Columbia Density visualization and analysis

Erin McMahon [*], **Dingzeyu Li** [*] , **William VanArsdall** [†]

[*] Computer Science, and [†] Applied Mathematics, Columbia University

Submitted to Modeling Social Data as final project report

Locating available study space at a university is very challenging, especially during midterm and final periods. Recently, Columbia University released access to some of the Wi-Fi usage data in its libraries and other study spaces, which led to the development of *Density* - an app that estimates the current percentage capacity of a study space by using the number of Wi-Fi devices currently connected. Our project, inspired by *Density*, analyzes the effectiveness of this metric, demonstrates the potential for better visualization, and proposes ways to improve its accuracy.

density | visualization | time series analysis

Abbreviations: CUIT, Columbia University Information Technology;

Introduction

As any academic knows, finding study spaces on any college campus, especially as finals approach, is a challenge in itself. Fortunately, in 2014, Columbia's IT Department (CUIT) released a real-time data application [1] with data on Wi-Fi usage for various libraries across campus. The Density application shows a basic visualizations and provides an open-source API for Wi-Fi usage patterns [2].

Motivation. In this project, we have utilized the *Density* API to analyze and visualize the real-time study space usage. We are interested in this topic because 1) it is relevant to our everyday lives; 2) this new data could lead to interesting applications; 3) and current tools do not fully exploit the data.

We also realize that the Density API is not a perfect predictor of library density. Figure 3 shows that during the month of September, the number of devices present *before* the opening of the library varied dramatically. There are many possible causes, such as different number of librarians, nearby classrooms, and so on. In addition, the Wi-Fi usage might not fully reflect the population in the study space – some people might have zero or more than one devices connected to Wi-Fi. In other words, we need to evaluate the efficacy of current metric.

Goal. Our goal in the project is the following:

- analyze the density data
- evaluate the efficacy of the app and its assumptions
- visualize the data in a more effective way
- investigate possible ways to improve the estimation

Data Acquisition

We collected data from two sources, the density website [2] and Columbia libraries entry statistics. For the detailed density website API documentation, please refer to the official website [3]. Here we will focus on the API documentation that is relevant. As Columbia students, we can request an API authentication key, with which we can send requests. In the database, there is usage data for 22 locations, grouped into 8 buildings; for example, group Butler has 6 locations, including Butler 2, Butler 3, and so on.

As for the library data analysis, all Columbia libraries record only entry swipes, but not exit swipes. Therefore we need to estimate the current population without data on the numbers of exits in the corresponding time frame. Similar research work like [4] tackles New York City MTA Subway estimation. Like our libraries, subway system only logs when people enter. Their system use a number of external data sources, including trip transition data, bus route analysis, and household travel surveying. Since we don't have that many resources available, we propose a simple model to estimate the exit amount, which we will explain in the Analysis section.

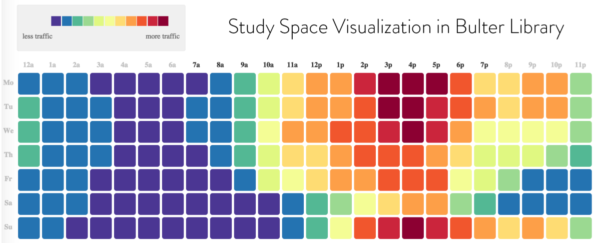

Fig. 1. We visualize the space utilization percent arranged by the hour for 7 days in a week, aggregated over a semester-long period. We can see a clear concentration in the afternoon and a decline on Friday/Saturday evening.

Data Cleaning from JSON API. To obtain the json response, we first determine the location or group that we are interested in. Then we specify the range of time to request. There are several ways to specify the time; we can specify the start date and end date and get the response in either 15-minute time frames or aggregated daily estimates. In the returned json file, there are two entries that are most relevant, namely “client counts” and ”percent full”. The former simply registers the total number of connected clients at a certain point whereas the latter further divides it by a maximum capacity. Maximum capacity is calculated per library as 95% of the maximum number of connected devices observed in the past year

Reserved for Publication Footnotes

(ignoring outliers)[1]. In this project we will primarily use the reported ”client counts” while we aim to improve the measure of maximum capacity.

