Does Digital Engagement Predict Enrollment: An Analysis of Applicants' Behavior on Union College's Website"

By

Christopher Garibaldi

* * * * * * * * *

Submitted in the partial fulfillment
of the requirements for
Honors in the Department of Economics

Union College
June, 2015

Abstract

GARIBALDI, CHRISTOPHER Does Digital Engagement Predict Enrollment: An Analysis of Applicants' Behavior on Union College's Website Department of Economics June 2015

This paper examines the behavior of admitted regular decision applicants on Union College's website. I focus on the period between the announcement of admission decisions until after the deadline for deposits. I find that website traffic surges following the release of admission decisions, and falls sharply after the deadline for deposits. Traffic comes from applicants who end up enrolling at Union, as well as those who ultimately enroll elsewhere. The share of traffic from applicants who enroll elsewhere gradually declines from about 75 percent at the beginning of our sample to about 60 percent by the deposit deadline, declining to 50 percent in the week after the deadline. In terms of destinations, traffic flows to admissions information (visit campus, financial aid). There is also strong interest in the academic part of the website including information on majors and minors. Perhaps surprisingly, visits to a variety of offices including the registrar and student activities are fairly common throughout the period. Overall, the pattern suggests that applicants use the website extensively to make decisions whether or not to enroll. Moreover, I find that frequent visits are strong predictors of enrollment, suggesting the potential to use website engagement for enrollment management purposes.

Table of Contents

List of Figures

Acknowledgements

I would like to recognize the guidance and support of Professor Dvorak and David Glasser. The insights and advice both Professor Dvorak and David provided were essential to the completion of this project. It has been an absolute pleasure working with them the last two terms and I wish them both all the best in the future.

Chapter 1

Introduction

According to Abrahamson (2000), prospective students rank college websites as the second most important source of information – right below a campus visit. While a website cannot change the fundamental characteristics of a college (e.g. location, reputation, selectivity), it has the potential to highlight the positive characteristics, and to direct the narrative that prospective students see. An effective website design requires an understanding how prospective students actually use college websites. In this study, I perform a digital engagement analysis of Union College's website to examine which destinations accepted applicants visit, and to determine if accepted applicants' website usage predicts enrollment.

The for-profit segment has been an early adopter of comprehensive digital engagement analysis as a means of driving business results and enhancing companies' brands. Digital technology allows companies to understand consumers' behaviors. Every time an Internet user visits a company's website, the company's server records the user's interaction in the form of server web logs (Farney and McHale 2013). For-profit businesses scrutinize server web logs because it supplies tacit knowledge of consumers. This knowledge gives companies insight on how to structure websites, create product offerings, and design marketing strategies that drive consumption (Wind and Vijay 2001).

Digital engagement analysis can also be employed in the non-profit segment to help meet critical objectives. For institutions of higher education, one key goal is to attract and enroll prospective students of high academic quality. A compelling website

can be a critical tool in this effort. Digital engagement analysis can provide institutions of higher education insight to the content most important to prospective students. This provides institutions the opportunity to focus their attention on the messaging and effectiveness of the destinations most frequently visited, thereby increasing the pool of applicants and enrolled students.

Institutions of higher education can also use the insights from digital engagement analysis in enrollment management. Each year institutions need to determine the appropriate number of applicants to admit. Current techniques of enrollment management focus on applicant demographic information, such as academic profiles and financial aid needs, to determine probability of enrollment. My digital engagement analysis of Union College suggests that an analysis of website traffic can help predict candidate decisions and therefore assist the administration in determining the admittance decisions.

Chapter Two

Literature Review

In the last two decades the evolution of the Internet significantly changed how businesses and institutions market products and services. The Internet provides companies with innovative ways to reach customers online, monitor consumers' online behaviors, and predict consumers' intent. Many studies discuss the evolution of Internet marketing and the capabilities it provides.

2.1. Digital Marketing Evolution

One of the earlier sources to describe digital marketing is "Digital Marketing: Global Strategies from the World's Leading Experts" by Wind and Vijay. Wind and Vijay discuss how advancements in technology, specifically the Internet, benefit companies through digital marketing. Wind and Vijay state the "Internet is a vast electronic marketplace with high liquidity, the economic feasibility of providing large amounts of information to agents at a very low cost, as well as the capability to deliver differentiated and customized information" (Wind and Vijay 2001). The Internet thus provides companies and institutions the ability to instantaneously and inexpensively market products and services to a large number of consumers. Digital marketing therefore can greatly help companies and institutions drive business.

Building on Wind and Vijay's work, Harden and Heyman's book, "Digital Engagement: Internet Marketing that Captures Customers and Builds Intense Brand Loyalty," offers a newer insight to the evolution of digital marketing. In the last two decades, major advertisers shifted from traditional to online advertising. From 2006 to 2011, the total spending of Internet advertising in the United States rose by 149% (from

$16.9 billion to $42.0 billion) (Harden and Heyman 2009). The sharp increase of digital marketing arose due to the growth of consumers' Internet usage and the development of Internet capabilities. Through the hypertext transfer protocol (HTTP) format of Internet server access logs, companies and institutions have the ability to monitor users' online behavior. Companies can see every destination a user visits, and the date, time and duration of the destination visit (Harden and Heyman 2009). With this information companies and institutions run digital engagement analysis to facilitate digital marketing strategies.

2.2. Analytics

In "Business Intelligence and Analytics: From Big Data to Big Impact," Chen, Chiang, and Storey discuss the applications and significance of business intelligence and analytics (BI&A). The authors define BI&A as "techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions" (Chen, Chiang, and Storey 2012). The authors demonstrate the significance of BI&A practices through a 2011 Bloomberg BusinessWeek study that shows 97% of companies with revenues over $100 million use some form of business analytics (Chen, Chiang, and Storey 2012). While large corporations utilize BI&A in everyday business practices, smaller companies and institutions also use forms of analytics, such as website based analytics, to make business decisions.

Elizabeth Black provides an example of a study that uses website analytics within the academic environment. In "Web Analytics: A Picture of the Academic Library Web Site User," Black uses website analytics to determine if users find content on the Ohio

State University's (OSU) Library website useful. Black uses AWStats, a log file analysis program, to analyze server log files from January 2005 to December 2006. Black finds average visits per month grew by 52% from 2005 to 2006 (Black 2009). Additionally, the pages with most visits are the homepage followed by citation guides and external database pages. Interestingly the homepage is the top entrance page while the citation guides and external database pages are the top exit pages (Black 2009). Black suggests that users locate their desired information on the top pages and therefore leave the website from that page. Black concludes users of the website find the content useful since the number of visits increases in the study period and the top content pages are the top exit pages.

"Applying Analytics for a Learning Portal: The Organic.Edunet Case Study" by Palavitsinis, Protonotarios, and Manouselis is a second example of a study that uses website analytics. The researchers investigate the effectiveness of informational sessions by monitoring the level of visitor traffic before and after each session. The learning portal of focus, Organic.Edunet Web Portal, is an international organic agriculture-based learning portal for high school and university level educators. The informational sessions, called Open Days, allow educators to test Organic.Edunet Web Portal with the help of portal professionals. Using Google Analytics, a website traffic analytics system, the researchers examine visitors' usage of the portal during two periods of two Open Days' sessions: March to May 2010 and August to September 2010.

The researchers focus on the frequency of use and the duration of the user experience. Researchers find that the number of users increases drastically from the day of the Open Days sessions to two months after the event. After the two-month period,

total visits drop significantly but at a level greater than that prior to the Open Days (Palavitsinis, Protonotarios, and Manouselis 2011). The researchers find the average time users spent on the portal in the two periods drops from six minutes to three minutes. Additionally, seven months after the Open Days the loyalty of users (measured by the percent of visitors that return more than once in a week) drops significantly (Palavitsinis, Protonotarios, and Manouselis 2011). The researchers conclude that Open Days drew positive increases in traffic immediately after each session, but led users to lose momentum in the long term.

2.3 College Website Analyses

Researchers in the case studies above present examples of how companies and institutions use website analytics to understand behaviors and preferences of their target audiences. In my research I found a lack of studies that analyze the effectiveness of college websites using website analytics. The few studies that do assess college websites predominantly use qualitative analysis.

In Jonathan Coffin's thesis, "Telling the Story of a Liberal Arts College: The College Website as a Window on Institutional Positioning," Coffin analyzes "how liberal arts institutions define institutional identity and establish differentiation within the higher education marketplace" (Coffin 2012). Coffin analyzes the language and visual presentation of ten randomly selected liberal arts colleges' homepages to identify the narratives and themes used to define each institution. Coffin finds that photography combined with brief text is the most commonly used technique to express identity. For example, when colleges display social life, colleges more likely show a plethora of pictures to highlight the ample and diverse activities students can take part in. When

displaying academics, colleges try to show happy students in a library to show academic rigor within an easy-going culture (Coffin 2012). The second most common technique to communicate identity is providing institutional voice. Coffin describes institutional voice as a statement(s) from a representative of the college who communicates a description of the institution. Coffin finds colleges to appear more authentic, appealing, and prestigious when institutional voices express identity in an indirect manner. Coffin concludes, to best express identity of a liberal arts college homepage, colleges should employ narratives through photography and language in indirect institution voice(s).

In Michael Poock's and Dennis Lefond's study, "How College-Bound Prospects Perceive University Websites," researches analyze how college-bound seniors perceive college websites. Poock and Lefond focus on finding what elements promote website browsing and which elements increase the likelihood of submitting applications. The study includes qualitative and quantitative testing of 55 college-bound seniors. Researchers first ask the students what their general Internet habits are and what their opinions are of college website homepages. Second, researchers ask students to view specific college website homepages and offer their opinions by ranking, on a five point scale, the importance of content, organization, graphics, distinctiveness, and download speed. Lastly, researchers ask students to locate specific information within a college's website while being timed.

Based on the opinions of the students in the first and second parts, content is the most important element to a college website with site architecture, ease of navigation and speed of downloads following in order. The two forms of content students expect to see most are environmental offerings, like athletics or extracurricular clubs, and admissions

information (Poock and Lefond 2001). The results of the ease of navigation test displayed frustration of students due to the amount of time it took to locate specific information. On average, it took the students 3 minutes and 45 seconds to locate desired information. Poock and Lefond attribute the higher than expected results to students needing to drill down through three or more levels of links to find the desired information. The researchers note that to find online applications students need to drill down through five to six links. Poock and Lefond conclude prospective students are more likely to browse websites and submit applications on college websites that offer institutional content and an intuitive organization

To supplement the studies that assess college websites using qualitative analysis, I perform a quantitative digital engagement analysis on the applicant traffic on Union College's website.

Chapter 3

Data

In this section I describe the three datasets I use to find destinations applicants visit and test the correlation between applicants' website interaction and their probability of enrollment. The first dataset comes from the applicant portal. The applicant portal provides identification numbers for every 2014 regular decision applicant and the Internet Protocol addresses (IP address) each applicant uses to access the portal. The second dataset comes from the website access log files (or web traffic) of Union College's website. This dataset provides IP addresses of visitors, dates of visits, and the destinations the visitors visit. The third dataset includes characteristics of each applicant and the identification number designated to him or her.

3.1 Applicant Portal

When applicants apply or review the status of their application online, they use the applicant portal. Each time an applicant log into the applicant portal with their user ID and password, the applicant's IP address(es) links to his or her user ID. The applicant portal includes 12,208 observations for 4,095 identification numbers and 11,379 IP addresses. This means that applicants view the applicant portal from more than one computer. On average applicants access the portal from approximately three different IP addresses. In few cases multiple applicants view the applicant portal from the same IP address. To accurately interpret applicant traffic I remove IP addresses linked to more than one ID. In the final version of the applicant portal dataset there are 10,941 IP addresses associated with 3,936 identification numbers.

3.2 Union College Website Access Log Files

The website access log files (web traffic), supplied by Union College's ITS office, reveals how applicants interact with the Union College website. The web traffic includes 148,468 total page views for 4,146 IP addresses from March 23rd to April 10th 2014.

Every time a computer connects to the Internet, that computer acquires a distinct IP address that identifies the computer on the Internet. Once the computer attempts to access a website page, the IP address of the computer sends a request to the website's server. The website's server then records each request the IP address makes. This process describes the creation of a website's server access logs (Goncalves and Ramasco 2008).

The format of Union College's website server access logs is the Common Log Format (CLF). The CLF represents each server access log by the following elements: the remote host, server name, log date, request, status, bytes, referrer and user agent. Descriptions of each element are below.

- Remote host: IP address of computer requesting website page
- Server name: Name of the website's server which an IP address seeks to connect with
- Log date: Date, time, and time zone of the IP address' request
- Request: Website page request (represented by the requested URL) of the IP address
- Status: Three digit HTTP status code indicating whether request was completed
- Bytes: Size of the requested object (in terms of bytes) that is sent to the IP address from the server
- Referrer: Name of the server that linked the IP address to the current server

- User agent: Name of the browsing or operation system the IP address used

(S. Bosnjak, Maric, and Z. Bosnjak 2009)

I use the remote host, log date, and request to indicate the website interaction of each applicant. The request represents the website destination of each applicant view. To use the request in the analysis, I split each request by "/"s into four levels. Level1 represents the type of request the visitor requests. The types of requests mainly come in two forms: "GET" and "POST." "GET" means a visitor requests to view a page and "POST" means a visitor requests to submit something. Level2 represents the high-level categorical buckets of the website's pages. For example, if a visitor views the "open-houses" page (www.union.edu/admissions/visit/open-houses/), "admissions" is the level2 item. Level2, level3 and level4 are the following categories down from level2. In the same example above, the level3 item is "visit" and the level4 item is "open-houses." Throughout the analysis, I focus on level2, level3, and level4 to identify applicants' website destinations.

3.3 Applicant Characteristics

The applicant characteristics dataset, provided by Union College's Admissions office, includes 56 variables describing the 4,626 regular decision applicants. The 56 variables provide a synopsis of the applicants' characteristics from admissions status to demographics. Of the 56 variables, I focus on admissions status.

Within admissions status are two variables; *admitted* and *enrolled*. The *admitted* variable indicates whether an applicant is admitted or not. In the dataset there are 4,626 applicants, but only 1,898 of them are admitted. Since the focus of the analysis is on web behavior of admitted applicants, I remove the 2,728 non-admitted applicants. The

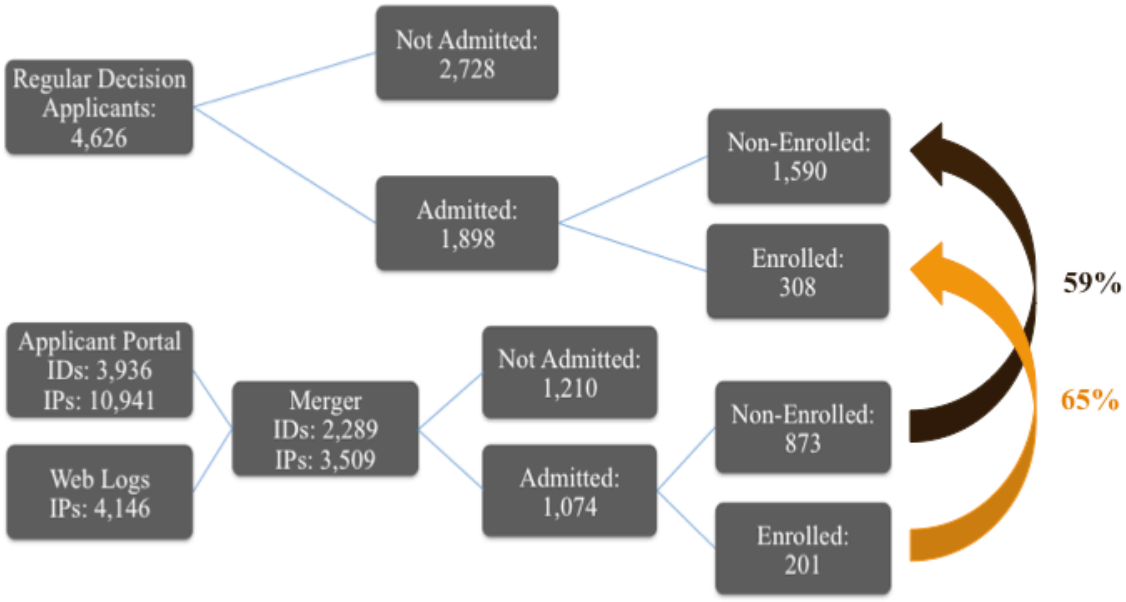
enrolled variable indicates whether admitted applicants enroll or do not. Of the 1,898 admitted applicants, 308 applicants enroll while the remaining 1,590 applicants did not.

3.4 Merging Datasets

I merge the three datasets in two steps to create the final dataset. First I merge the applicant portal data and the web traffic data by matching IP addresses. Of the 10,941 IP addresses in the portal, only 3,509 are found in the traffic log. These 3,509 IP addresses correspond to 2,289 IDs. The number of IDs is lower than in the applicant portal because not all applicants visit the website and the portal from the same IP address.

Second, I merge the dataset with web traffic information with the applicant characteristics dataset by matching IDs. Of the 1,898 accepted applicant IDs in the applicant characteristics, only 1,074 have matching IDs in the web traffic dataset. Within the 1,074 IDs, 201 enrolled and 873 did not enroll. In the original applicant characteristics dataset, with 4,626 total applicant IDs and 1,898 accepted applicant IDs, 308 IDs enrolled. The final dataset thus includes 65% of the applicant IDs who enrolled and 59% of those who did not enroll. Figure 3.1 summarizes the merger process.

