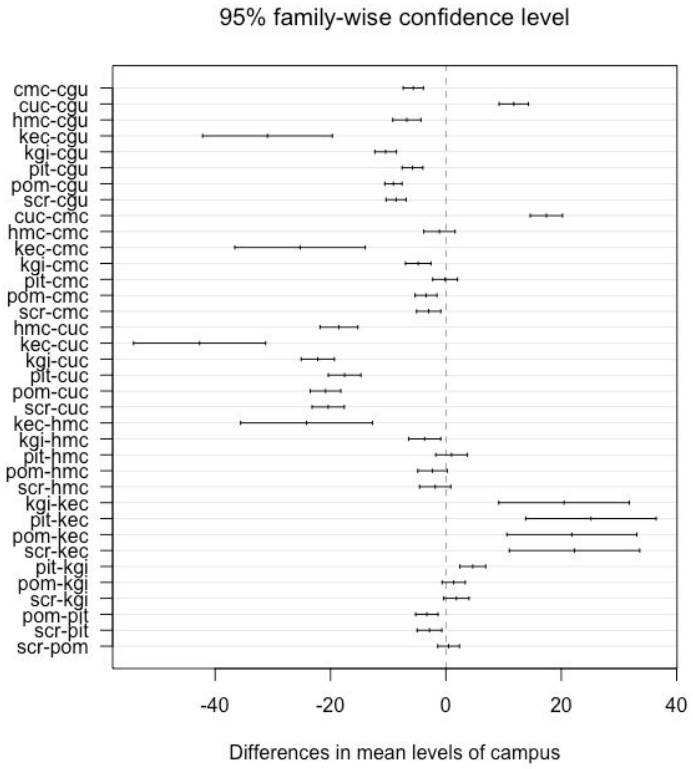

Figure 11: Tukey HSD confidence intervals for mean session duration by campus.

Question 2: Which campuses have the highest number of unique sessions? How does this change throughout the year?

Figure 12 shows how the number of sessions varies by campus across the year. As seen in other figures, CGU has by far the greatest number of sessions across the entire year. Looking across the year we can see some interesting trends. For the undergraduate colleges, there are generally more sessions in the Fall semester than in the Spring semester. However, this is opposite for Mudd. Pomona, Pitzer, Scripps, CMC, and KGI all have their peaks in Fall semester, but Mudd has the most sessions in March and April. There are a couple of other expected patterns to note.

There is generally less usage over the Summer months, and the highest usage months is December (note that students are only in school for the first half of December). However, there are a couple of exceptions to these patterns. Pomona seems to have a less significant drop-off during the Summer. Additionally, CUC has a much larger jump in usage in December than the other schools.

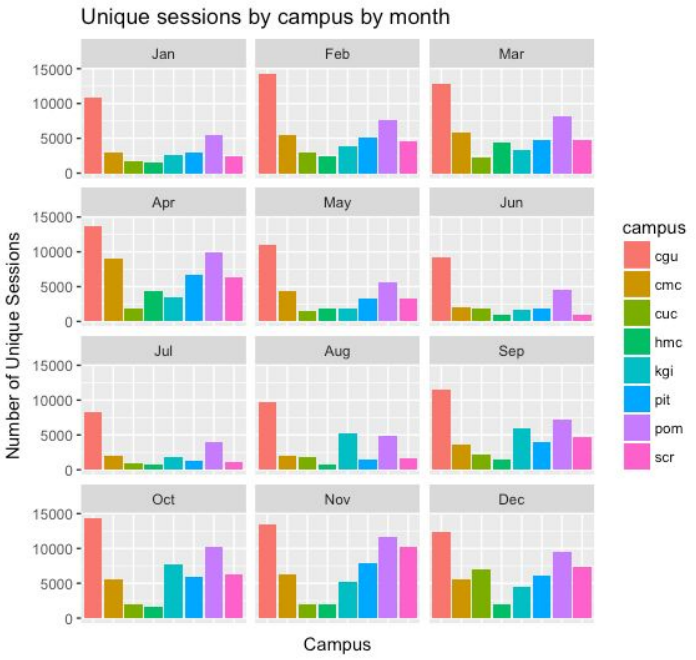

Figure 12: Number of unique sessions each month by campus.

