# Beyond COUNTER: Using IP Data to Evaluate Our Users

Timothy R. Morton
*University of Virginia*, morton@virginia.edu



Timothy R. Morton, "Beyond COUNTER: Using IP Data to Evaluate Our Users" (2013). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315274

# Beyond COUNTER: Using IP Data to Evaluate Our Users

*Timothy R. Morton, Electronic Resources Librarian, University of Virginia Library*

## Abstract

Traditional library statistics, whether counting our collections, our users, or our services, are typically concerned with answering questions such as “What?” or “How much?” or “When?” COUNTER-compliant statistics, the very welcome and useful standard for electronic resource providers, have allowed libraries to bring that same paradigm to bear on their digital collections, answering such questions as “What journals and e-books are our users downloading?” “How often are they searching this database?”, and even “When do they access this content?” However, what COUNTER and other traditional methods often fail to do is provide data that would allow libraries to answer questions such as “Who is using our resources?” and “Where are they when they access our licensed content?” By gathering detailed usage data by IP address from several electronic resource providers, and comparing those datasets with a well-developed network infrastructure, one can take steps to determine the “who” and “where” questions of e-resource usage at the University of Virginia.

## Introduction

This project had its genesis in the University of Virginia’s (UVA) well-publicized restructuring of its financial model. At the time, the plan was for various administrative units which are not revenue producing, such as the library, to be funded by a “tax” on revenue-producing units, such as the constituent colleges and schools. To inform this discussion, the library sought out numerous data points, including enrollment, circulation, and interlibrary loan requests, which could show comparative use by each school. During this information gathering, it was discovered that several of our electronic resource providers offered usage statistics by IP address. Since the university’s network infrastructure assigns those IP ranges by building, and the buildings are generally associated with a single school, this allowed the gathering of rudimentary information about the relative use of a couple of e-resources by the various schools at the university.

As the Electronic Resources Librarian, I was responsible for gathering these data and quickly realized the potential information that could be gleaned by looking at this IP address usage data in greater breadth and depth. Whereas the previous effort looked at broad school groupings for a couple of databases for financial purposes, this study aggregated the data on a much more granular level, not focused solely on academic units, and attempted to discern user characteristics rather than just organizational affiliation.

## Vendor Data

This project collected datasets from 12 different electronic resource providers, and represented 18 months of usage from January 2012 through June 2013. The datasets varied wildly between vendors—at a minimum, they consisted of an IP address and the number of full-text downloads and/or user sessions associated with that address for the given time period. Three datasets contained user session data, six contained full-text download data, and three contained both. In terms of market share, these 12 providers together accounted for approximately 54% of UVA’s electronic resource usage over that 18-month period.

I attempted to get IP address data from our most used electronic resource providers. The providers ultimately included in this study were chosen based on their ability and/or willingness to provide the necessary detailed IP address data. Several providers allowed me to harvest IP data via their existing data collection interfaces. Where this option did not exist, I e-mailed the providers directly, explained the study, and requested a custom data report. Several supplied the data after this initial contact, but several other large


http://dx.doi.org/10.5703/1288284315274

vendors whom I had hoped to include did not. Some initially declined due to privacy concerns, some indicated that technical limitations prevented that level of data collection, and some gave no reasons for being unable to provide data. I renegotiated with those vendors who initially balked due to the sensitive nature of the data and was able to reach compromises which would allow for their data to be used, while maintaining user privacy.

## Privacy Concerns

Like most librarians, I am cognizant of the need for patron privacy in their use of our resources and realized from the outset that I would need a plan to sufficiently protect that privacy in this project. As mentioned above, this was not simply good professional practice, but in fact became necessary for the cooperation of a few of the electronic resource providers. In order to maintain patron privacy, the IP addresses in each dataset were immediately anonymized to the level of the third octet, so that 128.143.1.1 became 128.143.1.xxx. After that, all data other than the IP address and the number of full-text downloads or user sessions was removed entirely.