Figure 3.1: The Distribution of all Regular Decision Applicants, and Traffic Data Availability



Chapter 4

Data Exploration

In the story below, I begin with an overview of accepted applicants' website behavior over time. Then I transition to an aggregate view of applicant visits by website destinations. Lastly I breakdown applicants' website traffic by level2, level3 and level4 destinations over time.

4.1 Traffic Overview by Time

Figure 4.1 displays how accepted applicants use Union College's website over time. The figure indicates the number of applicants on the website each day in two ways: one, total number of applicants (combined height of red and blue bars), two, breakdown of applicants by enrollment status (red bars represent non-enrolled applicants and blue bars represent enrolled applicants).

Figure 4.1

Figure 4.1 displays notable days accepted applicants visit the website. Those days are 3/24, 3/25, 4/26 to 5/1, and post 5/1. For the most part these days are fairly significant to the Admissions offices' regular decision timeline. Specifically, 3/24 and 3/25 fall four/five days after regular decision letters are mailed and two/three days after applicant decisions are posted on the applicant portal. It's no surprise that 3/24 and 3/25 are the days accepted applicants visit the website most (331 applicants on 3/24 & 275 applicants on 3/25) due to the close proximity to the decision announcements. On these days non-enrolled applicants visit the website more than enrolled applicants, but the share of enrolled applicant views (enrollment share) (26% on 3/24 and 27% on 3/25) is greater than the yield of accepted applicants in the study (18.7%).

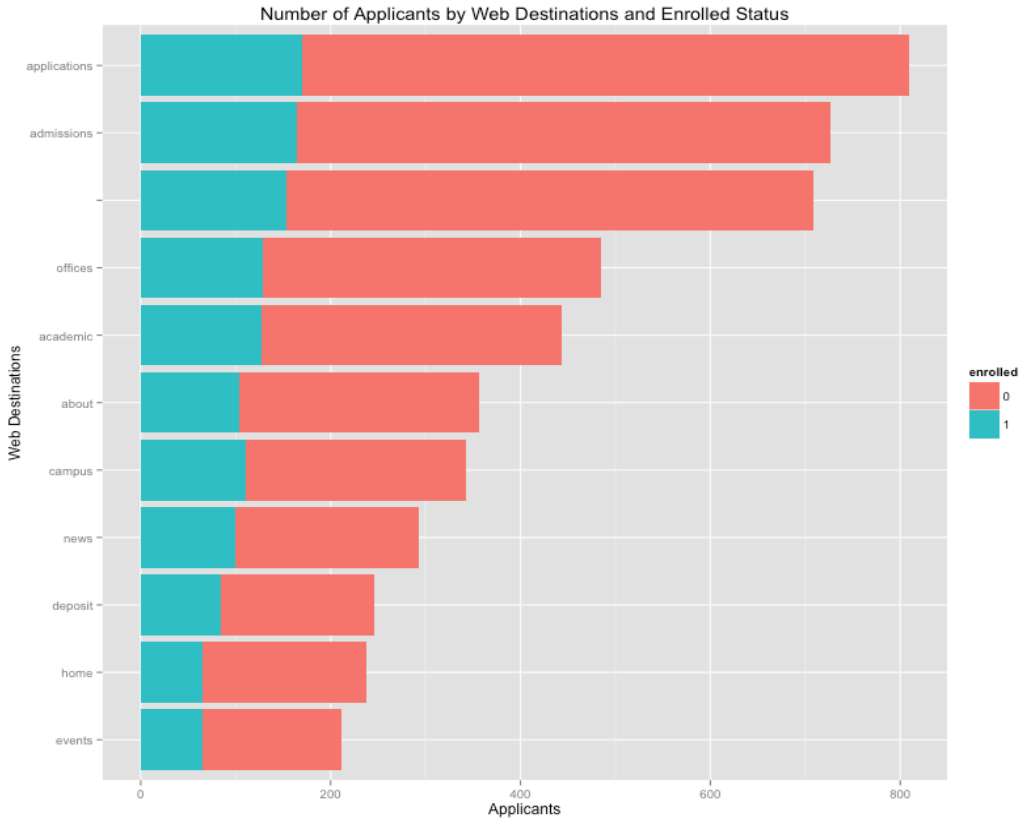
From 4/26 to 5/1 (week leading up to deposit deadline), applicants visit the website increasingly day after day. The deposit deadline represents applicants' final day to enroll, so the rise of applicant views indicate either the website is a resource in applicants' final decision and/or applicants submitting deposits online.

From 5/1 to the end of the time plot represents the period after applicants enroll or do not. In this period there is a steep drop in the number of applicant views. This could indicate applicants no longer needing the website as a decision making resource. Interestingly, non-enrolled applicants continue to visit the website. On average 47 non-enrolled applicants visit the website each day after the deposit deadline.

4.2 Traffic Aggregated by Destination

The figures in this section breakdown destinations by levels, applicant views, and enrollment status.

Figure 4.2



Note: Figure only shows level2 destinations with applicant views greater than 200

Figure 4.2 indicates the level2 destinations with most applicant views are "applications" (809 applicants), "admissions" (727 applicants), the website homepage (709 applicants), "offices" (485 applicants), "academic" (444 applicants), "about" (356 applicants), and "campus" (342 applicants). There are two reasons "applications" receives the most applicant views: one, to check the status of applications, applicants visit "applications" (specifically the applicant portal); two, I gather applicants' IP addresses from the applicant portal. Thus I know each applicant visits "applications" at least once in the study period.

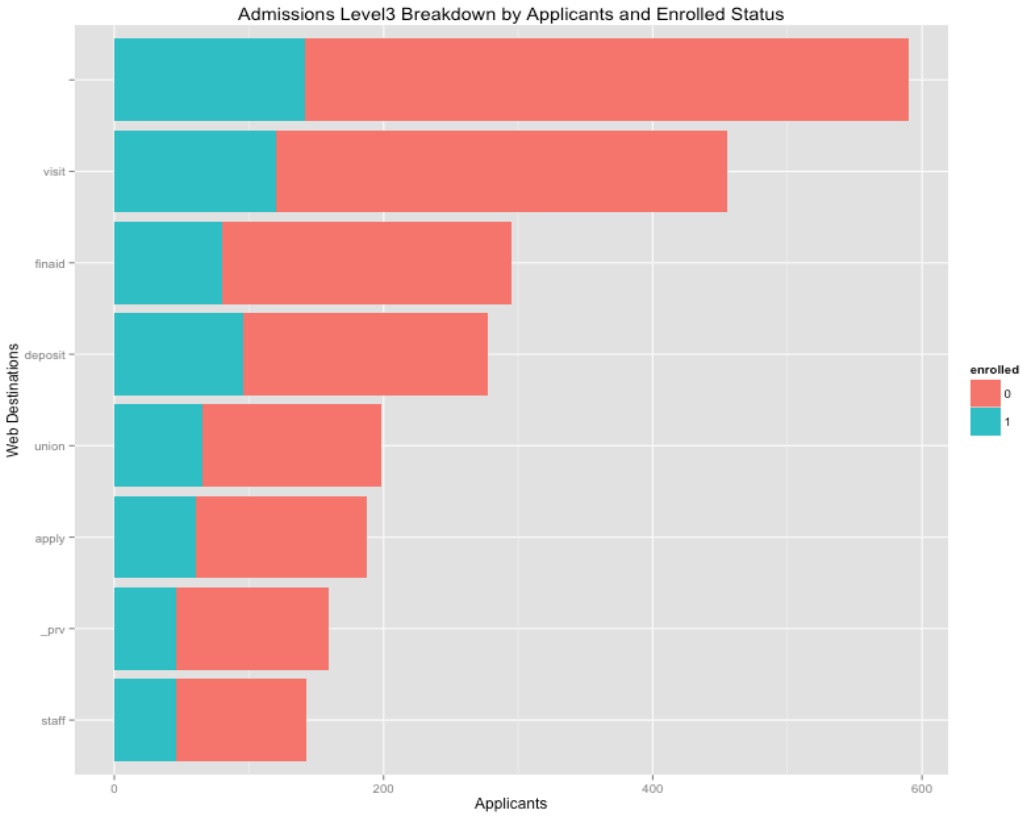
Of the top level2 destinations, I focus on the ones that contain the most meaningful level3 destinations: "admissions," "offices," "academic," "about," and

"campus." The enrollment share of these destinations exceeds the yield of accepted applicants further indicating their significance.

4.2.1. Admissions

In Figure 4.3, "admissions" is broken down by level3 destinations and enrollment status.

Figure 4.3



Note: Figure only shows level3 destinations with applicant views greater than 100

Figure 4.3 indicates the most viewed level3 destinations are the admissions homepage (590 applicants), "visit" (455 applicants), "finaid" (295 applicants), and "deposit" (277 applicants).

The level3 destination of most interest is "visit." The "visit" page provides applicants information on ways to experience Union College through campus tours, open houses, interviews, info sessions, or day programs. To ascertain how applicants prefer to

experience Union College on "visit," I break the page down by level4 destinations in Figure 4.4.

Figure 4.4

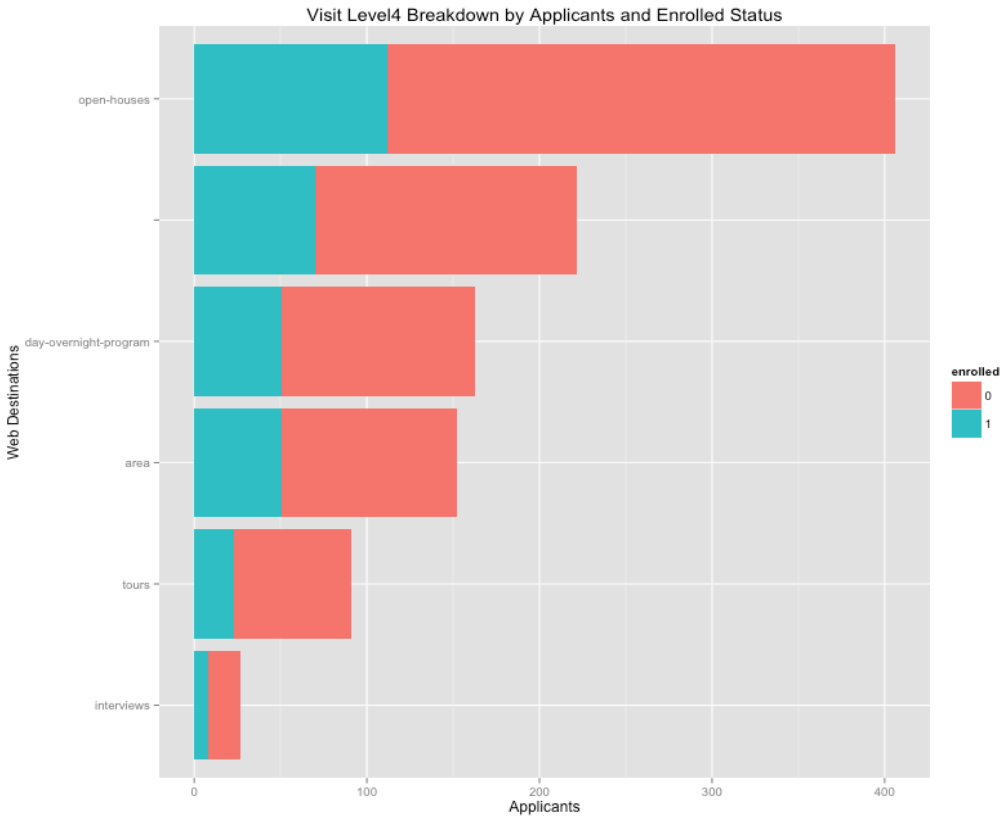
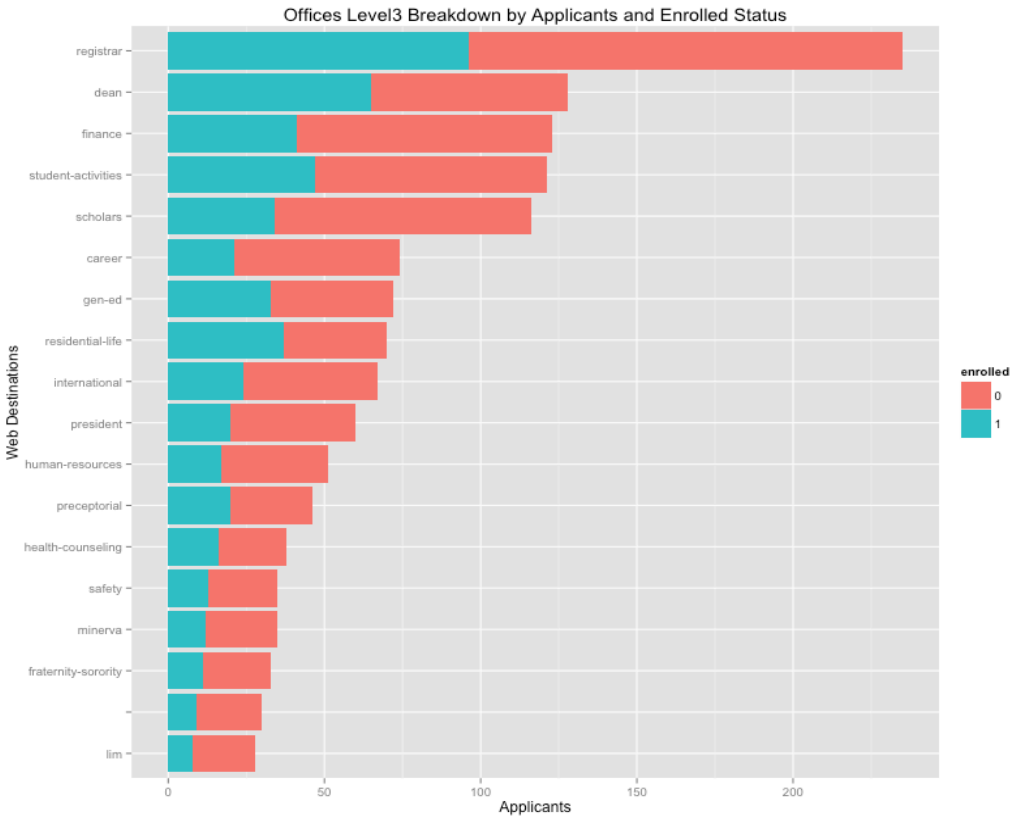


Figure 4.4 indicates that applicants view "open-houses" the most with 406 applicants. Though the enrollment share of "open-houses" is low in comparison to the other level4 destinations, more enrolled applicants visit this page than the others with 112 applicant views. This may suggest that applicants desire to attend open house events to help decide whether to enroll or not.

4.2.2. Offices

In Figure 4.7, "offices" is broken down into level3 destinations and enrollment status.

Figure 4.5



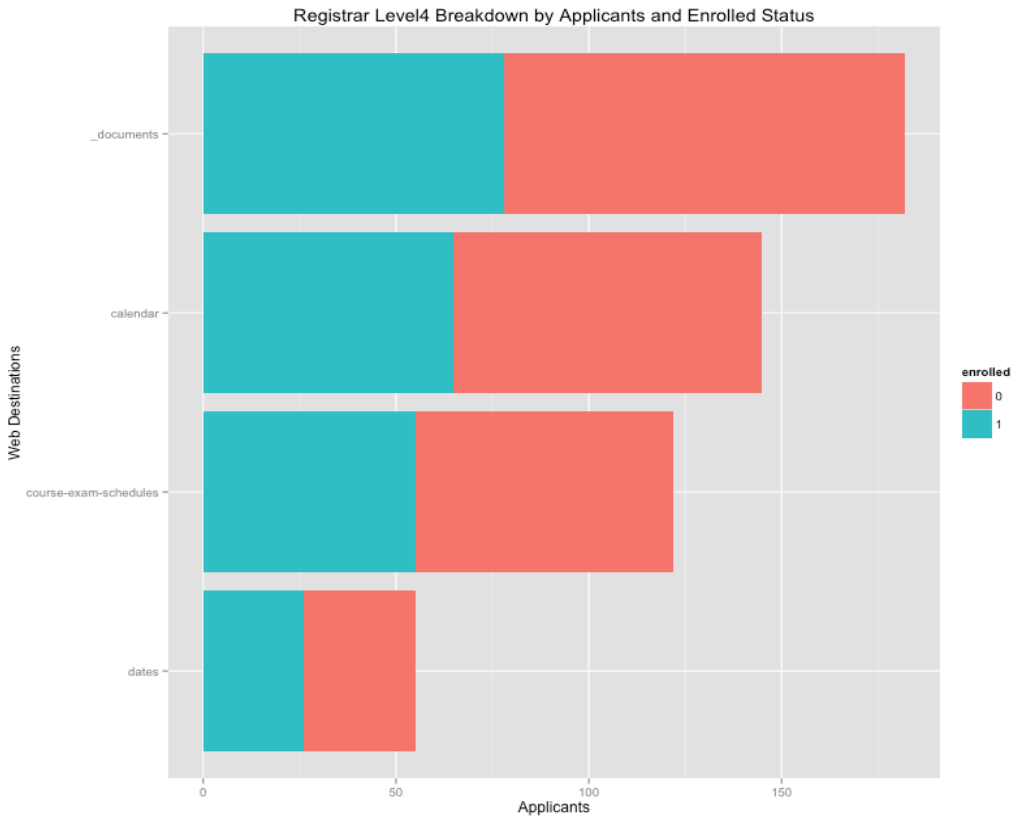
Note: Figure only shows level3 destinations with applicant views greater than 25

Figure 4.7 demonstrates that total applicants as well as enrolled applicants visit "registrar" the most (235 total applicants (96 enrolled applicants)). The "registrar" page provides academic resources for students regarding curriculum, dates, advising and much more. Additionally, applicants also visit "dean," "student-activities," and "finance" in large numbers. The figures below breakdown these pages into level4 destinations.