C. Sessions by IP Address

We can also analyze session by the IP address associated with them. This IP address tells us the location of the session (i.e. whether it is on-campus, off-campus, or VPN).

Question 1: Do different campuses have different usage patterns for on-campus, off-campus, and vpn access?

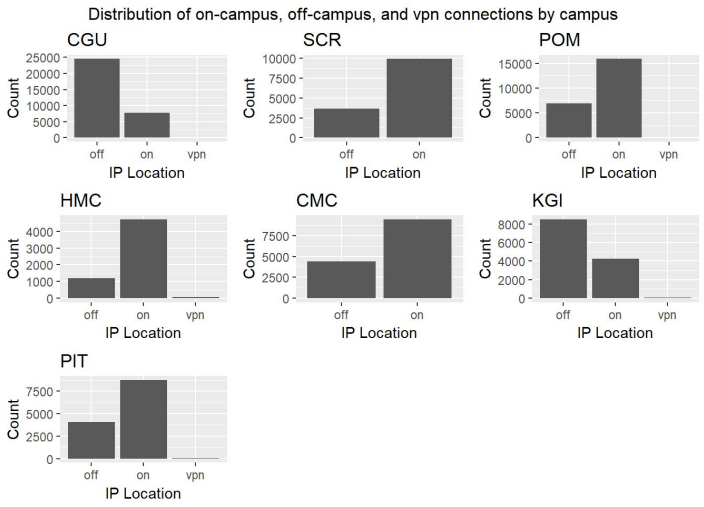

Figure 13: Bar charts showing proportion of on-campus, off-campus and VPN sessions for each campus.

We can see in Figure 13 that different campuses have different proportions of on-campus, off-campus, and VPN sessions. Not surprisingly, KGI and CGU have much higher off-campus usage. Interestingly, there are no VPN sessions from CMC or Scripps, and Mudd has a significant number of VPN sessions, despite its small size.

Question 2: Are different domains used different amounts at different IP locations?

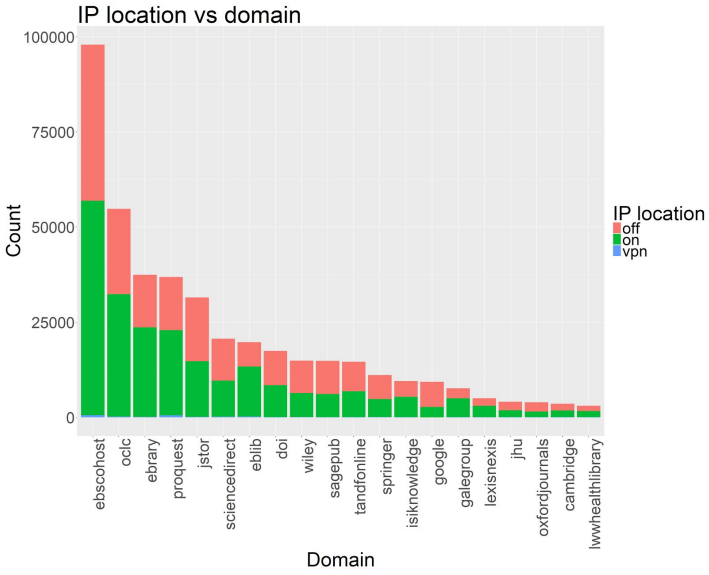

Figure 14: Stacked bar chart showing the top 20 domains and the proportion of their accesses that were on-campus, off-campus, or VPN.