These two steps were taken for both technical and practical reasons. First, by removing the fourth octet, I not only eliminated the ability to identify an individual computer, but also brought the data in line with the dataset outlining the network infrastructure at UVA. Our IP addresses are assigned to each building on campus at the level of the third octet of the IP address, making the fourth octet extraneous when it comes to matching them to a physical location. Second, some of the datasets were highly detailed, going well beyond the desired IP address/usage total. Some included such additional information as the exact title accessed and the date/time of the visit. From a privacy point of view, if simply knowing the number of times a particular computer accessed a database might be troublesome, then knowing each and every title accessed by that computer was unacceptable. Third, some vendors only provided data down to the third octet by default, and still others refused to provide any data at all unless it was anonymized to that level before delivery. In the end, not only did editing the datasets protect patron privacy, but it had a secondary benefit of making the datasets less cumbersome to analyze.

## Institutional Data

UVA organizes their network infrastructure along the same lines as many peer institutions. The IP addresses owned by UVA are divided up into 475 three-octet blocks and assigned to either individual large buildings or clusters of smaller buildings. Some buildings may only have one block, while others, such as the hospital, may have dozens. Generally speaking, these blocks are specified down to the third octet of the IP address (i.e., the first three sets of numbers). For example:

> 128.143.141.* is assigned to Cabell Hall
>
> 128.143.142.* belongs to Garrett and Maury Halls
>
> 128.143.143.* goes to Newcomb Hall and the Bookstore

In addition to being divided by location, the university's IP address blocks are also divided into two distinct networks, the Less Secure Network (LSN) and the More Secure Network (MSN). The LSN is the only network available to students and guests. Personally owned faculty and staff computers, as well as shared UVA-owned computers, such as those in research labs, libraries, or classrooms, are also restricted to the LSN. The MSN, on the other hand, is accessible only by UVA-owned computers that are assigned to an individual faculty or staff member as their exclusive work computer.

A final general division in the university's network setup is the allocation of a large block of IP addresses for network address translation (NAT) for the university's wireless network. All wireless users, regardless of physical location or user status, are assigned an IP address from this block which is reserved solely for wireless access. This networking scheme proved to be a mixed blessing for this study. On the one hand, it allowed the separation of wireless usage and wired usage, so we could draw definitive conclusions about the relative importance of the wireless network (at

least for research purposes). On the other hand, this wireless NAT block is effectively a black box, with no way to discern a user's location or status, at least on the scale of this project. The Health System has a similar IP block dedicated to NATing not only their wireless network, but also some wired connections as well as their own Health System VPN profile.

Taking these different considerations into mind, I created a classification scheme and assigned the entirety of the university's master IP address list accordingly. Each allocated IP address block/building was assigned to one of several categories, as outlined in the Table 1.

## Assumptions

Unfortunately, while the results of this study are highly suggestive of who our users are and from where they are conducting their research, there were a few necessary assumptions that prevented perfect accuracy of the results.

First, I had to assign each building to a single organizational category, even when it housed multiple organizations. The IP address blocks are assigned by building, not by organization, so for several libraries which are collocated with their constituent departments/schools, it is impossible to tell which IP addresses in the building are assigned to which unit. Based on the relative size of the library and remainder of the building, some buildings were assigned to the library category and others were allocated to the school or college. This means that there will inevitably be some library usage counted as part of an academic unit, and vice versa.

Second, I assumed that LSN usage from most academic buildings was overwhelmingly graduate students. Since all wireless usage is segregated into its own category, the remaining usage must come from hardwired connections. However, several years ago, public computer labs were removed from all buildings, aside from those few housed in the libraries. As a result, the only LSN wired network connections remaining in academic buildings should be those in research labs and graduate student departmental offices. While there will undoubtedly be some faculty and undergraduate presence in these research labs, based on the relative proportions of the staffing in those labs, I assume that grad students account for the lion's share of the usage. The only exception to this assumption is the Commerce School, which exclusively serves undergraduate students.

To increase confidence in this assumption, I looked at a previous study of UVa graduate students conducted by library colleagues in 2009. This study attempted to describe the graduate student research process via in-depth interviews with at least one Masters and one PhD student from each department on campus. When asked about where they were physically located when conducting research involving library resources, 38% said their departmental office/lab was their primary research location, second only to working from home at 44%.