Figure 4.6 shows the level4 destinations in "registrar." The results indicate that enrolled applicants, and total applicants in general, visit "calendar" (65 enrolled applicants) and "course-exam-schedules" (55 enrolled applicants) the most. "Calendar" provides important academic dates for the current and two upcoming academic years. "Course-exam-schedules" provides links for exam schedules and current course

offerings. This result indicates academic dates and course offerings are an importance resource for applicants who tend to enroll.

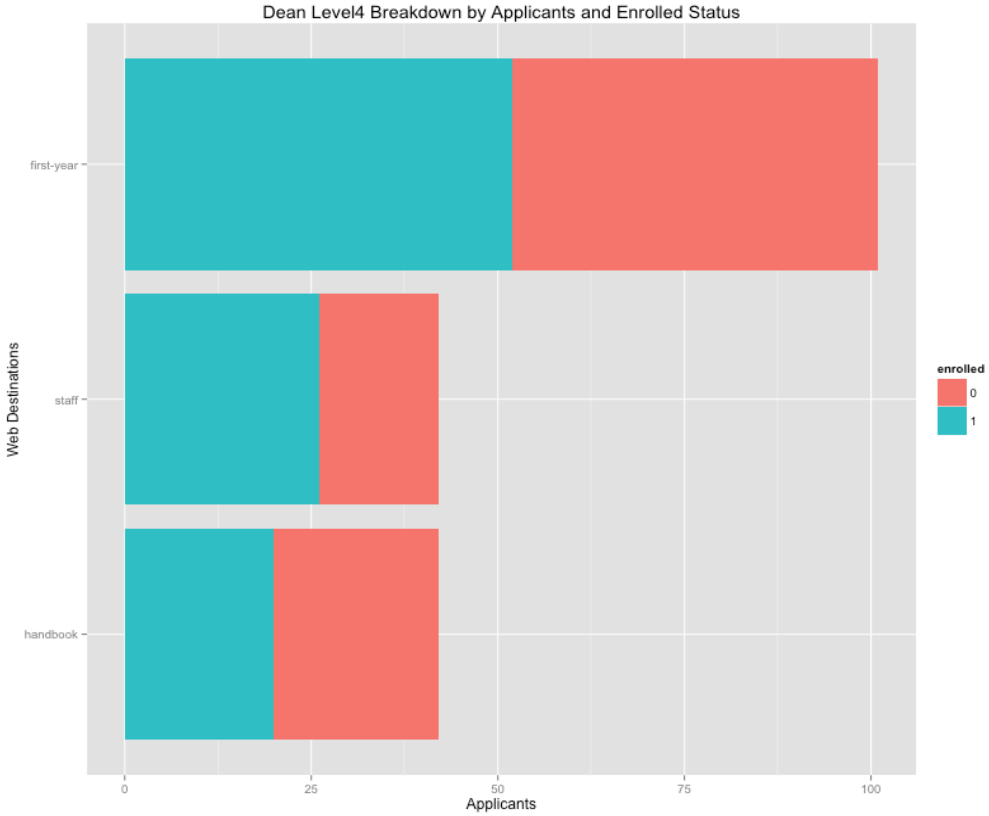
Figure 4.6



Note: Figure only shows level4 destinations with applicant views greater than 50

Figure 4.7 shows the level4 breakdown of "dean." The figure indicates that applicants visit "first-year" the most with 101 applicants. This result is no surprise since applicants are directed to the "first-year" page after making an online deposit. What's surprising is that of the 101 applicant views, only 52 correspond to enrolled applicants. This demonstrates that "first-year" is a resource to all applicants.

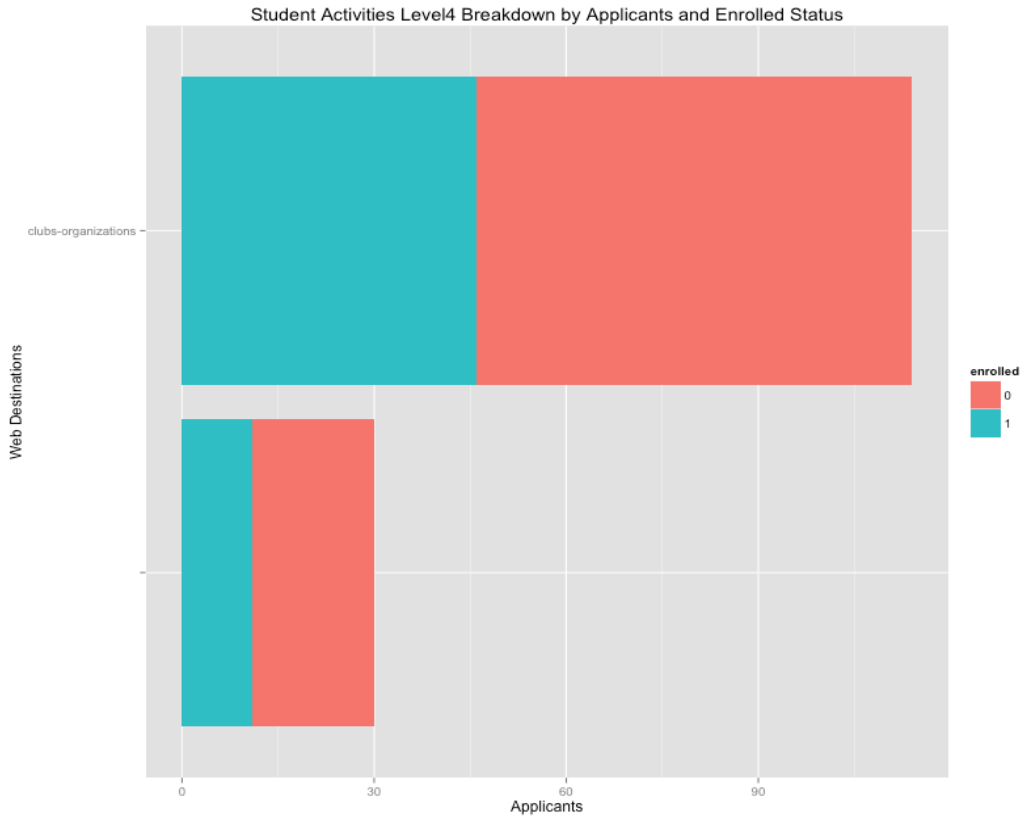
Figure 4.7



Note: Figure only shows level4 destinations with applicant views greater than 25

Figure 4.8 shows the level4 breakdown of "student-activities" by enrollment status. The figure demonstrates that applicants, specifically enrolled applicants, visit "clubs-organizations" the most (114 total applicants and 46 enrolled applicants). On the "clubs-organizations" page it lists and describes Union's 100 plus clubs and organizations. The destination thus acts as an extracurricular resource for applicants.

Figure 4.8



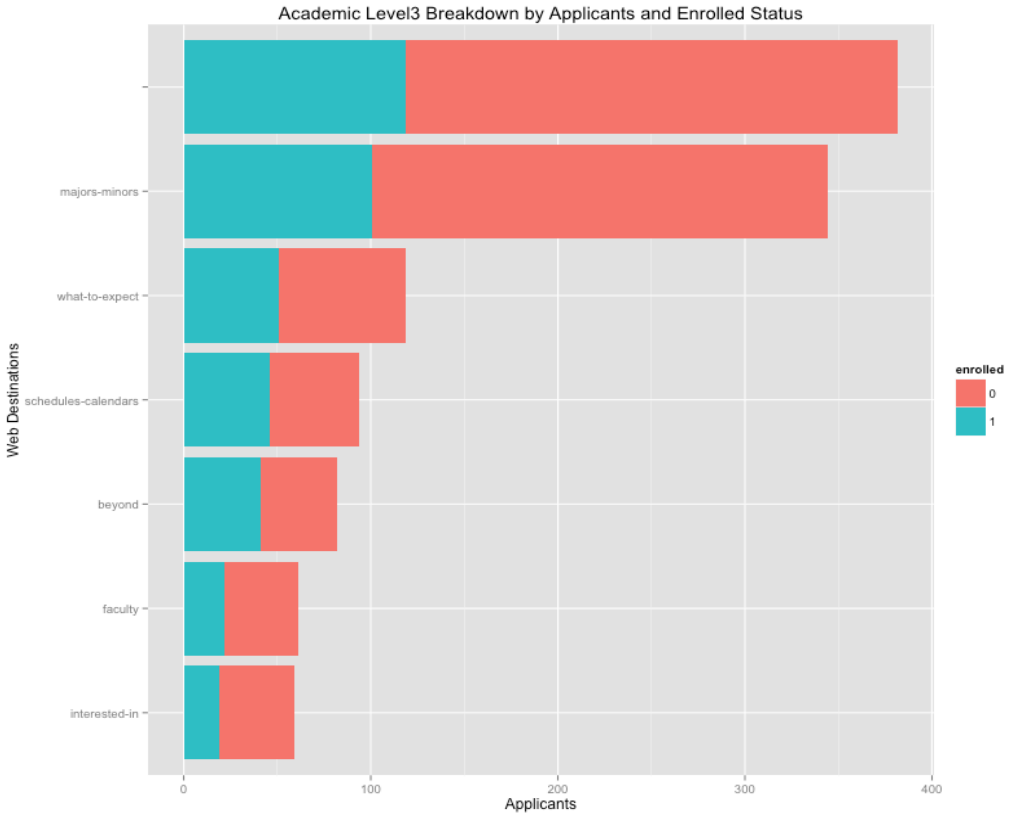
Note: Figure only shows level4 destinations with greater than 10 applicant views

The last level3 in "offices" is "finance." In "finance" the only level4 destination that receives more than 10 applicant views is "student-accounts" with 116 total applicants (38 enrolled applicants). The "student-accounts" page provides account billing information and links for information on financial aid and payment options.

4.2.3. Academic

Figure 4.9 displays the level3 destination breakdown of "academic" by enrollment status. The main destinations applicants, specifically enrolled applicants, visit are the "academic" homepage (382 total applicants & 119 enrolled applicants) and "majors-minors" (344 total applicants & 101 enrolled applicants).

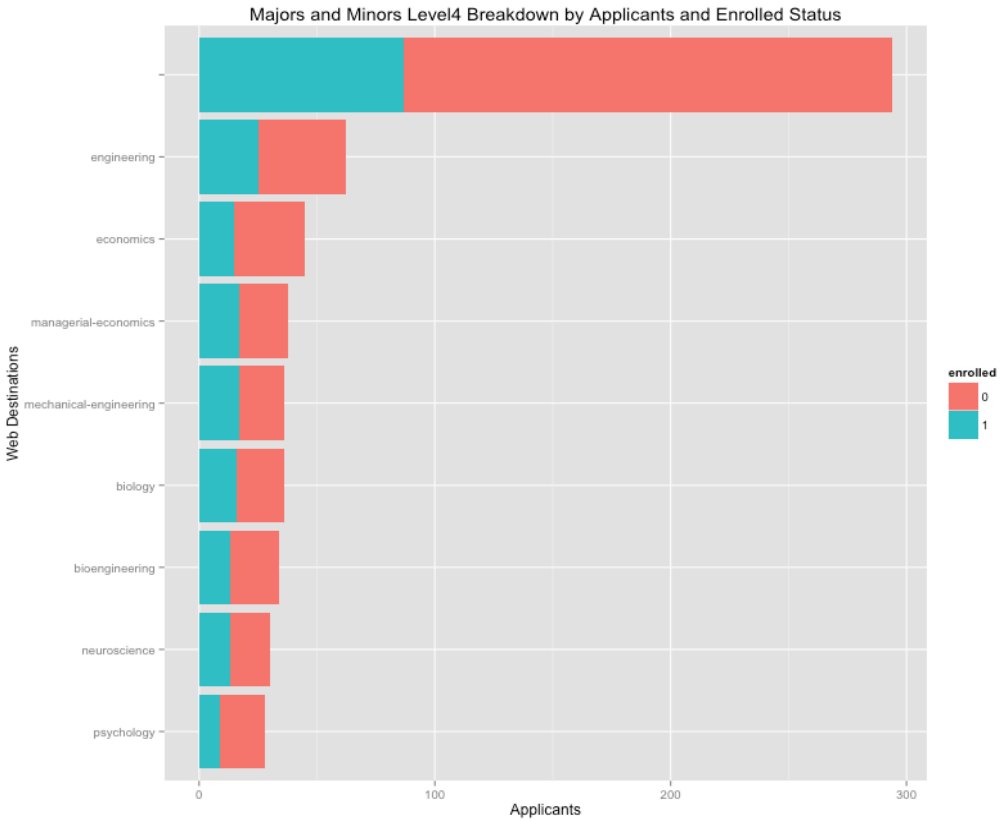
Figure 4.9



Note: Figure only shows level3 destinations with applicant views greater than 50

The "majors-minors" page is significant because it lists and provides links for the 44 majors and minors offered at Union College. I break down "majors-minors" by level4 destinations in Figure 4.10 to determine the major's and minor's applicants are most interested in.

Figure 4.10



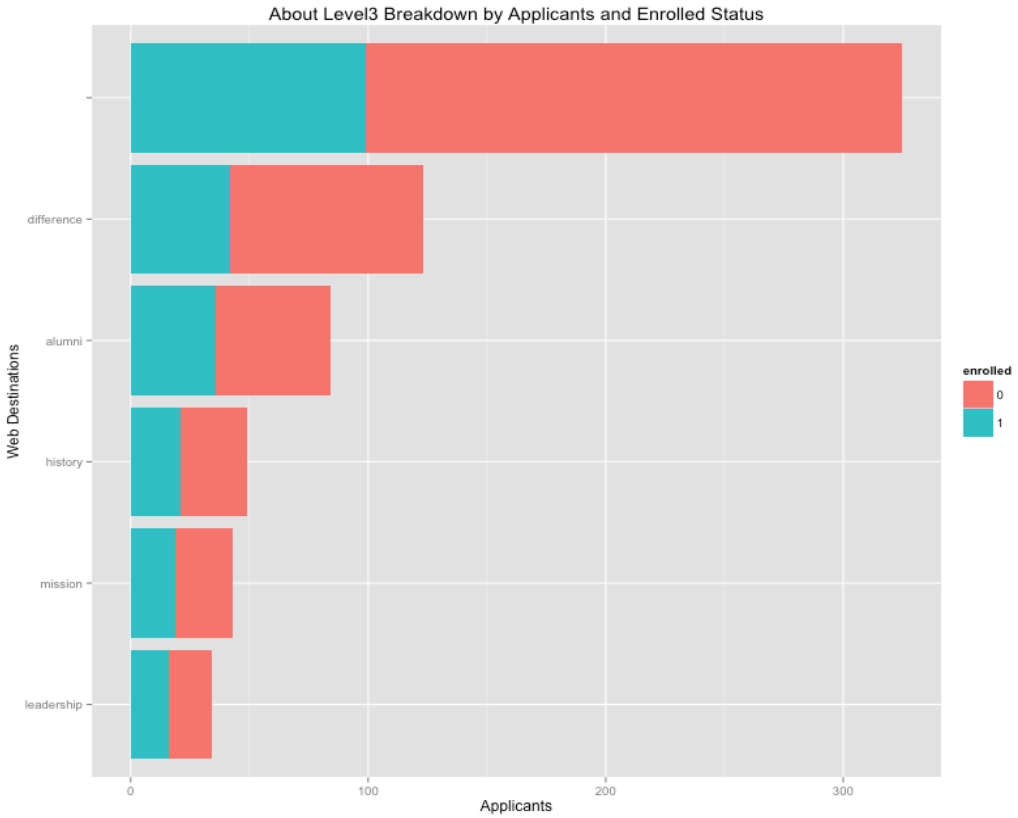
Note: Figure only shows level4 destinations with applicant views greater than 25

Figure 4.10 indicates that the majority of applicants who visit the "majors-minors" page mainly look at the homepage (294 applicants). The result highlights that applicants are more interested in viewing the list of major and minors than viewing specific major's and minor's pages. Applicants who did visit a major's or minor's page visit "engineering" (62 applicants) followed by "economics" (45 applicants) the most. Enrolled applicants visit "engineering" (25 applicants) followed by "managerial-economics" (17 applicants) and "mechanical-engineering" (17 applicants) the most. Interestingly, enrolled applicants visit the "managerial-economics" page slightly more than the "economics" page.

4.2.4. About

Figure 4.11 displays the level3 breakdown of "about" by enrollment status.

Figure 4.11



Note: Figure only shows level3 destinations with applicant views greater than 20

The figure indicates that applicants visit the "about" homepage the most with 325 applicant views. The "about" homepage, unlike other level3 homepages, provides a considerable amount of information. Specifically the page describes the college's identity, missions, and factors that differentiate the school from competitors.