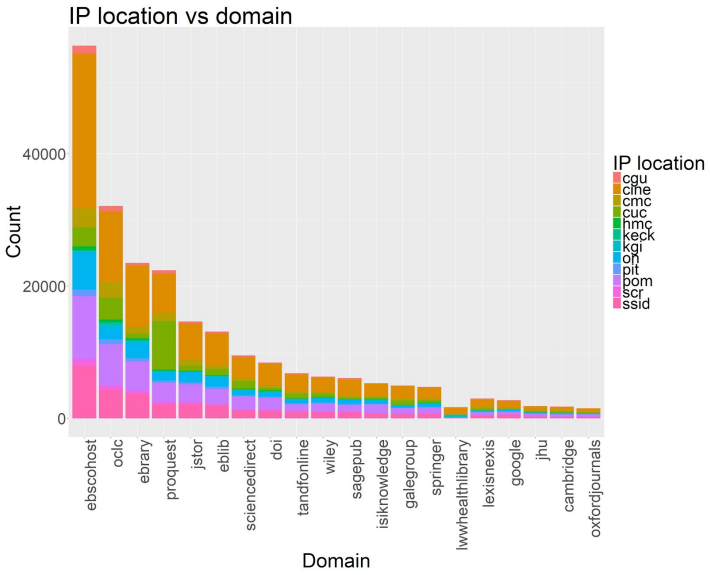

Figure 15: Stacked bar chart showing the top 20 domains and the proportion of their accesses that were in each IP address range.

In Figure 14 we can see the on and off campus usage of the top domains, and in Figure 15 we can see the usage of the top domains by different on campus IP locations. Although for many of the most popular domains the distribution looks about proportional to each school's overall usage, there is one domain that stand out. *Proquest* is used much more frequently by CUC IP addresses in comparison to other domains. In future analysis, it may also be informative to look at the top domains for each IP location.

Question 3: Are different campuses more or less likely to have sessions from IP addresses on their home campus?

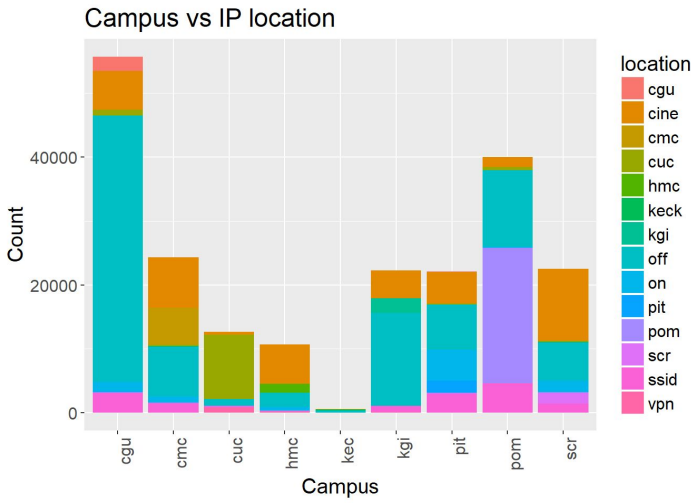

Figure 16: Stacked bar chart showing the the proportion of sessions in each IP address range for each campus.

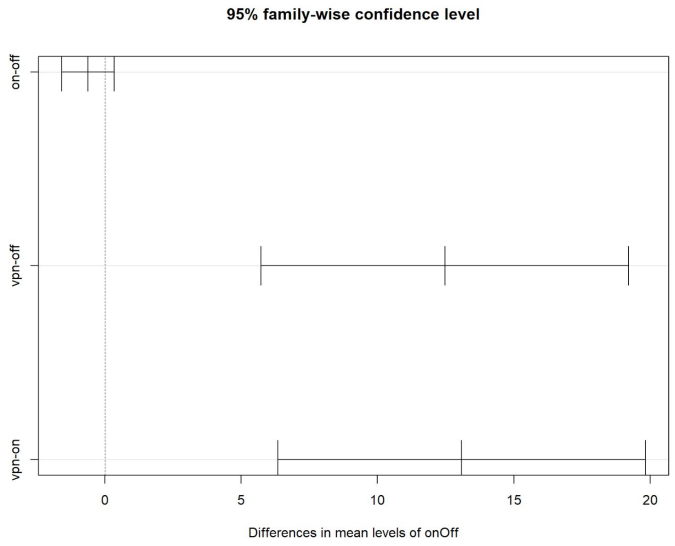

Figure 17: Tukey HSD confidence intervals for mean session duration by on-campus, off-campus, or VPN IP addresses.