Third, there were some areas primarily used by a combination of two of our three user groups, wherein it was impossible to differentiate the use between those two groups. For instance, LSN

| Location Category | Constituent Locations/Buildings |
|---|---|
| Remote | VPN, Proxy Server |
| Libraries | 5 large, non-collocated libraries |
| Dorms | Dorm wired and wireless ranges |
| Health System | Hospital, Health System, School of Medicine (excl. Health Sciences Library) |
| Administration | Independent research labs, Facilities, IT, University Administration, other university-wide services |
| Wireless | Wireless access points (excl. wireless in Dorms and Health System) |
| Academic Units (x10) | Buildings associated with each of the 10 colleges/schools (excl. School of Medicine) |
| Other | Dining Halls, Athletic Facilities, Bookstore, Student Centers |

**Table 1. Location Categories and Affiliated Buildings**

| User Group | Constituent IP Addresses |
|---|---|
| Likely Faculty | All MSN |
| Likely Graduate Students | LSN from all school except Commerce |
| Likely Undergraduate Students | LSN from Commerce and Other, Dorms |
| Likely Faculty/Grad Students | LSN from Administration |
| Likely Grad/Undergrad Students | LSN from Continuing Studies |

**Table 2. Assignment of IP ranges to User Group Categories**

| Location Category | Full-Text | User Sessions |
|---|---|---|
| Remote | 488,005 (39%) | 243,263 (40%) |
| Health System | 252,469 (20%) | 70,943 (12%) |
| Wireless | 249,779 (20%) | 111,621 (18%) |
| Academic Units | 119,986 (10%) | 67,122 (11%) |
| Dorms | 68,746 (6%) | 60,051 (10%) |
| Libraries | 39,457 (3%) | 46,420 (8%) |
| Admin | 24,544 (2%) | 7,516 (1%) |
| Other | 1,662 (<1%) | 2,035 (<1%) |

**Table 3. E-Resource Use by Location**

addresses in the Administration category represent a wide variety of facilities operating shared computers, many of which are interdisciplinary research labs. By their nature, they are not likely to have a significant undergraduate presence, but rather a combination of faculty and graduate students. However, since shared lab computers are not eligible for MSN access, I cannot differentiate between faculty and grad students within those labs. Additionally, many of these likely grad student-heavy labs are collocated with other faculty/staff-heavy administrative units. Similarly, LSN usage at the School of Continuing and Professional Studies represents remote campuses which maintain both graduate and undergraduate programs, but contain no full-time faculty presence. As such, this usage will be either undergraduate or graduate students, but likely not faculty.

Fourth, I assumed that usage coming from the MSN will be faculty, given that the only computers able to access this network are nonshared faculty computers. However, while the MSN is restricted to university-owned computers assigned to an individual member of the faculty or staff, there is no guarantee that every computer has been properly configured to use the MSN. The migration of the computer from the LSN to the MSN requires a few additional steps, which are generally performed by the local IT support partner assigned to a department. In conversations with IT staff across the university, MSN penetration is very high, but by no means exhaustive. As a result, there will be some small amount of faculty resource usage that gets tallied under a building's LSN IP ranges rather than the MSN range.

## Results

*By Location*

When looking at the data tallied by location, remote access clearly outstrips all other locations for the consumption of electronic resources, accounting for 40% of full-text downloads and 39% of user sessions. Interestingly, even this large number is likely an undercount. As previously mentioned an enormous IP range assigned to the Health System is used for NATing wireless, wired, and VPN connections, and the Health System also maintains its own additional VPN profiles. As a result, some unknowable number of the Health System uses are themselves remote access, meaning that potentially half of the total e-resource consumption is taking place from off campus.

| School/College | Full-Text | User Sessions |
|---|---|---|
| Engineering | 63,795 (52%) | 7,961 (12%) |
| Arts & Sciences | 40,833 (24%) | 14,225 (21%) |
| Law | 4,568 (4%) | 3,612 (5%) |
| Education | 3,096 (3%) | 3,952 (6%) |
| Nursing | 2,776 (2%) | 2,910 (4%) |
| Business | 2,344 (2%) | 23,436 (35%) |
| Commerce | 1,810 (1%) | 7,308 (11%) |
| Architecture | 1,734 (1%) | 3,111 (5%) |
| Continuing Studies | 646 (<1%) | 441 (1%) |
| Public Policy | 194 (<1%) | 166 (<1%) |