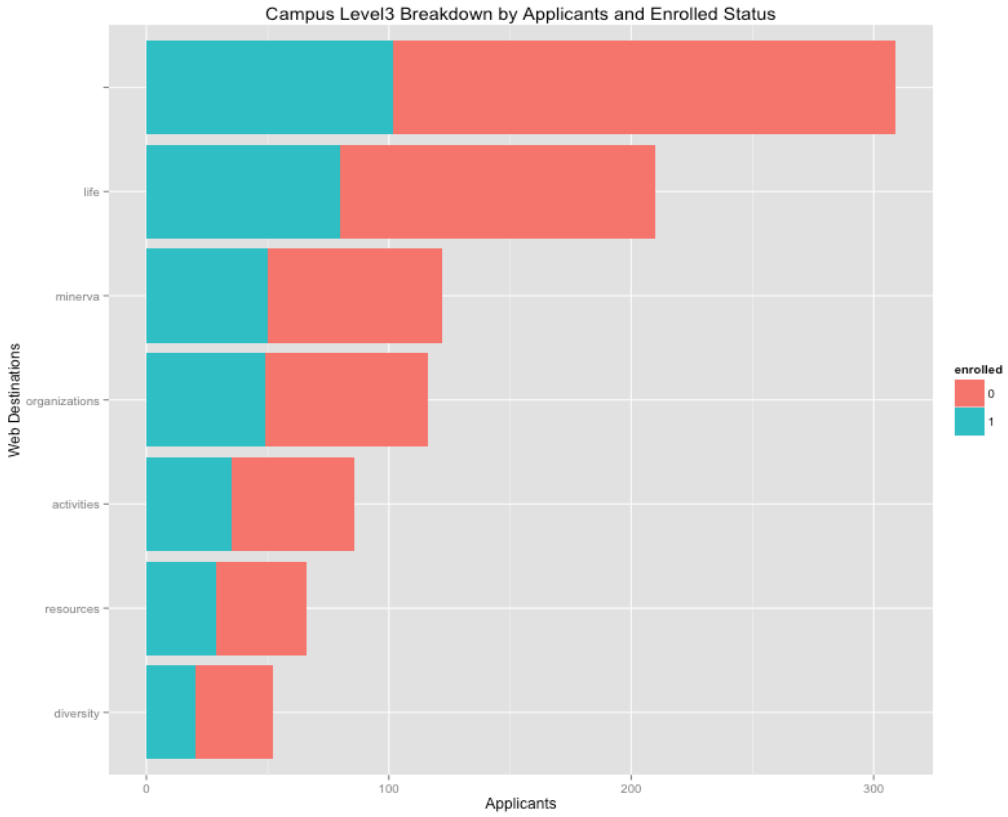
The homepage also receives the most enrolled applicant views (99 applicants) followed by "difference" (42 applicants) and "alumni" (36 applicants). Since total applicants, and specifically enrolled applicants, visit the "about" homepage the most, I do not breakdown any of level3 destinations further. This result demonstrates that the

"about" homepage is the main resource for applicants interested in learning about the college's identity.

4.2.5. Campus

The final level2 destination of focus is "campus." Figure 4.12 displays the level3 breakdown of "campus" by enrollment status.

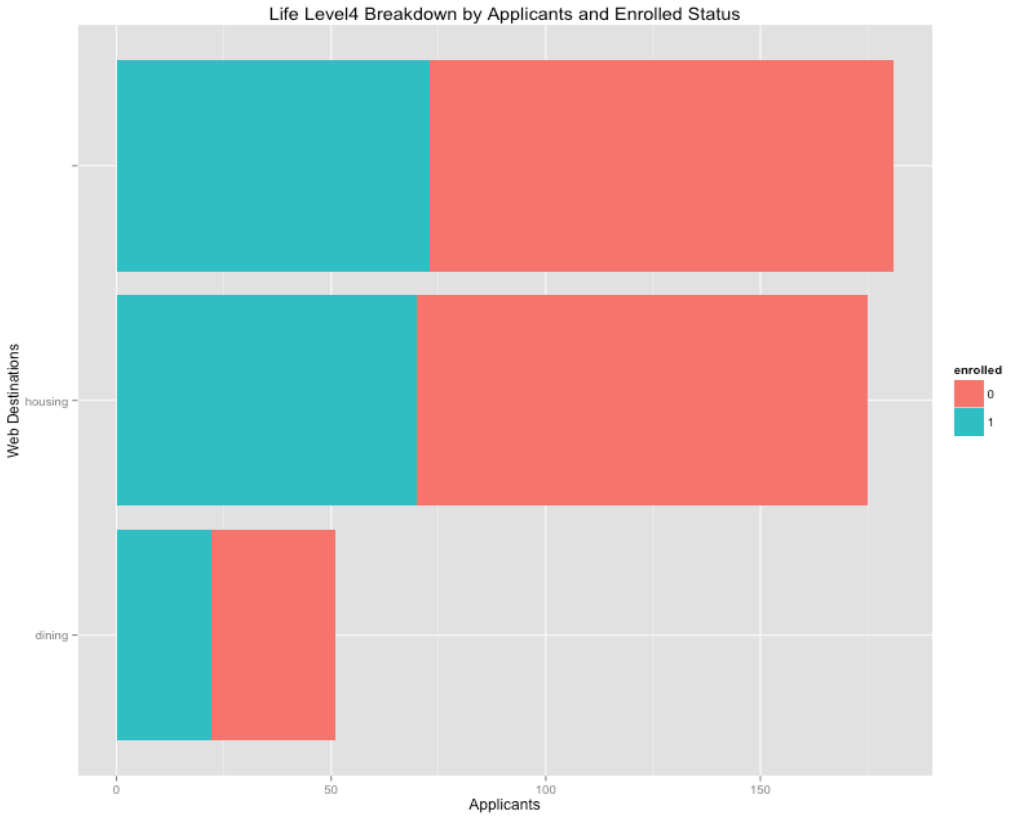
Figure 4.12



Note: Figure only show level3 destinations with applicant views greater than 50

The figure indicates that the "campus" homepage (309 applicants) and "life" (210 applicants) are the pages applicants view most. Similarly, enrolled applicants also visit the homepage (102 applicants) and "life" (80 applicants) the most. Since the homepage cannot get broken down further, I breakdown "life" into level4 destinations and enrollment status in Figure 4.13.

Figure 4.13



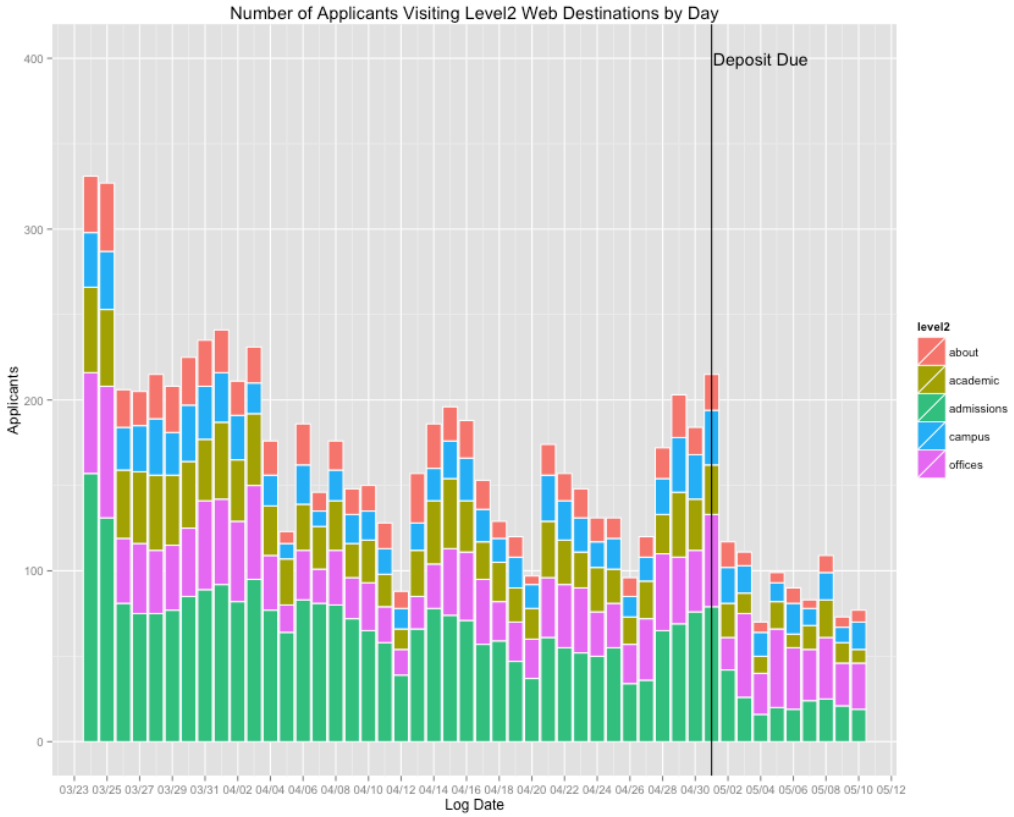
"Life" is an interesting level3 destination as there are only three level4 destinations within it; the "life" homepage, "housing," and "dining." Figure 4.13 indicates that applicants visit "housing" just as much as the homepage. This result could indicate that applicants primarily visit the "life" page to learn about the college's housing options rather than the dining.

4.3. Traffic by Destination and Time

In this section I combine the analysis from the previous sections by analyzing applicants' website destinations over time. First I display how applicants visit the top five-level2 destinations by time. Second I break down each level2 destination into level3 and level4 destinations over time to indicate the variation of destination visits in the time plot.

The first figure, Figure 4.14, displays how applicants view the top level2 destinations in the time plot.

Figure 4.14



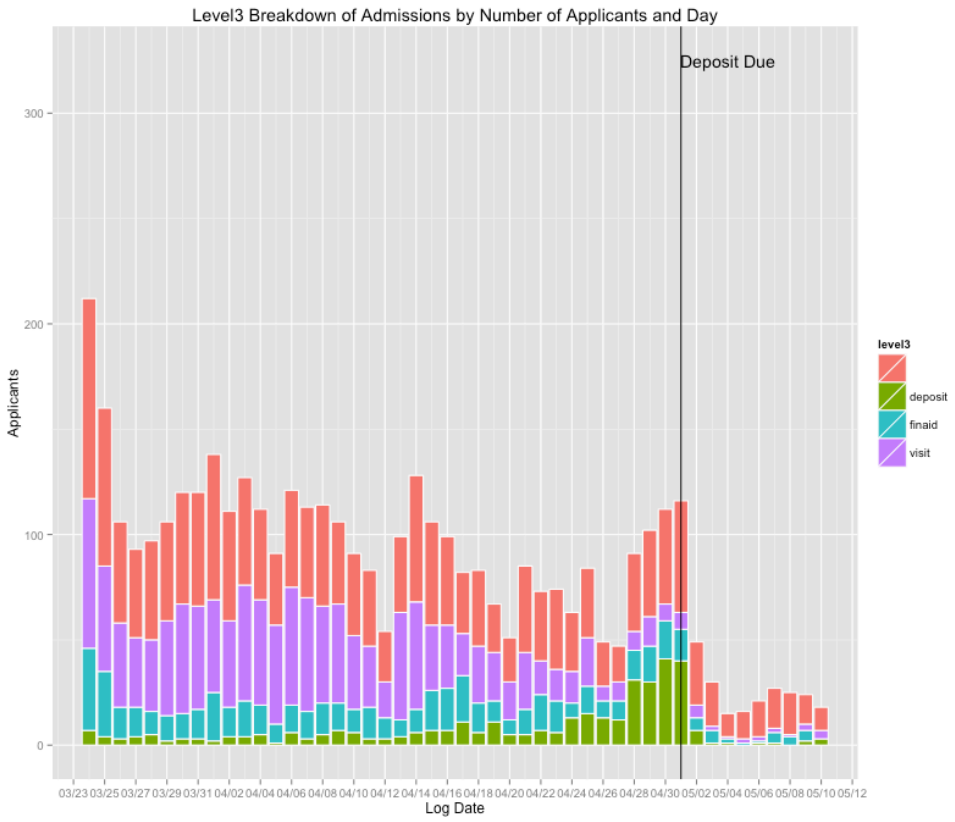
Note: Figure only shows level2 destinations with applicant views greater than 700. Figure also does not show "applications" or the Union College homepage

In Figure 4.14 there are noticeable patters to highlight. First, applicants view "admissions" more in the beginning of the time plot up to the deposit deadline. Once the deposit deadline passes, "admissions" views sharply decline, but applicant views of "offices" remain consistent.

4.3.1. Admissions

Figure 4.15 displays the level3 destination in "admissions" that applicants view over time.

Figure 4.15



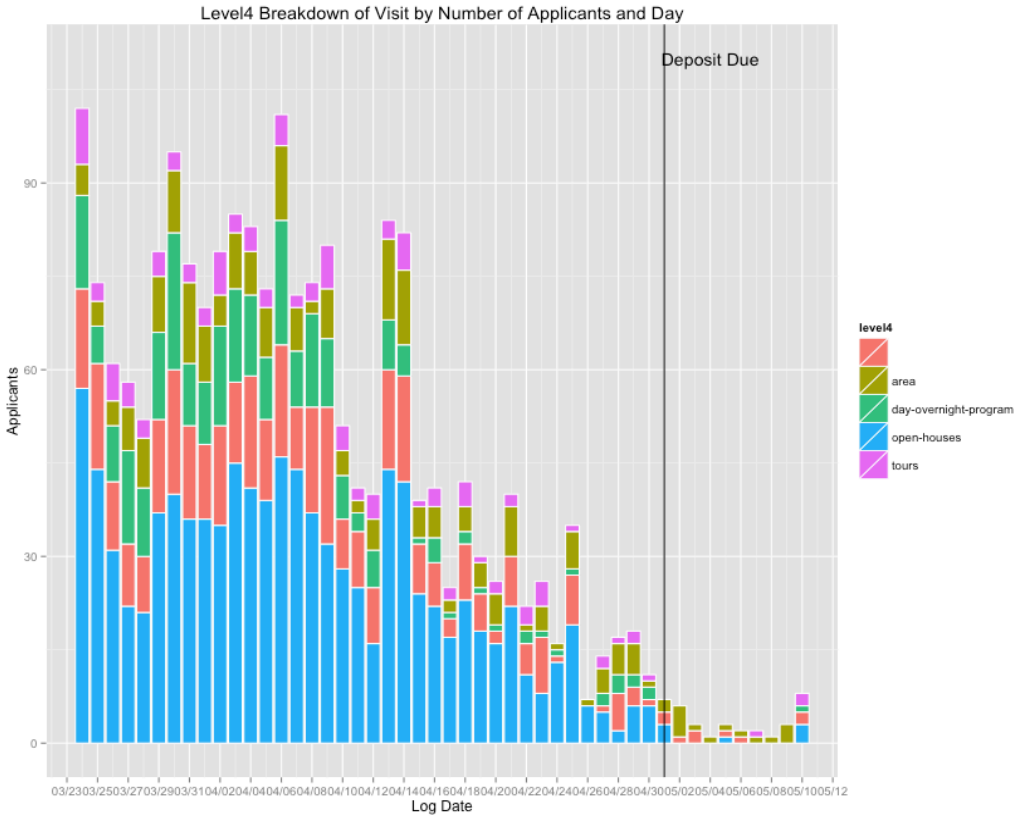
Note: Figure only shows level3 destinations with applicant views greater than 350

Looking at Figure 4.15 there are a few noticeable trends relating to "visit" and "deposit." From the beginning of the time plot to the week prior to the deposit deadline, applicants visit "visit" significantly. Since "visit" provides different ways to experience and learn more about Union College it's no surprise applicants view the page in large numbers. What's interesting is that applicant views of "visit" drop steeply a week prior to the deposit deadline. The level4 breakdown of "visit" over time (seen in Figure 4.16) displays this decline most notably with "open-houses."

The steep decline in applicant views could indicate the period applicants begin to decide if they are going to enroll or not. This is further highlighted by the applicant views of "deposit" consistently increasing on 4/26. Since the "deposit" page is where

applicants make their online deposit, the period from 4/26 to 5/1 could represent the time the majority of applicants make the final decision to enroll or not.

Figure 4.16

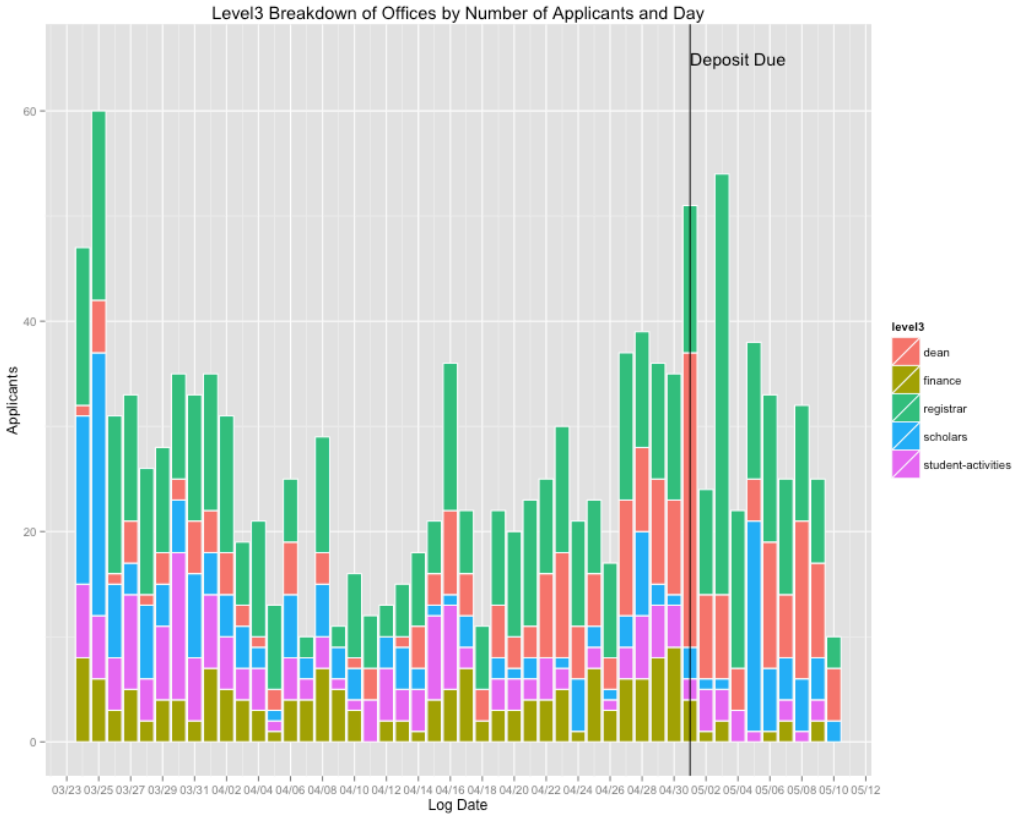


Note: Figure only shows level4 destinations with applicant views greater than 100

4.3.2. Offices

Figure 4.17 displays the level3 breakdown of "offices" over time.