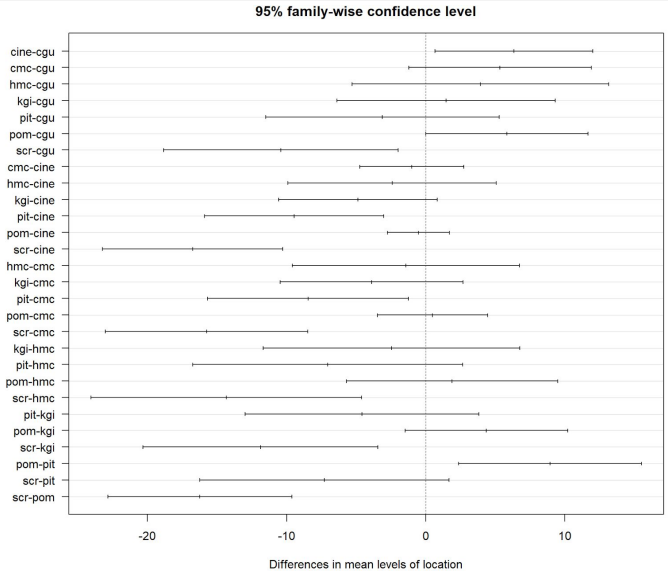

Figure 18: Tukey HSD confidence intervals for mean session duration by different IP address locations.

Figure 16 shows the number of sessions from each campus, broken down by the actual location of the IP address used for the session. We see that CGU and KGI are much more likely to have sessions at an off-campus IP address. For each campus, we can also look at the proportion of sessions that are from the home campus. We see that CUC and Pomona have high proportions of accesses from their home campuses. CMC also follows this pattern to a smaller extent. However, HMC, Pitzer, and Scripps actually have quite small proportions from their respective campuses IP addresses, and are more likely to be starting sessions from CINE IP addresses.

Question 4: Do sessions in different locations have different durations?

Figures 17 and 18 show confidence intervals for the difference in mean duration between different IP address locations.

We can see a strong difference in mean duration between VPN sessions and non-VPN sessions. On average, VPN sessions last about 13 minutes longer than non-VPN sessions. There is not a significant difference between the duration of on-campus sessions, and the duration of off-campus sessions. If we look more closely at session duration of on-campus sessions, we can see some significant differences. Scripps tends to have sessions that are about 15 minutes shorter than sessions from other IP

address locations. CINE tends to have slightly longer sessions than sessions from other IP addresses. There may be other insights to be gleaned from Figure 18.

D. Resource Usage

To better understand what domains are visited and for how many times/how long they are visited, we ask the following questions regarding domain usage and answer them with plots and analysis.

Question 1: On average, how many domains does each session visit? How does this vary by campus, IP location, and patron?

There are a total of 439 unique domains visited in the given data, with a minimum of 1 domain per session and a maximum of 20 per session. On average, 1.585 domains are visited in one session.

Looking at Figure 19, we first show a histogram of the distribution of number of domains visited per session across all connections.

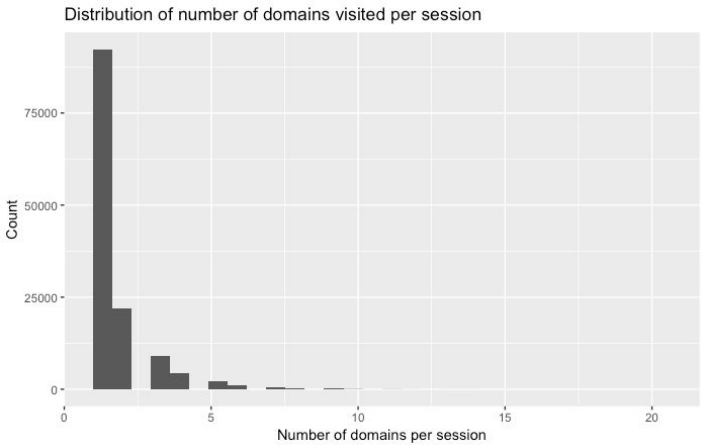

Figure 19: Histogram for number of domains visited per session.

When each campus is looked at separately, we have the following summary (Figure 20).