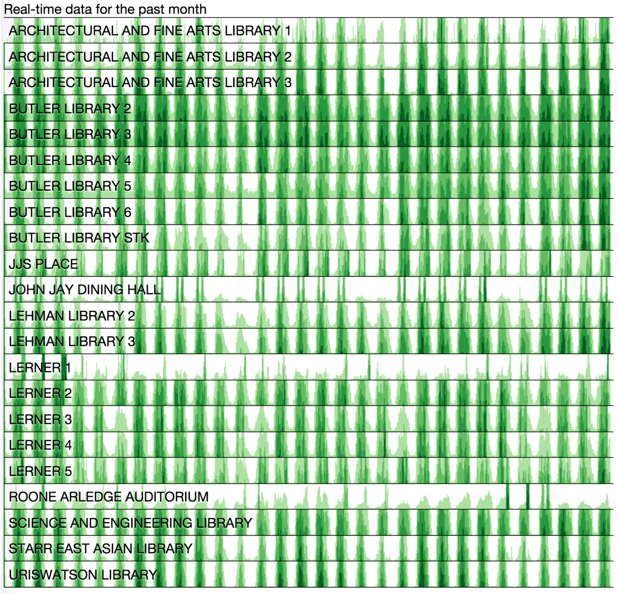

Fig. 2. Cubism time series visualization.

Data Cleaning from Libraries. In addition to routers, CUIT has a comprehensive database of all swipes into various libraries across campus. Unlike the *Density* data, this number is not decomposed by specific room or floor. We aggregate the entry data into 15-minutes blocks of unique ID cards swipes. (For example, if a student entered a library twice in a 15-minute period, the student is only counted once.) Although we might have looked at a running time frame (15-minutes from when the user entered) or a longer-time period (one-hour window), we believe the 15-minute blocks well represent ‘real’ library entries.

Visualization

Data visualization is a key component of this project. Students, and the public in general, are more receptive to data if it is in a visual and digestible format. Therefore, we created an interactive website so that students can investigate trends, see usage in real-time, and create other application add-ons that use our data.

Hourly Heatmap. Over the course of a semester, we compute the average percentage full for each libraries and display those values on an interactive heatmap. The grids are divided by hour and day of the week. Figure 1 shows one sample plot of Butler library. Through this straightforward visualization, we observe the apparent peak around 3pm every day and a significant drop in the evening on Friday and Saturday. On our website, users can specify building through an interactive map, choose the semester, and even get plots based on aggregated time or time for each day.

Cubism Time Series. Cubism is a JavaScript library that allows for effective visualization of real-time data. Figure 2 shows a screenshot of the real-time browsing system. (If the plots are too small, please visit our website for the interactive version.) The x-axis is the time variable, sampling at every hour; the y-axis is the number of clients, darker means more connections. All the libraries are stacked together. From this plot, we observe several interesting things. First, dark areas concentrate during weekdays and fade out quickly over the weekends. Second, in study places without dinner halls, there are two peaks in a day, one after lunch, one after dinner. The meal time also corresponds to John Jay dinner hall’s distribution. Third, in libraries with dinning halls, there is only one peak in a day, because students don’t have to leave the building for food.

Analysis

There are two points of contention in the current API model. First, devices do not necessarily equal students. Second, maximum capacity estimates might be high because they are based on 95% of the maximum number of clients ever connected.

Connected Devices as a Indicator of Students. A number of basic observations call into question the effectiveness of ‘connected devices’ as an estimator of the actual library population. For example, Figure 3 displays the average device count for the month of September at 7:30 a.m. in Uris/Watson library. The library, however, opens at 8:00 a.m. Therefore as much as 10% of Uris’ reported 966 maximum capacity could come from this data that clearly does not reflect the fullness of the library that is empty in reality.

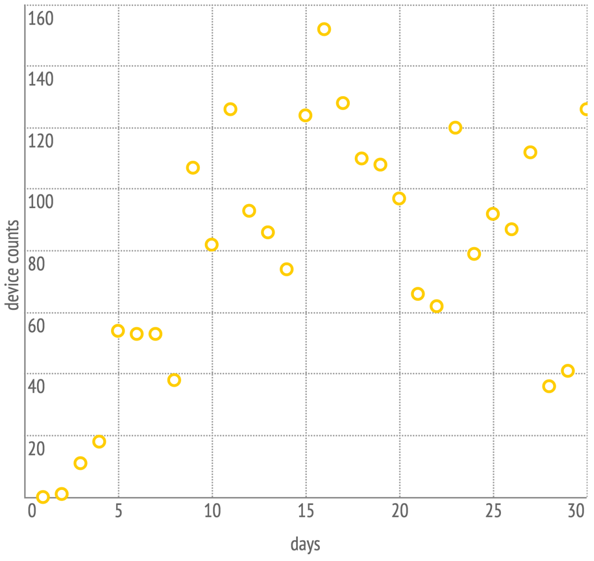

Fig. 3. Reported device counts in Uris/Watson Library at 7:30 a.m. for September 2014. The library opens at 8 a.m.