Figure 4.17



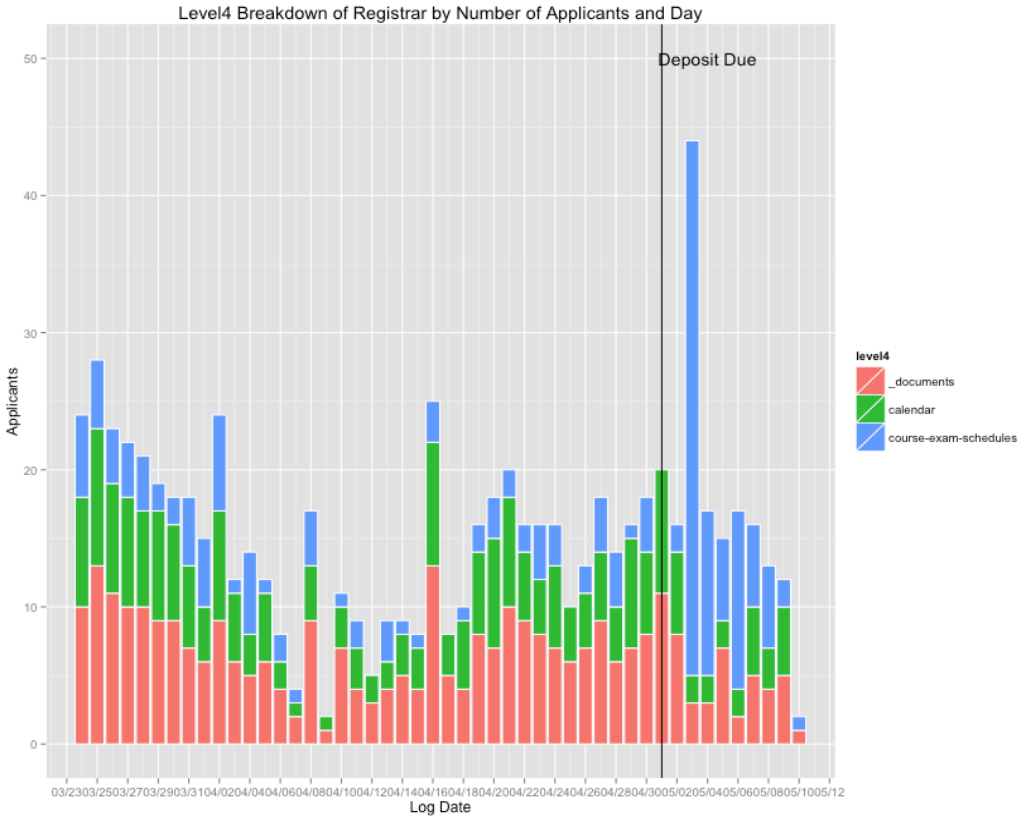
Note: Figure only shows level3 destinations with applicant views greater than 160

The figure highlights patterns corresponding with "registrar," "finance," "deans," and "scholars." Throughout the time plot "registrar" is noticeably the most viewed level3 destination as applicants consistently view the page in the beginning, middle, and end of the time plot. Applicants view "finance" almost entirely in one period, from the beginning of the time plot up to the deposit deadline. Applicants view "dean" from the week prior to the deposit deadline through the end of the plot. Lastly, applicants view "scholars" almost entirely on three separate days in the time plot.

In Figure 4.18, "registrar" is broken down by its level4 destinations. Prior to the deposit deadline applicants are highly interested in "calendar." But after the deadline, applicants became more interested in "course-exam-schedules." This result demonstrates

that applicants find the academic calendar more of a resource in deciding to enroll or not. Conversely "course-exam-schedules" represents a page applicants view after enrolling, to explore in more detail the possible courses they might take as a Union student.

Figure 4.18



Note: Figure only shows level4 destinations with applicant views greater than 70

Within "finance" applicants mainly visit "student-accounts" from the beginning of the time plot up to the deposit deadline. Interestingly, the week leading up to the deposit deadline applicant views consistently increase each day. On the deposit deadline applicant views sharply decline and remain very low throughout the rest of the plot. This result indicates that "student-accounts" is also resource for applicants in deciding to enroll or not.

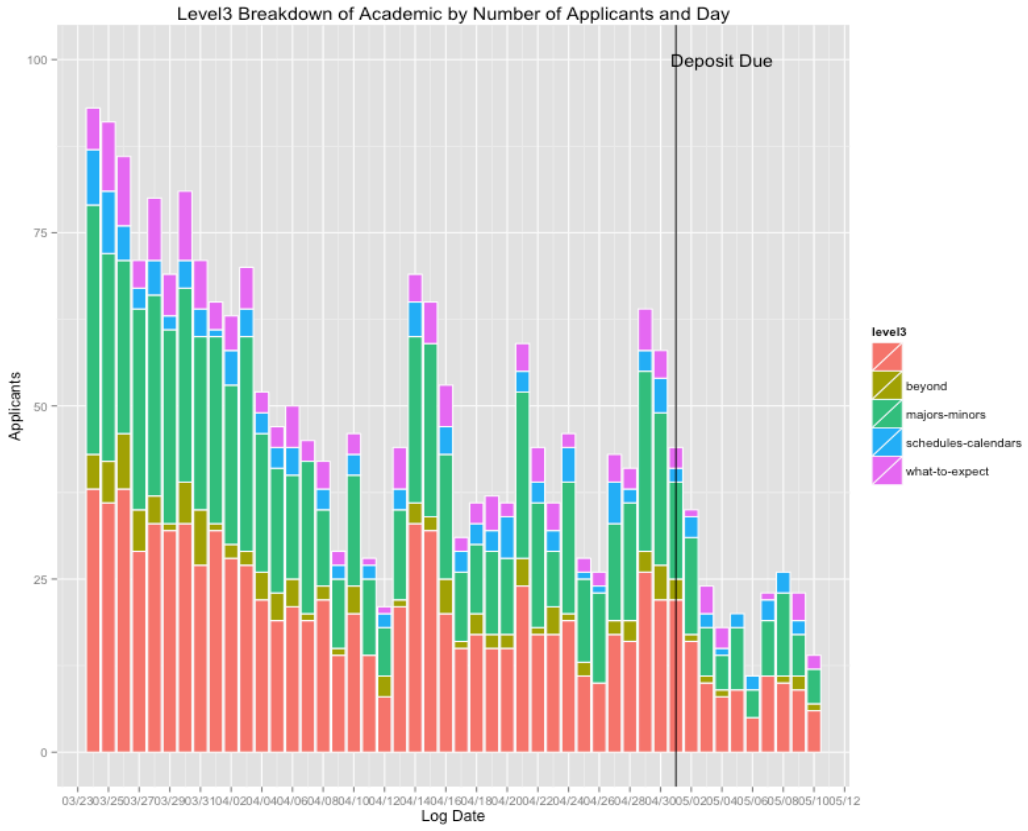
Within “dean” applicants visit “first-year” almost entirely in one period: week prior to the deposit deadline up to the last day in the time plot. Since non-enrolled applicants visit “first-year” just as much as enrolled applicants, the page could be a decision-making resource used directly before the deadline. But this is uncertain as enrolled applicants are directed to page after making online deposits.

The last level3 destination of focus is “scholars.” Interestingly the “scholars” page receives low levels of applicant views on all days except 3/24, 2/25, and 4/4. Since applicants of high academic standards are made aware of the scholars program, the visits on these dates could correspond to announcements made by the admissions office.

4.3.3. Academic

Figure 4.19 displays the level3 breakdown of “academic.”

Figure 4.19



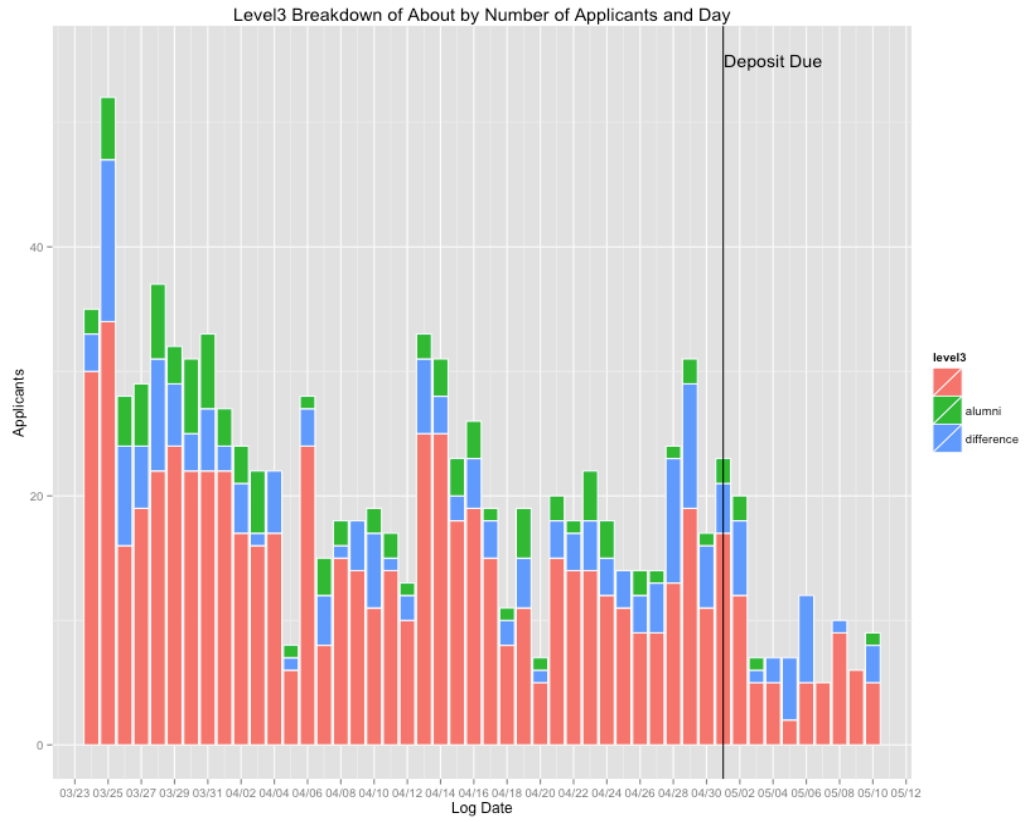
Note: Figure only shows level3 destinations with applicant views greater than 100

Figure 4.19 indicates that the "academic" homepage consistently receives views throughout the time plot. The "majors-minors" page, on the other hand, receives most of its views from the beginning of the time plot up to the deposit deadline. Applicant views of the "majors-minors" page also tend to increase directly prior to the deposit deadline. The sharp increase of views could demonstrate applicants finalizing if Union College meets their academic interests and desires. After the deposit deadline applicant views of "academic" decline significantly. This indicates the importance of "academic," specifically "majors-minors," in applicants' decision to enroll or not.

4.3.4. About

In Figure 4.20 the level3 destination breakdown of "about" is shown.

Figure 4.20



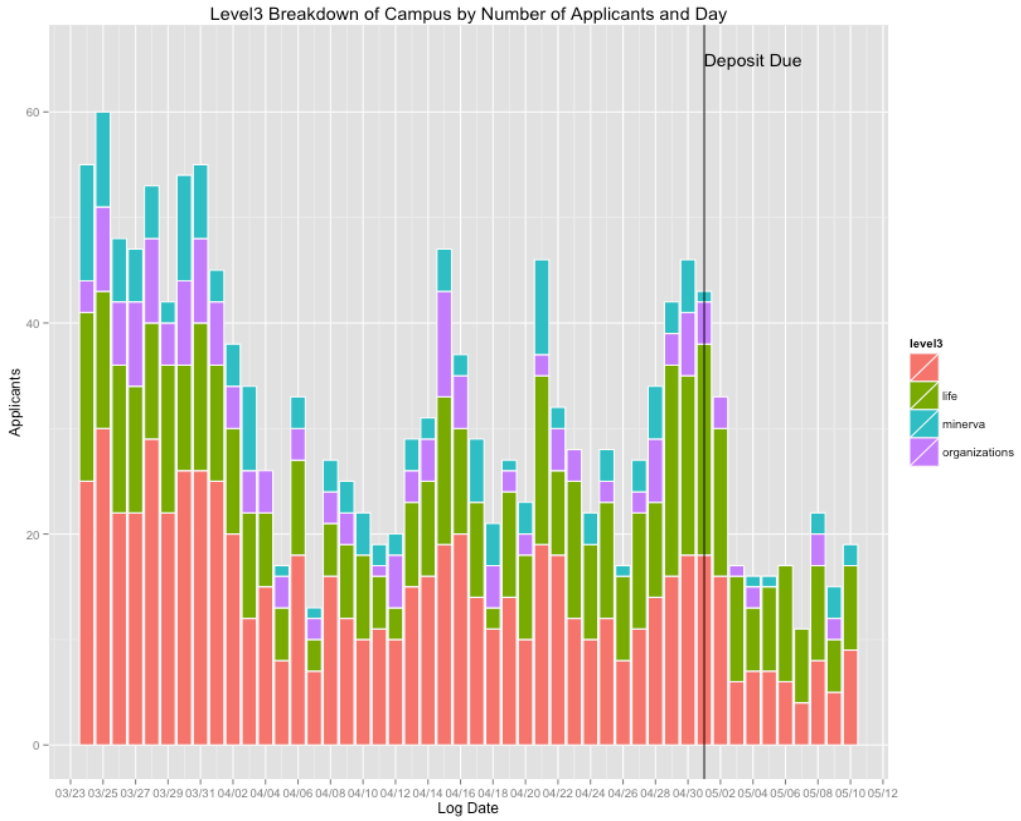
Note: Figure only shows level3 destinations with applicant views greater than 100

Applicants view the "about" homepage mostly from the beginning of the time plot up to the deposit deadline. Besides the homepage the only other level3 destination with noticeable applicant views is "alumni." Applicants visit "alumni" mainly after the announcement of the admission's decisions. This result indicates that applicants prefer learning about the college's successful alumni immediately after hearing of their admittance.

4.3.4. Campus

Figure 4.21 displays the level3 breakdown of the last level2 destination of focus, "campus."

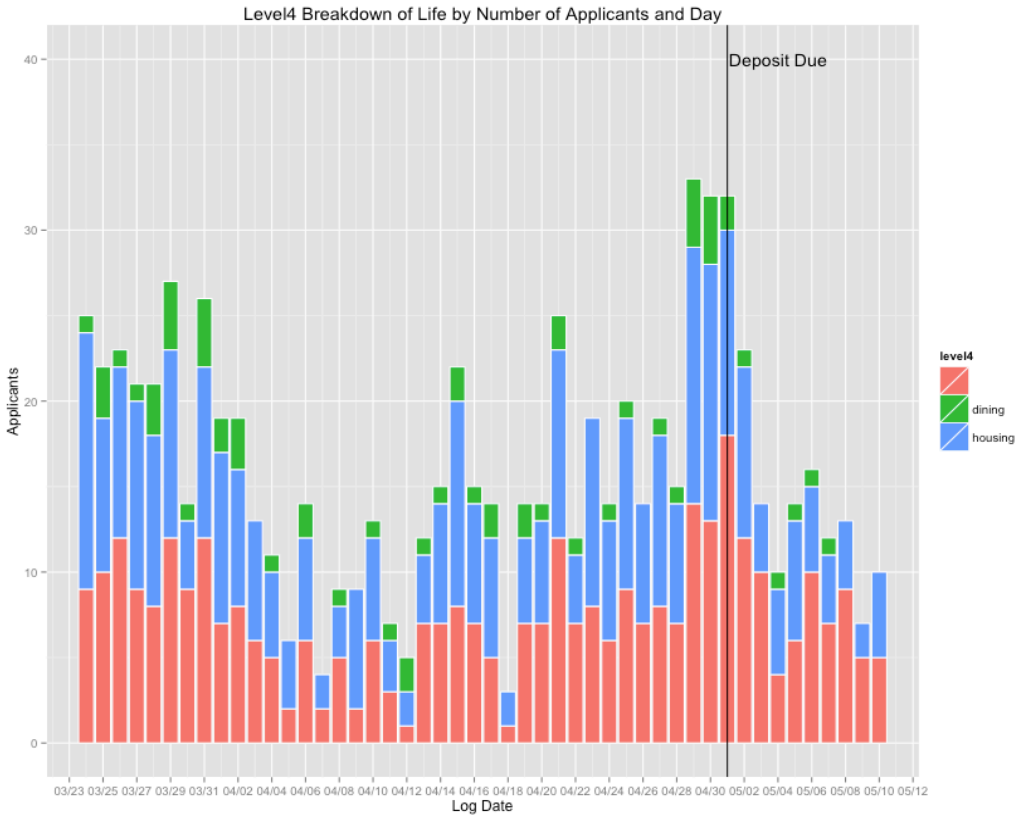
Figure 4.21



Note: Figure only shows level3 destinations with applicant views greater than 130

Applicants visit the "campus" homepage and "life" most frequently in the first 10 days of the plot and the week leading up to the deposit deadline. Specifically applicants view "housing" the most during these periods as seen in Figure 4.22. The concentration of applicant views in these periods demonstrates applicants' interest in the "campus" homepage and "housing" around announcement and decision periods.