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
ALL	1.000	1.000	1.000	1.585	2.000	20.000
POM	1.000	1.000	1.000	1.558	2.000	14.000
CMC	1.000	1.000	1.000	1.528	2.000	13.000
HMC	1.000	1.000	1.000	1.467	2.000	13.000
PIT	1.000	1.000	1.000	1.584	2.000	12.000
SRC	1.000	1.000	1.000	1.695	2.000	15.000
CGU	1.000	1.000	1.000	1.665	2.000	15.000
CUC	1.000	1.000	1.000	1.432	2.000	18.000
KEC	1.000	1.000	1.000	1.453	2.000	6.000
KGI	1.000	1.000	1.000	1.546	2.000	20.000

Figure 20: Table of summaries of number of domains visited per session for each campus.

Here we see the majority of sessions only visit 1 domain, making the minimum value, first quartile, median and third quartile not very interesting. Thus we also provide a histogram for each campus to visualize the distributions.

We can see that all distributions are skewed, with KGI having a longer tail, and CUC having a relatively higher peak.

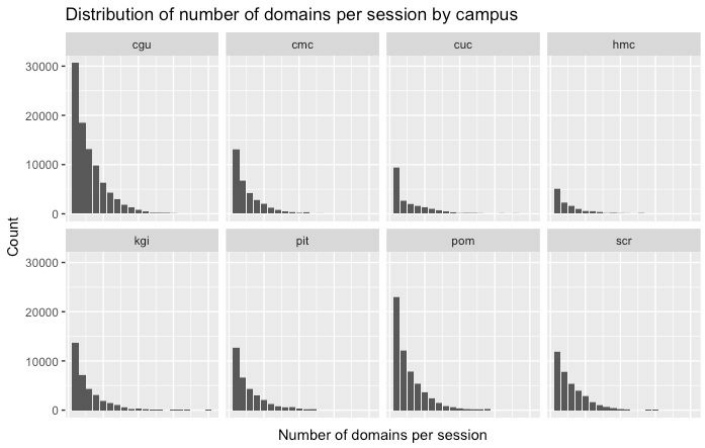

Figure 21: Histograms of the number of domains visited per session for each campus. The histogram for Keck Science Department was removed given its small number of observations.

To see whether the number of domains visited per session is different for different campuses, we compare Tukey HSD confidence intervals on the number of domains visited versus campus. The result is shown in Figure 22.

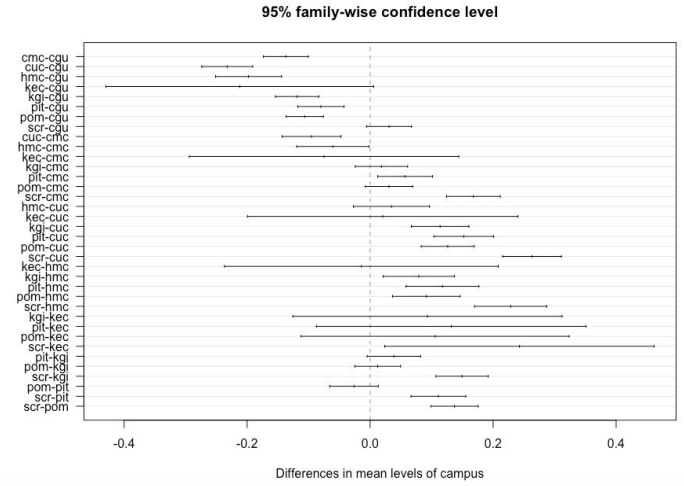

Figure 22: Tukey HSD confidence intervals for mean number of domains visited by different campuses.

We can see that KEC is different from all other campuses. Besides that, SCR-CGU, KGI-CMC, POM-CMC, HMC-CUC, POM-KGI and POM-PIT also seem to visit significantly different numbers of domains per session.

Question 2: Which domains are visited most frequently/for the longest duration? How does campus influence this?