**Table 4. E-Resource Use by School/College**

| User Group | Full-Text | User Sessions |
|---|---|---|
| Likely Grads | 110,855 (50%) | 53,032 (38%) |
| Likely Grad/Faculty | 22,107 (10%) | 6,296 (4%) |
| Likely Faculty | 13,731 (6%) | 12,410 (9%) |
| Likely Undergrads | 73,842 (34%) | 69,357 (49%) |
| Likely Ugrad/Grad | 646 (<1%) | 441 (<1%) |

**Table 5. E-Resource Use by User Group**

One very interesting result was the disparity in usage between the various schools and colleges as evidenced by full-text downloads and user sessions. When looking at downloads, the College of Arts and Sciences and the School of Engineering and Applied Sciences account for 86% of the use, and this heavy use might be expected since they make up two-thirds of the university's population. The School of Engineering and Applied Sciences also showed interesting results, in that they were responsible for only 12% of the user sessions, but 52% of the full-text downloads. Conversely, the Darden School of Business is responsible for 35% of the user sessions, but only 2% of the full-text downloads, and represents 4% of the university population.

*By User Group*

Based on the user group results, the vast majority of the research (as measured by electronic resource usage) on campus is conducted by students, both undergraduates and graduates. Faculty, on the other hand, account for roughly 10% of the total use, whether full-text downloads or user sessions. Just as with the location results, there are some intriguing disparities between the different user groups in terms of full-text download and user session data. Graduate students account for only a third of the user sessions, but half of the downloads, whereas undergraduates are responsible for half the sessions, but only one third of the downloads.

## Next Steps

The results of this project suggest a few possible avenues for further study, some of which will further complete the picture of who is using our resources and where they are, others of which are completely unrelated and were discovered in the course of this project.

The first and most logical follow-up would be to undertake a closer examination of the remote access to our electronic resources. Whether sessions or full-text downloads, remote access makes up at least 39% of our overall use. However, as stated above, since over half of the Health System use comes from a massive IP range

that includes NATing for the Health System VPN, even that 40% figure is likely an undercount. It's entirely possible that half of the electronic resource usage at UVa occurs off campus. In order to get a more complete view of the user groups, we would need to analyze our proxy and VPN logs, comparing them to our institutional directory. I've already taken limited steps in this regard, and an examination of a small selection of our proxy logs show that roughly 20% of the remote use is coming from faculty, which is approximately twice their share of the on-campus use. This rough data fits with the anecdotal picture painted by several subject librarians based on conversations with their faculty.

Second, the results suggest investing more time and resources into understanding and engaging with our graduate students. Graduate students seem to be conducting the majority of the research at UVa, at least as measured by electronic resource usage, however there is no systematic campaign to reach them and market the library's collections and services to them. The previously referenced 2009 study is the only significant attempt to study graduate students and their relationship with the library. In the meantime, library instructional planning has focused almost exclusively on lower-level undergraduates, the library has hired an Undergraduate Services Librarian, and the library has sponsored collaborative seminars with other units from across campus to understand the undergraduate student population. All of those steps are worthy and laudable, but there is a strong argument to be made for equivalent undertaking with the graduate student population.

Third, the results have already made an impact by pointing towards a new way to manage collections expenditures for digital resources. This is a wholly unexpected outcome from this project. Traditionally, larger interdisciplinary resources were funded centrally, while those more tailored to specific departments were funded by the subject allocations. While many of our largest and broadest electronic resources are obviously used across departments, I was surprised to find that even the narrowly focused resources were used across the institution. For instance, the Royal Society of Chemistry's collections are largely paid for by the Physical Sciences Librarian's allocation, and their obvious constituents are the Chemistry Department in the College of Arts & Sciences. However, we found that use by that intended constituency only accounted for half of the total use. With such a usage pattern we determined that it does not make sense for a single selector to pay the majority of the cost even while the majority of the use came from another selector's constituent departments. In response, the University of Virginia Library has shifted to central funding for all e-resources.