Figure 4.22



4.4 Summary

Accepted applicants visit the Union College website in somewhat predicable patterns. There is a rush of visits following the release of admission decisions, another increase immediately prior to the deposit deadline, and a sharp drop in the number of visits thereafter. The destinations that students visit, in particular, the variation in the destinations over time paints a clear picture of what online resources students use. For

example, applicants visit “visit” and “finance” before but not after the deposit deadline. Similarly applicants visit “deposit” and “life” increasingly during the critical decision period.

The data also reveals some interesting surprises. Applicants view “registrar” the most of all the offices, but in a pattern: applicants view “calendar” before the deposit deadline and “course-exam-schedules” after. When applicants visit “life” they view “housing” substantially more than “dining” indicating applicants care more about campus housing than food. The last surprises deal with non-enrolled applicants. Contrary to prior belief non-enrolled applicants visit “deposit” and “first-year” just as much as enrolled applicants. Additionally, on average 47 non-enrolled applicants continue to visit the website each day after the deposit deadline.

Chapter 5

Does web traffic predict enrollment?

Each year Union College's Admission's office admits a specific number of regular decision applicants to fill the incoming class. Since it's not a guarantee every admitted applicant will enroll, the Admission's office must admit more students than it needs to fill the class. The Admission's office determines the number of applicants to admit using a predictive model. Traditional variables that go into such a predictive model include applicant characteristics like academic profiles, financial needs, and demographics. I test if digital engagement affects applicants' probability of enrollment to determine if digital engagement should be a part of the predictive model.

5.1. Data

I only test digital engagement up to the deposit deadline. For the predictive model to be helpful, I want to consider traffic that happens before applicants decide to enroll or not. Since few applicants only visit the website after the deadline, the number of observations is slightly lower than in preceding sections. Figure 5.1 shows the descriptive statistics of the traditional variables as well as the digital engagement variables.

In the first row of Figure 5.1 is the dummy variable *enrolled*. *Enrolled* equals one if an applicant enrolls. The average of *enrolled* is equivalent to the yield of accepted applicants, which I early show is 18.7%. This implies that if Union College admits 2,000 regular decision applicants, 374 applicants would enroll.

The traditional variables I use are institutional need (*instneed*), campus visit (*campusvisit*), and applicant rating (not shown in Figure 5.1). Institutional need shows

the financial aid amount the Admission's office determines an applicant needs (in thousands of dollars). The average institutional need of applicants in the study is around $14,000 – this includes applicants with zero financial need. Campus visit indicates if an applicant visits Union College's campus at some point before the deposit deadline. On average, 70% of Union College's applicants visit campus. Applicant rating (*aprat*) is a composite measure of an applicant's academic profile. Each applicant, depending on their applicant rating, is placed in one of four groups. Applicants with higher applicant ratings are put in *aprat4* while applicants with low applicant ratings are put in *aprat1*. Each *aprat* group holds approximately 268 applicants.

The digital engagement variables I use are *hits*, *days*, *dining*, *majors_minors*, *calendar*, *course_exam_schedules*, *first_year*, *what_to_expect*, *finance*, *finaid*, *deposit*, *clubs_organizations*, *residential_life*, *engineering*, *economics*, *managerial_economics*, *alumni*, and *minerva*. *Hits* refer to the number of page views an applicant accumulates throughout the time period. On average applicants accumulate approximately 68 page views in the time period. *Days* on the other hand refer to the number of days an applicant visits the Union website in the time period. On average applicants visit the website 5.4 days in the time plot. The remaining variables (destination variables) are dummies indicating whether or not an applicant visits a particular destination. If *housing* equals one the applicant of focus visited the level4 destination "housing" at least once in the time plot. The range of the destination dummies is from 0.3 for *majors_minors*, *finaid*, and *deposit* to 0.03 for *managerial_economics*.

Figure 5.1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
enrolled	1,056	0.2	0.4	0	1
instneed	1,056	14	19.7	0	70.3
campusvisit	1,056	0.7	0.5	0	1
hits	1,056	67.7	129.1	1	1,676
days	1,056	5.4	5.9	1	38
housing	1,056	0.2	0.4	0	1
dining	1,056	0.04	0.2	0	1
majors_minors	1,056	0.3	0.5	0	1
first_year	1,056	0.1	0.3	0	1
what_to_expect	1,056	0.1	0.3	0	1
finance	1,056	0.1	0.3	0	1
calendar	1,056	0.1	0.3	0	1
course_exam_schedules	1,056	0.1	0.3	0	1
finaid	1,056	0.3	0.4	0	1
deposit	1,056	0.3	0.4	0	1
clubs_organizations	1,056	0.1	0.3	0	1
residential_life	1,056	0.1	0.2	0	1
engineering	1,056	0.1	0.3	0	1
economics	1,056	0.1	0.2	0	1
managerial_economics	1,056	0.03	0.2	0	1
alumni	1,056	0.1	0.3	0	1
minerva	1,056	0.1	0.3	0	1

5.2. Empirical Results

5.2.1 Does visiting the website predict enrollment?

I first test five specifications using probit analysis to demonstrate the correlation of traditional variables affecting the probability of enrollment. In each of the specifications the dependent variable is *enrolled*. The independent variables are *instneed*, *campusvisit*, *aprat*, *hits*, and *days*. Figure 5.2 shows the results.

In the first specification I regress the *enrolled* dummy on *instneed*. The coefficient on *instneed* is positive and statistically significant. This indicates that applicants who require higher levels of financial aid are more likely to enroll. This is

expected as the lower the price of tuition incentivizes applicants to enroll. In the second specification I add *campusvisit* while controlling for *instneed*. The coefficient on *campusvisit* is also positive and statistically significant thus indicating that applicants are more likely to enroll if they visit Union College's campus. In the third specification, I add *aprat*. Since *aprat* groups applicants into one of four buckets based on their numeric applicant rating, probit analysis creates three *aprat* variables. Meaning *aprat2* compares applicants in the second quartile to those in the first. The results of the third specification indicate applicants with applicant ratings in the second, third and fourth quartiles are less likely to enroll than applicants with applicant ratings in the first quartile. This is to be expected as applicants with higher applicant rating likely have other options besides Union.

I incorporate general website engagement variables in the fourth and fifth specifications. In the fourth specification I add *days* while controlling for the traditional variables. The coefficient of *days* is positive and significant indicating that the more days applicants visit the website the more they are likely to enroll. In the fifth specification I incorporate *hits* while controlling for *days* and the traditional variables. The results demonstrate that the number of pages applicants visit is only marginally significant when controlling for *days*. This indicates that the number of days applicants are on the website is a better predictor of enrollment than the number of pages they view overall.

Figure 5.2: Regression Results - Traditional and Digital Engagement Variables

Dependent variable: enrolled

Variables	(1)	(2)	(3)	(4)	(5)
instneed	0.011***	0.011***	0.014***	0.014***	0.014***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
campusvisit		0.922***	0.936***	0.850***	0.856***
		(0.122)	(0.123)	(0.127)	(0.127)
aprat2			-0.229*	-0.254*	-0.233*
			(0.136)	(0.131)	(0.136)
aprat3			-0.517***	-0.464***	-0.451***
			(0.138)	(0.142)	(0.142)
aprat4			-0.357***	-0.356**	-0.358**
			(0.138)	(0.143)	(0.143)
days				0.058***	0.036**
				(0.008)	(0.014)
hits					0.001*
					(0.001)
Constant	-1.059***	-1.767***	-1.552***	-1.857***	-1.831***
	(0.058)	(0.118)	(0.134)	(0.147)	(0.148)
Log Likelihood	-497.99	-464.05	-456.562	-426.188	-424.412
Akaike Inf. Crit.	999.98	934.099	925.125	866.377	864.824

Note: *p<0.1; **p<0.05; ***p<0.01

The figure also reports AIC, a measure of the goodness of fit. The lower the AIC the better the fit of the model. In Figure 5.2, the AIC declines with each new specification indicating the improvement of fit. The decline of AIC has the greatest significance in the fourth specification. In the fourth specification, AIC decreases by 59 units when *days* is included in the model. This decline is significant in comparison to the 9 unit decline of AIC in the third specification when incorporating *aprat*. Therefore the decline of AIC in the fourth specification indicates that digital engagement variables significantly improve the predictive model.

While the fourth specification significantly improves the model, the fifth only does so slightly. Controlling for the traditional variables and *days*, the incorporation of

hits only declines AIC by 1.5 units. This indicates that *hits* does not add much explanatory power.

5.2.2 Does visiting a specific destination on the website predict enrollment?

I further investigate the effects of visiting a specific destination on the probability of enrollment. I test eighteen specifications retaining *enrolled* as the dependent variable. In each specification I control for the traditional variables and *days*. This allows me to determine if specific destinations affect the probability of enrollment holding all else constant. Figure 5.3 and Figure 5.4 display the results.

The results indicate that when controlling for *days,* visiting a specific destination is not a significant predictor with the exception of few destinations. These exceptions include the destinations corresponding to *what_to_expect*, *deposit*, *calendar*, *first_year*, *clubs_organizations*, and *alumni*. Each of the destination variables' coefficients are significant and positive indicating that the probability of enrollment increases each time an applicant views each destination, ceteris paribus.

In the final specification, I include the traditional variables, *days,* and all destination variables. Interestingly, the only destination variables that remain significant are *deposit* and *alumni*. It is no surprise *deposit* is significant as applicants use "deposit" to enroll online. But *alumni* is a different story. The only reason applicants would visit "alumni" is to use it as a resource. This result possibly indicates that applicants are more likely to enroll once learning about notable alumni and their accomplishments.

Figure 5.3: Regression Results – Destination Variables

Dependent Variable: enrolled

Variables	(1)	(2)	(3)	(4)	(5)	(6)
traditional variables	yes	yes	yes	yes	yes	yes
days	yes	yes	yes	yes	yes	yes
housing	0.239					
	(0.148)					
dining		0.004				
		(0.225)				
minerva			0.18			
			(0.154)			
majors_minors				0.005		
				(0.12)		
engineering					0.209	
					(0.148)	
economics						-0.009
						(0.196)
Constant	-1.859***	-1.857***	-1.852***	-1.858***	-1.852***	-1.852***
	(0.148)	(0.147)	(0.147)	(0.147)	(0.147)	(0.147)
Log Likelihood	-424.897	-426.188	-425.522	-426.188	-425.219	-426.187
Akaike Inf. Crit.	865.794	868.376	867.043	868.375	866.439	868.375

Note: *p<0.1; **p<0.05; ***p<0.01
Estimated coefficients of *days* and traditional variables not shown, but still in regression equation

Dependent Variable: enrolled

Variables	(7)	(8)	(9)	(10)	(11)	(12)
traditional variables	yes	yes	yes	yes	yes	yes
days	yes	yes	yes	yes	yes	yes
managerial_economics	0.135					
	(0.24)					
what_to_expect		0.289*				
		(0.158)				
finaid			-0.174			
			(0.125)			
deposit				0.418***		
				(0.114)		
finance					0.003	
					(0.151)	
calendar						0.334**
						(0.15)
Constant	-1.851***	-1.868***	-1.852***	-1.894***	-1.857***	-1.853***
	(0.147)	(0.148)	(0.147)	(0.149)	(0.147)	(0.148)
Log Likelihood	-426.035	-424.549	-425.214	-419.671	-426.188	-423.731
Akaike Inf. Crit.	868.07	865.098	866.428	855.342	868.376	863.462

Note: *p<0.1; **p<0.05; ***p<0.01
Estimated coefficients of *days* and traditional variables not shown, but still in regression equation

Figure 5.4: Regression Results – Traditional, Digital Engagement, and Destination Variables

Dependent Variable: enrolled

Variables	(13)	(14)	(15)	(16)	(17)	(18)
traditional variables	yes	yes	yes	yes	yes	yes
days	yes	yes	yes	yes	yes	yes
housing						0.002
						(0.18)
dining						-0.286
						(0.248)
minerva						0.046
						(0.175)
majors_minors						-0.1
						(0.144)
engineering						0.217
						(0.172)
economics						-0.338
						(0.34)
managerial_economics						0.28
						(0.409)
what_to_expect						0.17
						(0.178)
finaid						-0.269*
						(0.146)
deposit						0.384***
						(0.12)
finance						0.027
						(0.181)
calendar						0.178
						(0.166)
course_exam_schedules	0.231					0.165
	(0.186)					(0.202)
first_year		0.429**				0.281
		(0.181)				(0.203)
clubs_organizations			0.265*			0.106
			(0.157)			(0.173)
residential_life				0.306		0.221
				(0.198)		(0.231)
alumni					0.431**	0.443**
					(0.172)	(0.185)
Constant	-1.857***	-1.843***	-1.845***	-1.851***	-1.865***	-1.856***
	(0.148)	(0.147)	(0.147)	(0.148)	(0.148)	(0.152)
Log Likelihood	-425.429	-423.466	-424.816	-424.987	-423.098	-408.447
Akaike Inf. Crit.	866.858	862.932	865.632	865.974	862.197	864.894

Note: *p<0.1; **p<0.05; ***p<0.01

Estimated coefficients of *days* and traditional variables not shown, but still in regression equation

Chapter 6

Conclusions

Overall, my results suggest that Union College accepted applicants use the college's website as a major resource when finalizing their choice for enrollment. This conclusion is based on the increase in applicant website traffic immediately following the communication of admittance up to the deposit deadline. An analysis of the surge in traffic shows that applicants use the website as a resource for a wide range of college information including information on admissions, offices, academics, and campus life. Applicants' traffic indicates a strong interest in the information on "visit," "calendar," "majors-minors," and "life." Knowing the destinations accepted applicants most commonly view provides Union College with an opportunity to focus their efforts and optimize their messaging. By doing so, the website could be more effective in enhancing applicants' desire to enroll.

In terms of using digital engagement analysis to predict enrollment, my results indicate that the frequency of website use is a highly predictive factor. In addition, the analysis shows that applicant views of specific destinations, "alumni" in particular, can also be a highly predictive factor. These observations along with other patterns found in the study demonstrate that digital engagement can help predict enrollment. This linkage would be especially useful to the Admission's office if a similar study were undertaken that analyzes website traffic in the period preceding the decision of admittance. This analysis could provide the Admissions office an opportunity to be more selective with the applicant pool.

These conclusions are subject to a number of limitations. First, the sample of applicants in the study only represents 65% of total enrolled applicants and 59% of total non-enrolled applicants. In the process of matching IP addresses to IDs in the applicant portal, many IP addresses corresponded to more than one id. This is but one of few cases where observations needed to be removed. Moreover, if applicants use a different device to browse the website than the device used to log into the applicant portal, their traffic is not in the data. Thus patterns associated with applicant views and probability of enrollment might not reflect the behavior of all applicants. Secondly, the direction of causality between enrollment and website traffic cannot be determined. In order for future research to determine causality, researches would need to run experiments on the website. This could be done by randomly switching website design or content on certain days and testing the difference of probability of enrollment. Thirdly, the traffic of applicants does not take into consideration when enrolled applicants submit their deposit. While most enrollees submit their deposit close to the deadline, a closer examination of behavior immediately prior to depositing could be useful.

Works Cited

Abrahamson, T. "Life and death on the Internet: To web or not to web is no longer the question." *The Journal of College Admissions 168* (2000): 6-11.

Black, Elizabeth L. "Web Analytics: A Picture of the Academic Library Web Site User." *Journal of Web Librarianship* 3.1 (2009): 3-14. Web. 17 Nov. 2014. <https://kb.osu.edu/dspace/bitstream/handle/1811/46648/BlackE_JournalWebLibrarianship_2009_v3n1_p3-14.pdf?sequence=1>.

Bosnjak, Sasa, Mirjana Maric, and Zita Bosnjak. "The Role of Web Usage Mining in Web Applications Evaluation." *Management Information Systems* 5, no. 1 (2010): 31-36. Accessed November 17, 2014. http://www.ef.uns.ac.rs/mis/archive-pdf/2010%20-%20No1/MIS2010_1_5.pdf.

Chen, Hsinchun, Roger H. Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36, no. 4 (December 2012): 1165-88. ACM Digital Library. Accessed November 17, 2014. http://dl.acm.org/citation.cfm?id=2481683.

Coffin, Jonathan C. "Telling the Story of a Liberal Arts College: The College Website as a Window on Institutional Positioning." M.A. thesis, Johns Hopkins University, 2012. http://advanced.jhu.edu/wp-content/uploads/2013/04/Jonathan_Coffin_sm2.pdf.

Farney, Tabatha, and Nina McHale. *Web Analytics Strategies for Information Professionals: A LITA Guide*. N.p.: ALA TechSource, 2014. *EBrary*. Web. 17 Oct. 2014. <http://site.ebrary.com/lib/unioncollege/detail.action?docID=10788467>.

Goncalves, Bruno, and Jose J. Ramasco. "Human Dynamics Revealed Through Web Analytics." *Physical Review E* 78 (August 26, 2008). Accessed December 17, 2014. http://journals.aps.org/pre/abstract/10.1103/PhysRevE.78.026123.

Harden, Lalend, and Bob Heyman. *Digital Engagement: Internet Marketing That Captures Customers and Builds Intense Brand Loyalty*. New York: Amacom, 2009. http://site.ebrary.com/lib/unioncollege/reader.action?docID=10292221.

Palavitsinis, Nikos, Vassilios Protonotarios, and Nikos Manouselis. "Applying Analytics for a Learning Portal: the Organic.Edunet Case Study." *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (2011). ACM Digital Library. Accessed November 17, 2014. http://dl.acm.org/citation.cfm?id=2090138.