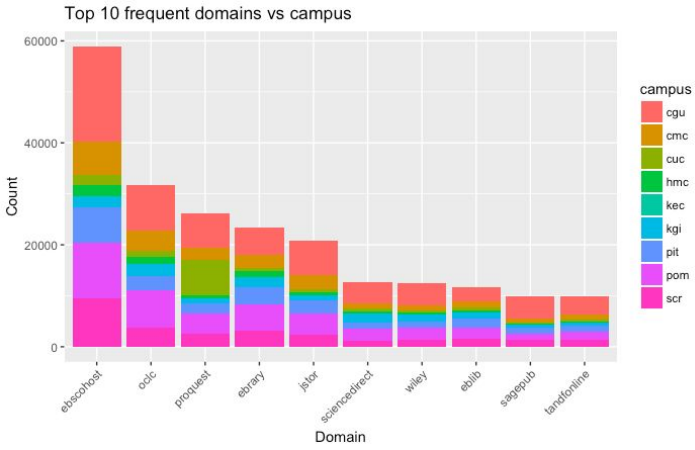

Figure 23: Stacked bar chart showing the proportion of sessions for each campus for the top 10 most frequently visited domains.

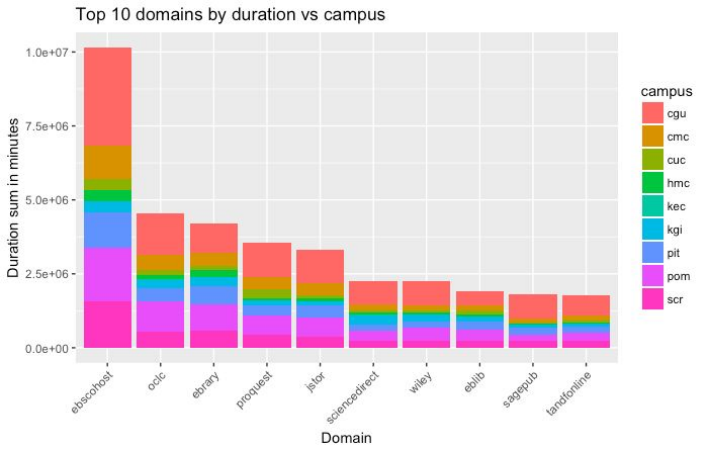

Figure 24: Stacked bar chart showing the the proportion of total duration of visit for each campus for the top 10 domains with respect to duration.

We can see that the top 10 domains by number of sessions and by total duration are exactly the same, with *ebscohost* and *oclc* being the two domains visited the most times and for the longest duration. Moreover, the proportion of number of visits for each domain and each campus and that of the total duration look very similar except for *Proquest*, for which CUC has a relatively high number of visits but short total duration.

To explore the relationship between total duration and number of sessions visited for each domain, we build a linear model with the total duration on each domain as the response and the number of sessions that visited the domain as the predictor.

```
Call:
lm(formula = duration ~ visits, data = dur_vsts)

Residuals:
    Min      1Q  Median      3Q     Max
-161818    -482    -109      46   88824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.4637   722.1380   0.062    0.951
visits      198.8722     0.2126 935.244   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14630 on 422 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 8.747e+05 on 1 and 422 DF,  p-value: < 2.2e-16
```

We can see that number of visits is indeed significant and we have a large r-squared, which might help explain the similarity in the previous two graphs.

Question 3: Which domains are most popular for each campus?

We found the top 5 popular domains for each campus. All campuses have the same top 5 domains. Figure 25 shows the proportion of each of the 5 domains for all campuses.

We can see that most campuses visit *ebscohost* the most, but CUC visits *proquest* the most.

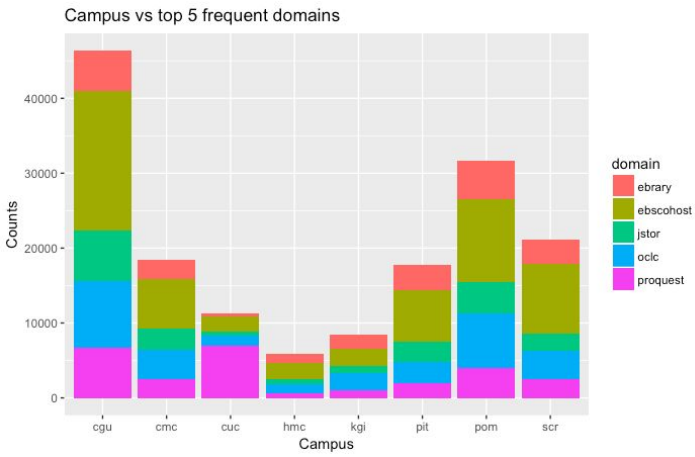

Figure 25: Stacked bar chart showing the the proportion of number of visits for each of the top 5 domains for each campus. The bar for Keck Science Department was removed given its small number of observations.