[1] the API documentations claims that there is a slightly more involved process for obtaining each maximum capacity value. However, this simple formula always replicates the given maximum capacity values with very high accuracy

Intuitively, a student might have zero or more than one device. We used actual library entry data as a means of comparison to see if we could get a better estimate of the student population in a library at a given time.

The library data provides us with a clean estimate of library swipes. However, the correlation between entries and 'seats' in a given library can be inaccurate for several reasons including students stepping outside of the library for a short time, missing exit data, and other factors. In addition, the ratio between 'seats' and entries might vary across libraries. In Uris, the bathroom is located outside the library; in contrast, a student could probably spend an entire week living in Butler without ever leaving the library. Even given these limitations, there is a fairly robust correlation between the library entry data and the *Density* data.

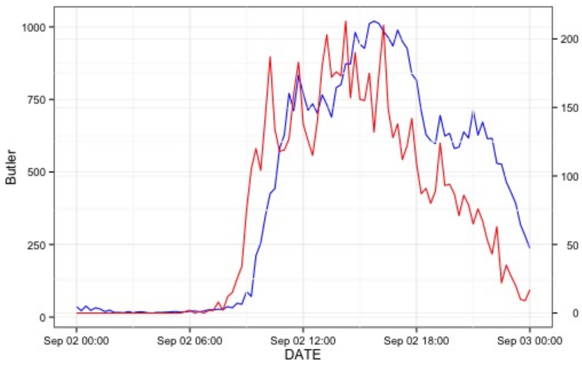

Fig. 4. Entry data shows a leading influence on Density data and high correlation between time series.

To derive an estimate of 'seats' from entries, we created a very simple model based on the library entry data. We assume that the number of exits depends *only* on when they entered the library. We tested eight models of exit distributions where each coefficient corresponded to a beta pdf estimate.

$$\mathrm{Seat(t)} = \sum_{i=0}^{t} \mathrm{Swipe}_i - \mathrm{Exit(i)}$$

$$\begin{aligned}\mathrm{Exit(i)} = {} & \beta_1 * \mathrm{Swipe}_{i-2} + \beta_2 * \mathrm{Swipe}_{i-4} + \beta_3 * \mathrm{Swipe}_{i-8} \\ & + \beta_4 * \mathrm{Swipe}_{i-12} + \beta_5 * \mathrm{Swipe}_{i-16} + \beta_6 * \mathrm{Swipe}_{i-20}\end{aligned}$$

β represents coefficients from six different beta models where the pdf varies from a left skew to a right skew. $\sum \beta = 1$
i represents the time period lags where $i - 1$ signifies 15 minutes before i.

We ran eight model variations against library and density data from 09-01-2014 to 12-31-2014. We used the maximum correlation between the entry model and the *Density* data to find the optimal model. Using the model with the highest correlation, we computed the ratio between Density 'seats' and the entry model 'seats' for each 15-minute time estimate and computed the mean ratio. The results are displayed below in the plot and table. The table represents the maximum correlation associated with the best-fit model. From this model, we generated the mean ratio of Density to Entry Model estimates.

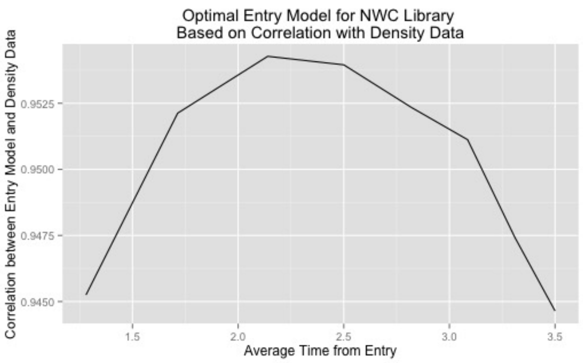

Fig. 5. The optimal model for NWC Library is model 3, which corresponds to an average exit time of 2.14 hours. This correlation yields a mean ratio of 0.524 as shown in Table 1.