Poock, Michael A., and Dennis Lefond. "How College-Bound Prospects Perceive University Websites: Findings, Implications and Turning Browsers into Applicants." *College & University Journal* (2001): 15-21. http://www.mstoner.com/wp-content/uploads/old/Poock&Lefond.pdf.

Wind, Jerry, and Vijay Mahajan. *Digital Marketing : Global Strategies from the World's Leading Experts*. New York: John Wiley & Sons, 2001. http://site.ebrary.com/lib/unioncollege/reader.action?docID=10001754.

Appendices: R Programming

Appendix A: Input and Clean Data

```
setwd("/Users/chrisgaribaldi/Documents/RStudio")
library(dplyr) #new package for manipulating datasets
library(tidyr) #new package for reshaping datasets, requires R3.1
library(lubridate) #needed for rounding dates
library(stringr) #for manupulating strings
library(ggplot2)
library(reshape2)
library(scales)

#**************************Portal Input*************************
#Input data into R
portal <- read.table("portallog.csv",header = TRUE, sep=",")
portal <- portal[c("id","ip_address")] #keep only id and ip_address
portal <- portal[!duplicated(portal),] #remove any duplicate observations
#11K ip addresses, and only 4095 ids, so students access the portal from lots of different
computers
#12K obs so there are ip addressess associated with more than one id
portal <- arrange(portal, ip_address, id) #sort by ip and id
portal <- portal %>%
  group_by(ip_address) %>%
  mutate(no_ids=n()) %>% #calculates number of id's  used by each ip address
  ungroup() %>%
  group_by(id) %>%
  mutate(no_ips=n()) #calculates number of ip's used by each id

#this is just to display some counts of different values of no_ids and no_ips
table(portal$no_ips) #10941 ip's were used by just one id, but 278 ips were used by two
id's and there is even 2 ips that were used by 12 ids
table(portal$no_ids) #1414 id's use just one ip, 1086 id's used 2 ips, 613 used 3ips, 326
used 4ips, 225 used 5, 114 used 6, 72 used 7

portal <- filter(portal, no_ids==1) #keep only ips that have just one id associated with it
portal <- select(portal, ip_address, id)

#display number of id's
length(unique(portal$id))  #3936 different id's, the number of obs is much bigger because
many id's use more than one ip

#**************************Traffic Input*************************
traffic <- read.table("March-April 2014 traffic only portal ip addresses.txt", header=
FALSE, sep = " ", na.strings = "-", fill = TRUE, col.names=c('ip_address','servername',
```

```
'remoteuser', 'logdate', 'timezone', 'request', 'status',  'responsesize', 'referrer', 'useragent',
'cookie'))

traffic <- traffic[c("ip_address","logdate","request")] #keep ip date and what they
requested
#there are 180K obs
#there are 4271 ip addresses, in traffic

traffic <- traffic[!duplicated(traffic),] #remove duplicates
traffic$logdate <- substr(traffic$logdate, start=2,stop=100) #remove the first character in
logdate because it is "["
traffic$logdate <- as.POSIXct(traffic$logdate, format="%d/%b/%Y:%H:%M:%S") #turn
logdate into time variable
traffic$logdate <- round_date(traffic$logdate, "day") #create a new variable that just has
date
traffic$logdate <- as.Date(traffic$logdate,format="%m/%d") #makes it class Date as
opposed to character, and as opposed to POSICTtraffic$logday <-
format(traffic$logday,"%m/%d")

traffic <- filter(traffic,logdate!="2014-05-11") #drop that last day
traffic <- filter(traffic,logdate!="2014-03-23") #drop that first day - because we are not
sure we have the full day

#create a character variable logday - sometimes this works nicer with graphs
traffic$logday <- format(traffic$logdate, "%m/%d")

#**************************Traffic Clean Up**********************
#Turn requests into string class
traffic$request <- as.character(traffic$request)
#a little trick to make sure that all destinations have bunch of levels
traffic$request <- paste(traffic$request, "/ / /")

#Split the requests by levels
traffic$levels <- strsplit(traffic$request,"/")
traffic$level1 <-unlist(lapply(traffic$levels,'[[',1))
traffic$level2 <-unlist(lapply(traffic$levels,'[[',2))
traffic$level3 <-unlist(lapply(traffic$levels,'[[',3))
traffic$level4 <-unlist(lapply(traffic$levels,'[[',4))

#****Clean Up Levels
traffic$level2 <- sub("HTTP","",traffic$level2)
traffic$level3 <- sub("HTTP","",traffic$level3)
traffic$level4 <- sub("HTTP","",traffic$level4)
#also get rid of 1.1 and replace with blank
traffic$level2 <- sub("1.1","",traffic$level2)
traffic$level3 <- sub("1.1","",traffic$level3)
```

```
traffic$level4 <- sub("1.1","",traffic$level4)
#also get rid of 1.0 and replace with blank
traffic$level2 <- sub("1.0","",traffic$level2)
traffic$level3 <- sub("1.0","",traffic$level3)
traffic$level4 <- sub("1.0","",traffic$level4)

#the function below gets rid of these trailing blanks
traffic$level2 <- gsub("^\\s+|\\s+$", "", traffic$level2)
traffic$level3 <- gsub("^\\s+|\\s+$", "", traffic$level3)
traffic$level4 <- gsub("^\\s+|\\s+$", "", traffic$level4)

#Convert all levels into lower case letters
traffic$level2 <- tolower(traffic$level2)
traffic$level3 <- tolower(traffic$level3)
traffic$level4 <- tolower(traffic$level4)

#Drop some of the really long levels/destinations (makes it easier to view the table)
#nchar is function that counts the number of characters in the variable
traffic <- filter(traffic, nchar(level2)<60)
traffic <- filter(traffic, nchar(level3)<60)
traffic <- filter(traffic, nchar(level4)<60)

#Drop level2 destinations - favicon & _banners
traffic <- filter(traffic,level2!="favicon.ico")
traffic <- filter(traffic,level2!="_banners")

#Replace level3 & level4 destination "index.php" with " "
traffic$level3 <- sub("index.php","",traffic$level3)
traffic$level4 <- sub("index.php","",traffic$level4)

#drop the list 'levels' otherwise dplyr does not work
traffic <- select(traffic, -levels)
#create a new variable that has number of obs in each group defined by level2, let's call it
hits2
traffic <- traffic %>% group_by(level2) %>% mutate(hits2=n())
traffic <- traffic %>% group_by(level3) %>% mutate(hits3=n())
traffic <- traffic %>% group_by(level4) %>% mutate(hits4=n())

##******Dummy variables to determine types of destination views of applicants
traffic$housing <- ifelse(str_locate(traffic$level4, "housing")[,1]=="NA",0,1)  #this puts
1 if request contains "housing", and NA otherwise
traffic$housing[is.na(traffic$housing)] <-0 #this puts 0 where it had NA

#dining -- life
traffic$dining <- ifelse(str_locate(traffic$level4, "dining")[,1]=="NA",0,1)  #this puts 1 if
request contains "housing", and NA otherwise
```

```
traffic$dining[is.na(traffic$dining)] <-0 #this puts 0 where it had NA

#majors-minors -- academic
traffic$majors_minors <- ifelse(str_locate(traffic$level3, "majors-
minors")[,1]=="NA",0,1)  #this puts 1 if request contains "housing", and NA otherwise
traffic$majors_minors[is.na(traffic$majors_minors)] <-0 #this puts 0 where it had NA

#calendar -- registrar
traffic$calendar <- ifelse(str_locate(traffic$request, "registrar/calendar")[,1]=="NA",0,1)
#this puts 1 if request contains "housing", and NA otherwise
traffic$calendar[is.na(traffic$calendar)] <-0 #this puts 0 where it had NA

#course-exam-schedules -- registrar
traffic$course_exam_schedules <- ifelse(str_locate(traffic$level4, "course-exam-
schedules")[,1]=="NA",0,1)  #this puts 1 if request contains "housing", and NA otherwise
traffic$course_exam_schedules[is.na(traffic$course_exam_schedules)] <-0 #this puts 0
where it had NA

#first-year -- dean
traffic$first_year <- ifelse(str_locate(traffic$request, "dean/first-year")[,1]=="NA",0,1)
#this puts 1 if request contains "housing", and NA otherwise
traffic$first_year[is.na(traffic$first_year)] <-0 #this puts 0 where it had NA

#what-to-expect -- academic
traffic$what_to_expect <- ifelse(str_locate(traffic$request, "academic/what-to-
expect")[,1]=="NA",0,1)  #this puts 1 if request contains "housing", and NA otherwise
traffic$what_to_expect[is.na(traffic$what_to_expect)] <-0 #this puts 0 where it had NA

#finance
traffic$finance <- ifelse(str_locate(traffic$level3, "finance")[,1]=="NA",0,1)  #this puts 1
if request contains "housing", and NA otherwise
traffic$finance[is.na(traffic$finance)] <-0 #this puts 0 where it had NA

#finaid -- admissions
traffic$finaid <- ifelse(str_locate(traffic$level3, "finaid")[,1]=="NA",0,1)  #this puts 1 if
request contains "housing", and NA otherwise
traffic$finaid[is.na(traffic$finaid)] <-0 #this puts 0 where it had NA

#deposit-- admissions
traffic$deposit <- ifelse(str_locate(traffic$level3, "deposit")[,1]=="NA",0,1)  #this puts 1
if request contains "housing", and NA otherwise
traffic$deposit[is.na(traffic$deposit)] <-0 #this puts 0 where it had NA

#clubs-organizations -- student-activities -- offices
```

```
traffic$clubs_organizations <- ifelse(str_locate(traffic$level4, "clubs-
organizations")[,1]=="NA",0,1)  #this puts 1 if request contains "housing", and NA
otherwise
traffic$clubs_organizations[is.na(traffic$clubs_organizations)] <-0 #this puts 0 where it
had NA

#residential life -- offices
traffic$residential_life <- ifelse(str_locate(traffic$level3, "residential-
life")[,1]=="NA",0,1)  #this puts 1 if request contains "housing", and NA otherwise
traffic$residential_life[is.na(traffic$residential_life)] <-0 #this puts 0 where it had NA

#engineering -- academic
traffic$engineering <- ifelse(str_locate(traffic$level4, "engineering")[,1]=="NA",0,1)
#this puts 1 if request contains "housing", and NA otherwise
traffic$engineering[is.na(traffic$engineering)] <-0 #this puts 0 where it had NA

#economics -- academic
traffic$economics <- ifelse(str_locate(traffic$level4, "economics")[,1]=="NA",0,1)  #this
puts 1 if request contains "housing", and NA otherwise
traffic$economics[is.na(traffic$economics)] <-0 #this puts 0 where it had NA

#managerial-economics -- academic
traffic$managerial_economics <- ifelse(str_locate(traffic$level4, "managerial-
economics")[,1]=="NA",0,1)  #this puts 1 if request contains "housing", and NA
otherwise
traffic$managerial_economics[is.na(traffic$managerial_economics)] <-0 #this puts 0
where it had NA

#alumni -- about
traffic$alumni <- ifelse(str_locate(traffic$request, "about/alumni")[,1]=="NA",0,1)  #this
puts 1 if request contains "housing", and NA otherwise
traffic$alumni[is.na(traffic$alumni)] <-0 #this puts 0 where it had NA

#minerva -- campus
traffic$minerva <- ifelse(str_locate(traffic$level3, "minerva")[,1]=="NA",0,1)  #this puts
1 if request contains "housing", and NA otherwise
traffic$minerva[is.na(traffic$minerva)] <-0 #this puts 0 where it had NA


##**********Merge Traffic with Portal
traffic <- merge(traffic,portal,"ip_address") #merges in portal data, keeps only
ip_addresses that are in BOTH datasets

#keep just id logday, level2 (word), hits2
traffic <- select(traffic, id, logday,logdate, level2, hits2, level3, hits3,level4,hits4,
housing, dining, majors_minors, calendar, course_exam_schedules, first_year,
```

```
what_to_expect, finance, finaid, deposit, clubs_organizations, residential_life,
engineering, economics, managerial_economics, alumni, minerva)

#create a dataset that just has the id's that are in traffic
ids_in_traffic <- traffic %>%
  group_by(id) %>%
  summarize(x=n())

#****************************Yield Input************************
yield <- read.csv("yield.csv", header=TRUE)
yield <- select(yield, id, admitted, enrolled, paiddate, applicantrating, instneed,
campusvisit)   #keep only the variables we need from yield

yield <- yield[yield$admitted=="Y",] #keep only people who were admitted
yield <-  select(yield,-admitted) #drop admitted variable

yield$enrolled <-ifelse(yield$enrolled=="Y",1,0) #create an enrolled dummy --> enrolled
= 1
yield$campusvisit <-ifelse(yield$campusvisit=="Y",1,0) #create a campus visit dummy -
-> Y = 1

yield$id <- as.character(yield$id)
yield$paiddate <- as.Date(yield$paiddate,format="%m/%d/%y") #makes it class Date as
opposed to character, and as opposed to POSICTtraffic$logday <-
format(traffic$logday,"%m/%d")

#Merge yield with ids in traffic
yield <- merge(yield, ids_in_traffic, by="id")
yield <- select(yield, -x)
yield$aprat <- as.factor(ntile(yield$applicantrating, 4)) #create a factor that identifies
wich quartile group the student falls in terms of his or her applicant rating

#****************************Traffic & Yield Merge*****************
MainData <- merge(traffic,yield, by="id")
length(unique(MainData$id))
save(MainData,file="MainData.Rda")
```

Appendix B: Plot and Analyze Data

```
#Plot and Analyze Data

setwd("/Users/chrisgaribaldi/Documents/RStudio")
library(dplyr) #new package for manipulating datasets
library(tidyr) #new package for reshaping datasets, requires R3.1
library(lubridate) #needed for rounding dates
library(stringr) #for manupulating strings
library(ggplot2)
library(reshape2)
library(scales)
load("MainData.Rda")
MainData$enrolled <- as.factor(MainData$enrolled)
#Plotting, Playing with, and Analyzing MainData


#******TIME PLOTS********TIME PLOTS********TIME PLOTS********TIME
#ORDER
#1) No Fill - Students & Page Views
#2) Enrolled vs. Not Enrolled - Students & Page Views
#3) Applicant Rating - Students & Page Views
#**********************************************************************
#Collapse MainData for Students (no fill) & Enrolled & Applicant Rating
MainData_col <- MainData %>%
  group_by(logday,logdate, id, enrolled, aprat) %>%
  summarize(hits234=n()) %>%
  ungroup()

#**********************************Students Over Time - No Fill************
#Plot of accepted student views by student count without fill
temp_plotSTUD <- MainData_col %>%
  group_by(logday,logdate) %>%
  summarize(hits=sum(hits234),students=n())

#Plot - students over time
ggplot(temp_plotSTUD, aes(x=logdate, y=students))+geom_bar(binwidth=1,
colour="white", stat="identity")+scale_x_date(labels = date_format("%m/%d"), breaks =
date_breaks("2
days"),limits=c(as.Date("2014/03/24"),as.Date("2014/05/10")))+labs(y="Applicants",
x="Log Date", title="Number of Applicants Visiting Website by Day") +
geom_vline(xintercept=as.numeric(as.Date("2014/05/01"))) + annotate("text",
x=as.Date("2014-05-04"), y=400 , label = "Deposit Due")

#*****************Enrolled vs. Not Enrolled - Students **********************
```

```
#Students & Page Views
temp_plot1 <- MainData_col %>%
  group_by(logday,logdate,enrolled) %>%
  summarize(hits=sum(hits234),students=n())

#Plot - Students
ggplot(temp_plot1, aes(x=logdate, y=students, fill=enrolled))+geom_bar(binwidth=1,
colour="white", stat="identity")+scale_x_date(labels = date_format("%m/%d"), breaks =
date_breaks("2
days"),limits=c(as.Date("2014/03/24"),as.Date("2014/05/10")))+labs(y="Applicants",
x="Log Date", title="Number of Applicants Visiting Website by Day and Enrolled
Status") + geom_vline(xintercept=as.numeric(as.Date("2014/05/01"))) + annotate("text",
x=as.Date("2014-05-04"), y=400 , label = "Deposit Due")

#Table for Analyzing

temp_plot11 <- recast(temp_plot1,logday + logdate ~ variable + enrolled, id.var =
c("logday","logdate","enrolled"))
temp_plot11[is.na(temp_plot11)] <-0 #replace NAs with zeros (the NAs happen if for a
paricular level23 only one group of students(enrolledor not renrolled) did not hit there)
#calculate aggregates, columns that have totals for enrolled and not-enrolled
temp_plot11 <- temp_plot11 %>%

mutate(tot_hits=hits_0+hits_1,tot_stud=students_0+students_1,hitsEnrol_share=(hits_1/t
ot_hits),studEnrol_share=(students_1/tot_stud)) %>%
  arrange(logdate)

#****DESTINATION HORIZONTAL PLOTS*****DESTINATION HORIZONTAL
PLOTS*****
#ORDER
#1) Total - no fill
#2) Enrolled Status fill per top5 level2 for both Students & Page Views
#3) Applicant Rating fill per top 5 level2 for both Students & Page Views
#*****************************************************************

#*****************************Enrolled vs. Not Enrolld - Students*************

#Enrolled vs. Not Enrolled
Destinations_AccApp <- MainData %>%
  group_by(level2,id,enrolled) %>%
  summarize(hits=n()) %>% #hits for each id per level2 per day
  group_by(level2,enrolled)%>%
  summarize(hits=sum(hits),students=n())%>% #total hits per level2 per day, total
number of students per level2 by day
  group_by(level2)%>%
```

```
  mutate(tot_hits=sum(hits),tot_students=sum(students))%>% # total hits and students
across entire period (**tot_students will double count students b/c visits on multiple days
are added together)
  ungroup() %>%
  arrange(desc(tot_hits))

#********Plot = Students - No fill
#In subset made enrolled == 1 --> this helped eliminate the problem of adding tot_stud
together for enrolled and non enrolled applicants
bar_tot_stud <- ggplot(subset(Destinations_AccApp,tot_students>200 & enrolled==1),
aes(x=reorder(level2, tot_students),y=tot_students))
bar_tot_stud + geom_bar(stat="identity", aes(order=desc(tot_students))) + coord_flip()+
labs(x="Web Destinations", y="Applicants", title="Number of Applicants by Web
Destinations")

#********Plot = Enrolled vs. Not Erolled - Students
#here I can plot the number of enrollend and not enrolled students that hit upon each
level23 destination
bar_stud_enrol <- ggplot(subset(Destinations_AccApp,tot_students>200),
aes(x=reorder(level2, tot_students),y=students, fill=enrolled))
bar_stud_enrol + geom_bar(stat="identity", aes(order=desc(enrolled))) + coord_flip() +
labs(x="Web Destinations", y="Applicants", title="Number of Applicants by Web
Destinations and Enrolled Status")

#Table for Analyzing
Destinations_AccApp <- select(Destinations_AccApp,-tot_hits,-tot_students)
Destinations_AccApp1 <- recast(Destinations_AccApp, level2 ~ variable + enrolled,
id.var = c("level2","enrolled"))
Destinations_AccApp1 [is.na(Destinations_AccApp1 )] <-0
#calculate aggregates, columns that have totals for enrolled and not-enrolled
Destinations_AccApp1 <- Destinations_AccApp1 %>%

mutate(tot_hits=hits_0+hits_1,tot_stud=students_0+students_1,hitsEnrol_share=(hits_1/t
ot_hits),studEnrol_share=(students_1/tot_stud)) %>%
  arrange(desc(tot_stud))


#*****************************Level3 Breakdowns**********************

#*********Overall & Enrolled fill***************
#Admissions level 3 breakdown
#No Fill & Enrolled Status
Destinations_AccApp <- subset(MainData,level2=="admissions") %>%
  group_by(level3,id,enrolled) %>%
  summarize(hits=n()) %>%
  group_by(level3,enrolled)%>%
```

```
  summarize(hits=sum(hits),students=n())%>%
  group_by(level3)%>%
  mutate(tot_hits=sum(hits),tot_students=sum(students))%>%
  ungroup() %>%
  arrange(desc(tot_students))

#********Plot = Admissions level3 breakdown by students
ggplot(subset(Destinations_AccApp,tot_students>100), aes(x=reorder(level3,
tot_students),y=students)) + geom_bar(stat="identity") + coord_flip() + labs(x="Web
Destinations", y="Applicants", title="Admissions Level3 Breakdown by Applicants")

#********Plot = Admissions level3 breakdown by students & enrolled status
ggplot(subset(Destinations_AccApp,tot_students>100), aes(x=reorder(level3,
tot_students),y=students, fill=enrolled)) + geom_bar(stat="identity",
aes(order=desc(enrolled))) + coord_flip() + labs(x="Web Destinations", y="Applicants",
title="Admissions Level3 Breakdown by Applicants and Enrolled Status")

#Table for Analyzing
Destinations_AccApp <- select(Destinations_AccApp,-tot_hits,-tot_students)
Destinations_AccApp1 <- recast(Destinations_AccApp, level3 ~ variable + enrolled,
id.var = c("level3","enrolled"))
Destinations_AccApp1 [is.na(Destinations_AccApp1 )] <-0
#calculate aggregates, columns that have totals for enrolled and not-enrolled
Destinations_AccApp1 <- Destinations_AccApp1 %>%

mutate(tot_hits=hits_0+hits_1,tot_stud=students_0+students_1,hitsEnrol_share=(hits_1/t
ot_hits),studEnrol_share=(students_1/tot_stud)) %>%
  arrange(desc(tot_stud))
```