E. Patron Types

Another interesting feature in the given library data is the patron type of the sessions. To provide a clearer analysis, we re-categorized the patron types into 8 categories instead of the original 16 categories. We use the following 8 categories: current faculty and staff, retired faculty and staff, undergraduate, graduate student, seniors, visiting scholar, alumni, and community and family.

Question 1: How does session duration compare across patron types?

We first examine the differences in average session durations for different patron types. Figure 26 shows that the community and family category has the shortest duration, while visiting scholars has the longest. There doesn't seem to be a significant difference between the average session duration for graduate students and undergraduate students.

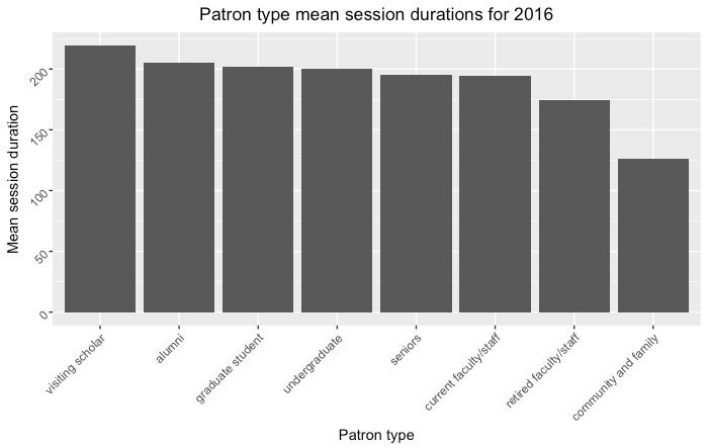

Figure 26: Bar graph showing the mean session durations for 2016 for all patron types.

Question 2: Which domains are the most popular for each patron type?

In Figure 27 and Figure 28, we show the top 10 domains by number of visits and duration, and the proportion of each campus for each domain.

Observe that the top 10 domains by number of sessions and by duration are the same, but the order is different for some: *proquest* has more visits than *ebrary*, but in terms of duration, it is not as long. We notice that current faculty/staff seem to constitute a lot of the visits with relatively short duration, which is reasonable given that CUC was seems to have a lot of short-duration visits.

Wiley and *Sciencedirect* also have reverse order in Figure 27 and Figure 28. However, the difference is relatively minimal compared to the difference between *ebrary* and *proquest*.

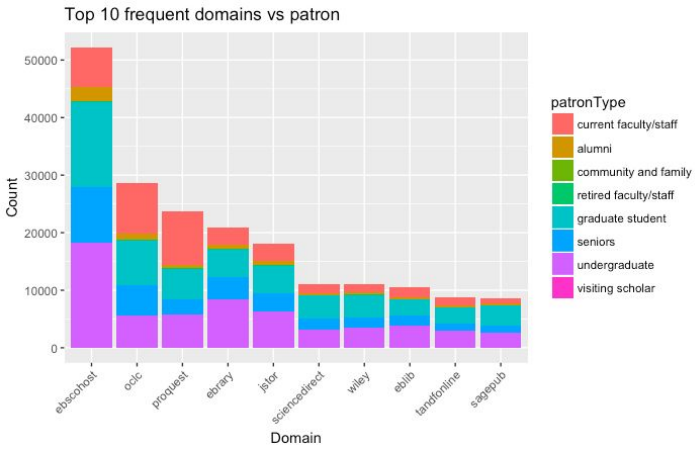

Figure 27: Stacked bar chart showing the proportion of sessions for each patron type for the top 10 most frequently visited domains.