Table 1. The maximum correlation between the model and Density data. The mean ratio represents our estimate of the number of Density clients to the number of 'seats' in the each library.

Library	Correlation	Mean Ratio
Butler	0.9402573	0.50
Uris	0.9738087	0.9856
NWC	0.9542753	0.524

The data from Uris suggests that the Density data is close to a 1-1 ratio of device per student while Butler and NWC suggest a ratio of 1 device per two students.

Maximum Capacity. The current method of obtaining the maximum capacity of each library is simply finding 95% of the maximum measured device count, excluding large outliers. By tailoring each maximum capacity value more specifically to each library, a more accurate value can be found. Using Uris as an example, the true 95th percentile value of the daily maximum (two standard deviations from the mean) is 878, while the maximum capacity given by the Density app is 993. Performing this analysis on the Northwest Corner and Butler Libraries, we obtain the following updated value for each library's maximum capacity:

Table 2. Maximum Capacity by Density and Our Method

Library	Density's	Ours
Butler	1789	1672
Uris	993	880.4
NWC	154	146.5

These new estimates only result in small changes to any given library's estimated fullness percentage; for example, if the app previously reported that Uris was 50% full, the app would now report that is instead 56%

Discussion

In many respects, the *Density* data is an accurate measure of current utilization of library study spaces. From our heat maps and cubism plots, we can see that library times spike after lunch and dinner. Graduate school libraries tend to have a 9-5 time frame while Butler shows more variation. However, although the trends seem accurate, we believe that the percentage numbers are flawed. First,we have used library entrance data to create a better estimate of the true number of students entering and leaving the library. Second, we created a new metric for maximum capacity for each library.

Devices per person. We use a simple model to approximate the Exit function. While the assumption that people will most likely leave in a few hours is reasonable, there are factors that we do not consider. Time of day might be reasonable to include as another parameter among others. It seems reasonable that devices per person would be lower than one since many students do not utilize devices while studying. In addition, there are many employees in each library who will not be using devices as well.

Maximum Capacity. We believe that the maximum capacity estimates used in Density's API do not accurately reflect the real capacity. An easy fix to this problem would be to calculate maximum capacity only using the difference in device count from opening time to the maximum. This would significantly change the maximum capacity metric for Uris, but would have no effect on libraries like Butler which do not typically close. This solution, however, is more of a quick fix than a better model. The fact that the device count can vary as much as Figure 3 (this particular data has a standard deviation of 40.8) during times with *no change* in study space density implies a more endemic problem with using device count as an estimate of the number of students in a library. Observations such as these motivated our search for a better means of calculating the capacity of each library.

Conclusion and Future Work

In this project, both data sources are incomplete in the sense that swipe data does not contain exit information and Wi-Fi data cannot reliably reflect the number of people. One straightforward approach is to add a vision-based tracking system to count the people in the library [5]. Installing surveillance might raise some privacy concerns as well as legal issues, but those are beyond the scope of our report.

Which maximum capacity metric is best? While the metrics proposed in this paper provide an improvement to the current model, the truly 'best' model would be an accurate real-time headcount of those in the library. In lieu of this lofty ideal, we believe that the average user's intuition on what it means for a library to be x% full would be a robust metric. Further improvements to the maximum capacity estimate could be made by appropriating a multi-armed bandit model using each maximum capacity estimate as a bandit. By first calculating the density estimates from the different maximum capacity metrics, and then polling students currently in each library on which one provides the most accurate estimate of the current density, we could determine which metric is most effective. This is

ACKNOWLEDGMENTS. We would like to thank Jake Hofman for preliminary discussion on the visualization and analysis; Hayrunnisa Bakkalbasi for providing us with the library entrance data; Trulia and Square for open-sourcing their visualization libraries.

1. Columbia ADI Labs. https://adicu.com/labs/ (2014)
2. Density by ADI Labs. http://density.adicu.com (2014)
3. Density API by ADI Labs. http://density.adicu.com/docs (2014)
4. Barry, James J., Robert Freimer, and Howard Slavin. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation Research Record: Journal of the Transportation Research Board 2112.1 (2009): 53-61.
5. Chan, Antoni B., Z-SJ Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Computer Vision and Pattern Recognition (CVPR) 2008. IEEE Conference on, pp. 1-7