#**NOTE: REPEAT PROCESS FOR OTHER LEVEL3 DESTINATIONS**

```
#******************Level4 Drill Down by Visit
Destinations_AccApp <- subset(MainData,level3=="visit") %>%
  group_by(level4,id,enrolled) %>%
  summarize(hits=n()) %>%
  group_by(level4,enrolled)%>%
  summarize(hits=sum(hits),students=n())%>%
  group_by(level4)%>%
  mutate(tot_hits=sum(hits),tot_students=sum(students))%>%
  ungroup() %>%
  arrange(desc(tot_students))

#********Plot = Visit level4 breakdown by students
ggplot(subset(Destinations_AccApp,tot_students>1), aes(x=reorder(level4,
tot_students),y=students)) + geom_bar(stat="identity") + coord_flip() + labs(x="Web
Destinations", y="Applicants", title="Visit Level4 Breakdown by Applicants")
```

```
#********Plot = Visit level4 breakdown by students & enrolled status
ggplot(subset(Destinations_AccApp,tot_students>1), aes(x=reorder(level4,
tot_students),y=students, fill=enrolled)) + geom_bar(stat="identity",
aes(order=desc(enrolled))) + coord_flip() + labs(x="Web Destinations", y="Applicants",
title="Visit Level4 Breakdown by Applicants and Enrolled Status")

#Table for Analyzing
Destinations_AccApp <- select(Destinations_AccApp,-tot_hits,-tot_students)
Destinations_AccApp1 <- recast(Destinations_AccApp, level4 ~ variable + enrolled,
id.var = c("level4","enrolled"))
Destinations_AccApp1 [is.na(Destinations_AccApp1 )] <-0
#calculate aggregates, columns that have totals for enrolled and not-enrolled
Destinations_AccApp1 <- Destinations_AccApp1 %>%

mutate(tot_hits=hits_0+hits_1,tot_stud=students_0+students_1,hitsEnrol_share=(hits_1/t
ot_hits),studEnrol_share=(students_1/tot_stud)) %>%
  arrange(desc(tot_stud))
```

#**NOTE: REPEAT PROCESS FOR OTHER LEVEL4 DESTINATIONS**

```
#****DESTINATION BY TIME*****DESTINATION BY TIME**************
#ORDER
#1) Top 5 level2 - Students & Page Views
#2) Level3 breakdowns - Students & Page Views - 5 per
#*************************************************************************

#********************Top 5 Level2 Destinations - Student & Page Views******

#Level2 Collapse
Destinations_AccApp <- MainData %>%
  group_by(logdate,level2,id) %>%
  summarize(hits=n()) %>% #hits for each id per level2 per day
  group_by(logdate,level2)%>%
  summarize(hits=sum(hits),students=n())%>% #total hits per level2 per day, total
number of students per level2 by day
  group_by(level2)%>%
  mutate(tot_hits=sum(hits),tot_students=sum(students))%>% # total hits and students
across entire period (**tot_students will double count students b/c visits on multiple days
are added together)
  ungroup() %>%
  arrange(desc(tot_students))

#Filter out "applications" and homepage from graphs
Destinations_AccApp <- filter(Destinations_AccApp,level2 !="")
```

```
Destinations_AccApp <- filter(Destinations_AccApp,level2 != "applications")


#Plot - Student
ggplot(subset(Destinations_AccApp,tot_students>700), aes(x=logdate, y=students,
fill=level2))+geom_bar(binwidth=1, colour="white",
stat="identity")+scale_x_date(labels = date_format("%m/%d"), breaks = date_breaks("2
days"),limits=c(as.Date("2014/03/24"),as.Date("2014/05/10"))) + labs(x="Log Date",
y="Applicants", title="Number of Applicants Visiting Level2 Web Destinations by
Day")+ geom_vline(xintercept=as.numeric(as.Date("2014/05/01"))) + annotate("text",
x=as.Date("2014-05-04"), y=400 , label = "Deposit Due")

#***************Level3 Destination Breakdown - Students ********

#Admissions

Destinations_AccApp <- subset(MainData,level2=="admissions") %>%
  group_by(logdate,level3,id) %>%
  summarize(hits=n()) %>%
  group_by(logdate,level3)%>%
  summarize(hits=sum(hits),students=n())%>%
  group_by(level3)%>%
  mutate(tot_hits=sum(hits),tot_students=sum(students))%>%
  ungroup() %>%
  arrange(tot_students)

#Plot - Students
ggplot(subset(Destinations_AccApp,tot_students>350), aes(x=logdate, y=students,
fill=level3))+geom_bar(binwidth=1, colour="white",
stat="identity")+scale_x_date(labels = date_format("%m/%d"), breaks = date_breaks("2
days"),limits=c(as.Date("2014/03/24"),as.Date("2014/05/10"))) + labs(x="Log Date",
y="Applicants", title="Level3 Breakdown of Admissions by Number of Applicants and
Day")+ geom_vline(xintercept=as.numeric(as.Date("2014/05/01"))) + annotate("text",
x=as.Date("2014-05-04"), y=325 , label = "Deposit Due")
```

#NOTE: REPEAT PROCESS FOR OTHER LEVEL3 DESTINATIONS

```
#**************Level4 Breakdown of Visit over time
Destinations_AccApp <- subset(MainData,level3=="visit") %>%
  group_by(logdate,level4,id) %>%
  summarize(hits=n()) %>%
  group_by(logdate,level4)%>%
  summarize(hits=sum(hits),students=n())%>%
  group_by(level4)%>%
  mutate(tot_hits=sum(hits),tot_students=sum(students))%>%
  ungroup() %>%
```

```
  arrange(desc(tot_students))

#Plot - Students
ggplot(subset(Destinations_AccApp,tot_students>100), aes(x=logdate, y=students,
fill=level4))+geom_bar(binwidth=1, colour="white",
stat="identity")+scale_x_date(labels = date_format("%m/%d"), breaks = date_breaks("2
days"),limits=c(as.Date("2014/03/24"),as.Date("2014/05/10"))) + labs(x="Log Date",
y="Applicants", title="Level4 Breakdown of Visit by Number of Applicants and Day")+
geom_vline(xintercept=as.numeric(as.Date("2014/05/01"))) + annotate("text",
x=as.Date("2014-05-04"), y=110 , label = "Deposit Due")
```

#**NOTE: REPEAT PROCESS FOR OTHER LEVEL4 DESTINATIONS**

Appendix C: Empirical Analysis Regressions

```
#Empirical Analysis

setwd("/Users/chrisgaribaldi/Documents/RStudio")
library(dplyr) #new package for manipulating datasets
library(tidyr) #new package for reshaping datasets, requires R3.1
library(lubridate) #needed for rounding dates
library(stringr) #for manupulating strings
library(ggplot2)
library(reshape2)
library(scales)
library(stats)
library(stargazer) #for nice summary tables
load("MainData.Rda")

#Filter out period post deposit deadline
MainData <- filter(MainData,logdate!="2014-05-2")
MainData <- filter(MainData,logdate!="2014-05-3")
MainData <- filter(MainData,logdate!="2014-05-4")
MainData <- filter(MainData,logdate!="2014-05-5")
MainData <- filter(MainData,logdate!="2014-05-6")
MainData <- filter(MainData,logdate!="2014-05-7")
MainData <- filter(MainData,logdate!="2014-05-8")
MainData <- filter(MainData,logdate!="2014-05-9")
MainData <- filter(MainData,logdate!="2014-05-10")


byid <- MainData %>%
  group_by(id, enrolled, aprat, logdate, instneed, campusvisit) %>%
  summarize(hits=n()) %>%
  group_by(id, enrolled, aprat, instneed, campusvisit) %>%
  summarize(hits=sum(hits), days=n()) %>%
  ungroup()
byid$instneed <- (byid$instneed / 1000)
byid <-as.data.frame(byid) #this is needed because stargazer needs a dataframe

#create descriptive stats table
stargazer(byid, type = "text", title="Descriptive statistics", digits=1,
out="descript_stats.txt")

#estimate the determinants of enrollment
probit1 <- glm(enrolled ~ instneed,  family = binomial(link = "probit"), data = byid)
summary(probit1)
probit2 <- glm(enrolled ~ instneed + campusvisit,  family = binomial(link = "probit"),
data = byid)
```

```
summary(probit2)
probit3 <- glm(enrolled ~ instneed + campusvisit + aprat,  family = binomial(link =
"probit"), data = byid)
summary(probit3)
probit4 <- glm(enrolled ~ instneed + campusvisit + aprat + days,  family = binomial(link
= "probit"), data = byid)
summary(probit4)
probit5 <- glm(enrolled ~ instneed + campusvisit + aprat + days + hits,  family =
binomial(link = "probit"), data = byid)
summary(probit5)

#results from mutiple models can be nicely summarized by stargazer
stargazer(probit1, probit2, probit3, probit4, probit5, type="text")

#**********REGRESSIONS - DESTINATIONS INCLUDED
#housing, dining, majors_minors, first_year, what_to_expect, finance, calendar

byid <- MainData %>%
  group_by(id, enrolled, aprat, logdate, instneed, campusvisit) %>%
  summarize(hits=n(), housing=max(housing), dining=max(dining),
majors_minors=max(majors_minors), first_year=max(first_year),
what_to_expect=max(what_to_expect), finance=max(finance), calendar=max(calendar),
course_exam_schedules=max(course_exam_schedules), finaid=max(finaid),
deposit=max(deposit), clubs_organizations=max(clubs_organizations),
residential_life=max(residential_life), engineering=max(engineering),
economics=max(economics), managerial_economics=max(managerial_economics),
alumni=max(alumni), minerva=max(minerva)) %>%
  group_by(id, enrolled, aprat, instneed, campusvisit) %>%
  summarize(hits=sum(hits), days=n(), housing=max(housing), dining=max(dining),
majors_minors=max(majors_minors), first_year=max(first_year),
what_to_expect=max(what_to_expect), finance=max(finance),
calendar=max(calendar),course_exam_schedules=max(course_exam_schedules),
finaid=max(finaid), deposit=max(deposit),
clubs_organizations=max(clubs_organizations), residential_life=max(residential_life),
engineering=max(engineering), economics=max(economics),
managerial_economics=max(managerial_economics), alumni=max(alumni),
minerva=max(minerva)) %>%
  ungroup()

byid$instneed <- (byid$instneed / 1000)

byid <-as.data.frame(byid) #this is needed because stargazer needs a dataframe

#create descriptive stats table
stargazer(byid, type = "text", title="Descriptive statistics", digits=1,
out="descript_stats.txt")
```

```
#estimate the determinants of enrollment

probit1 <- glm(enrolled ~ instneed + campusvisit + aprat + days + housing,  family =
binomial(link = "probit"), data = byid)
summary(probit1)
probit2 <- glm(enrolled ~ instneed + campusvisit + aprat + days + dining,  family =
binomial(link = "probit"), data = byid)
summary(probit2)
probit3 <- glm(enrolled ~ instneed + campusvisit + aprat + days + minerva,  family =
binomial(link = "probit"), data = byid)
summary(probit3)
probit4 <- glm(enrolled ~ instneed + campusvisit + aprat + days + majors_minors,
family = binomial(link = "probit"), data = byid)
summary(probit4)
probit5 <- glm(enrolled ~ instneed + campusvisit + aprat + days + engineering,  family =
binomial(link = "probit"), data = byid)
summary(probit5)
probit6 <- glm(enrolled ~ instneed + campusvisit + aprat + days + economics,  family =
binomial(link = "probit"), data = byid)
summary(probit6)
probit7 <- glm(enrolled ~ instneed + campusvisit + aprat + days +
managerial_economics,  family = binomial(link = "probit"), data = byid)
summary(probit7)
probit8 <- glm(enrolled ~ instneed + campusvisit + aprat + days + what_to_expect,
family = binomial(link = "probit"), data = byid)
summary(probit8)
probit9 <- glm(enrolled ~ instneed + campusvisit + aprat + days + finaid,  family =
binomial(link = "probit"), data = byid)
summary(probit9)
probit10 <- glm(enrolled ~ instneed + campusvisit + aprat + days + deposit,  family =
binomial(link = "probit"), data = byid)
summary(probit1)
probit11 <- glm(enrolled ~ instneed + campusvisit + aprat + days + finance, family =
binomial(link = "probit"), data = byid)
summary(probit11)
probit12 <- glm(enrolled ~ instneed + campusvisit + aprat + days + calendar, family =
binomial(link = "probit"), data = byid)
summary(probit12)
probit13 <- glm(enrolled ~ instneed + campusvisit + aprat + days +
course_exam_schedules, family = binomial(link = "probit"), data = byid)
summary(probit13)
probit14 <- glm(enrolled ~ instneed + campusvisit + aprat + days + first_year, family =
binomial(link = "probit"), data = byid)
summary(probit14)
```

```
probit15 <- glm(enrolled ~ instneed + campusvisit + aprat + days + clubs_organizations,
family = binomial(link = "probit"), data = byid)
summary(probit15)
probit16 <- glm(enrolled ~ instneed + campusvisit + aprat + days + residential_life,
family = binomial(link = "probit"), data = byid)
summary(probit16)
probit17 <- glm(enrolled ~ instneed + campusvisit + aprat + days + alumni, family =
binomial(link = "probit"), data = byid)
summary(probit17)
probit18 <- glm(enrolled ~ instneed + campusvisit + aprat + days + housing + dining +
majors_minors + calendar + course_exam_schedules + first_year + what_to_expect +
finance + finaid + deposit + clubs_organizations + residential_life + engineering +
economics + managerial_economics + alumni + minerva, family = binomial(link =
"probit"), data = byid)
summary(probit18)

#results from multiple models can be nicely summarized by stargazer
stargazer(probit1, probit2, probit3, probit4, probit5, probit6, probit7, probit8, probit9,
probit10, probit11, probit12, probit13, probit14, probit15, probit16, probit17, probit18,
type="text")
```