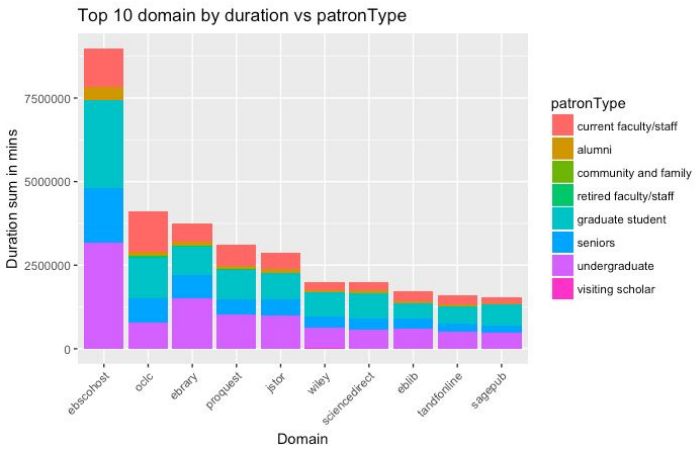

Figure 28: Stacked bar chart showing the proportion of sessions for each patron type for the top 10 visited domains by duration.

For the most frequent domains by patron type, we exclude retired faculty/staff, community and family, and visiting scholar because of the small number of observations. The 5 most visited domains are *ebscohost, oclc, proquest, ebrary* and *jstor* for all patron types except for alumni, which has *ebscohost, isiknowledge, oclc, ebrary* and *jstor* as the 5 most visited domains.

In Figure 29, we show the proportion of domains for each patron type for domains *ebscohost, oclc, proquest, ebrary*, *jstor* and *isiknowledge*. We see that current faculty/staff has a relatively large proportion of visits to *oclc* and *proquest* compared to other patron types, and undergraduate has a relatively large proportion of visits that go to *ebrary*.

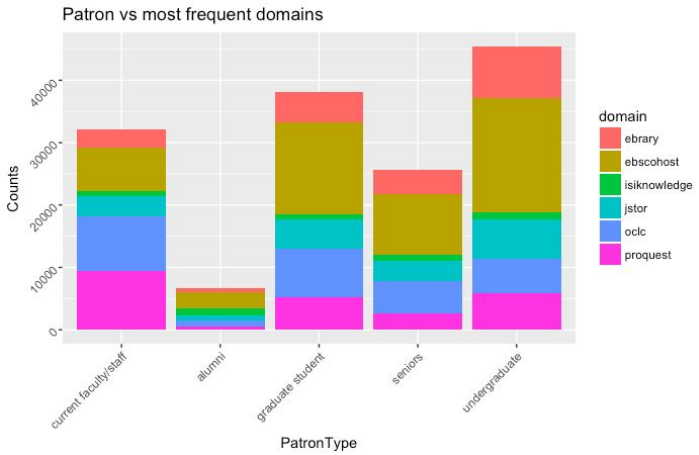

Figure 29: Stacked bar chart showing proportion of most visited domains for each patron type. Bars for retired faculty/staff, community and family, and visiting scholar are removed given the small number of observations.

V. Suggestions

While there are many potential insights to be gained from the graphical and statistical analysis above, we would like to highlight a few suggestions for the library staff to either implement or look into.

We noticed in the data an apparent minimum session duration of two hours. We suggest that the library staff look into what is causing this, as it potentially obscures a lot of information in the data. Additionally, there was a reporting error in the wireless data in mid-September (Sep 11 - Sep 18), causing a gap in the data.

Since there are more disconnections at closing than connections at opening, we suggest emphasizing closing later, rather than opening earlier, particularly on Thursday. Additionally, the data supports the library's current 24/7 approach to finals week.

In looking at usage by campus, the data continues to show that CGU uses library resources significantly more than other schools, particularly when also considering the relative population size of the campuses. There are a few trends across the year in Figure 12 that seem worth looking into, particularly where different campuses have their usage peaks.

In the IP location analysis section, we found that VPN sessions tend to be longer, and the different campuses seem to use VPN different amounts. Additionally, in Figure 16, we see that different campuses are more likely to have sessions originating from their home campus.

We found that these domains are the 10 most used across all categories: *ebscohost, oclc, proquest, ebrary, jstor, sciencedirect, wiley, eblib, tandfonline,* and *sagepub*. We also note that the top 5 domains used are the same across all patron type categories, except for alumni, who visit *isiknowledge* frequently.

VI. Appendix

The code used to preprocess the data and generate the figures used in the report can be found at the following Github repository: https://github.com/lbynum/statlin-library-project.