Paper 197-2012

"A Week in the Life": A Visual Analysis of Internet Use by School-Age Students

Aaron Daniels, Simon King, Jacob Warwick
Cary Academy, Cary, NC, USA

ABSTRACT

This study took data of the Internet web hits for one week for the students at Cary Academy, a grades 6–12 independent public school in Cary, North Carolina. All students have their own school-issued tablet PC. Social networks are blocked while students are in school. Student web hits were monitored from 8 a.m. to midnight for a school week. It was discovered that student Internet use was up to twice as frequent outside of school, with the use of social network sites increasing greatly for upper-school students, with only a small increase for middle school students. All students showed a significant use of "streaming" websites, a more recent Internet trend. Conclusively, Internet use is not "out of control" in the classroom.

INTRODUCTION

Cary Academy is a grades 6-12 independent college preparatory school founded by Mr. and Mrs. James and Ann Goodnight and Mr. and Mrs. John and Ginger Sall in 1996. Cary Academy holds a student body of approximately 700 and a faculty and staff of 150. Cary Academy supplies each student and teacher with a tablet PC, which students and faculty take home each day and over school breaks. In classes, students often use online resources instead of paper worksheets, as CA provides high-speed wireless internet access on campus.

All Cary Academy student internet usage, including usage at home, is logged by the industry leading security software WebSense (http://www.websense.com), which assigns each web request a usage category, such as "Education" or "Productivity," to be logged with the request. Certain categories, such as "Violence" or "Drugs" are blocked at all times, while others such as "Tasteless" or "Social Networking" are only blocked during school hours. Via a web interface, teachers are able to view students' web activity during the school day, while parents are able to view all internet usage at home or at school.

This system of web tracking allows for a new type of statistical examination of teenage internet use based on a census of the population, instead of making inferences from a sample. Previous studies of teenage internet use have relied on inferential results based off of random cell-phone surveys, as the most recent 2009 Pew Research Center Internet & American life Project's Parent-Teen Cell Phone Survey (3). The project surveyed 900 teens ages 12-17 about their internet and cell phone use, and found that 89% of teens interviewed use the internet on a daily or weekly basis, and the majority of teen internet use is centered around social networking sites (73%), news and information sites (62%), and online shopping (48%).

METHOD

From October 29th, 2011, to November 5th, 2011, every web request by a Cary Academy student was tracked and logged on Cary Academy's servers. Cary Academy was selected as it is one of the few schools with web monitoring capabilities that make this type of data collection possible. The students of Cary Academy were the main focus of this study. WebSense is a tool for monitoring web usage over a network, and blocking requests for sites in certain categories. When a web "hit" is recorded by WebSense, it is automatically categorized into one of the following categories:

Abortion	News and Media
Adult Material	Non-HTTP
Advocacy Groups	Parked Domain
Bandwidth	Productivity
Business and Economy	Racism and Hate
Drugs	Religion
Education	Security

Entertainment	Shopping
Extended Protection	Social Organizations
Gambling	Society and Lifestyles
Games	Sports
Government	Tasteless
Health	Travel
Illegal or Questionable	Job Search
Information Technology	Internet Communication

Figure 1: WebSense-assigned Web Hit Categories

A web hit is a record of the request a browser makes to the internet, and all the subsequent requests necessary in the loading of a page. Cary Academy records web hits in a relational MSSQL database which includes the target URL, the name and grade of the student making the request, the date and time of the request (accurate to the second), and the WebSense-assigned category. In order to avoid bias and preserve confidentiality the students' names were replaced with a randomly generated ID numbers.

The data extracted from WebSense were dumped into CSV files and converted into SAS® Datasets using PROC IMPORT. DATA steps and PROC MERGE were used to combine the relational database into one large SAS Dataset. After extracting the web hits from WebSense, there were noticeable anomalies within the data. There were some grade levels equal to "13"- an impossibility - and some were null values. These two grade levels were later identified as teacher web hits and administrator activity. These web hits were excluded from the dataset, which narrowed down the total number of web hits from 16,308,931 to 16,202,435.

The data also included redundant web hits recorded by WebSense - one user accessing one website multiple times witin a single second. It was thought that these hits represented additional web requests to backround sites, such as advertisement servers, made in the process of loading a page. In order to eliminate these redundant web hits from the dataset the following SAS procedure was applied:

```
proc sort data=webdata.combined_data out=webdata.cleaned_data nodup;
by DateTime UserID;
run;
```

This procedure had the effect of removing all web hits by an individual user at a particular second, except the first hit recorded. As such, this could have had the most potential to affect the data. However, as the combined dataset was sorted on the DateTime variable, the procedure above would have kept only the first request, which is assumed to be the page originally requested by the user. This further reduced the total number of web hits from 16,202,435 to 9,778,829. That resulting dataset was used in conjunction with SAS 9.3 to analyze internet use.

DISCUSSION

This method also collected a census data instead of just a sample from which inferential statistics could be made. By gathering a census instead of a sample, inferences were not needed, as the entire population was included within the data. Since no inferences were made upon the population, sampling bias was eliminated from the study. Because a census was collected and Cary Academy consists of a unique student population, any attempt to infer the results from this study on other populations would be pointless and would prove to be inaccurate. This data was collected before the experiment was devised, so it is practically impossible that experimental bias could have been introduced into the data.

The only subjective part of this study was the classifying of websites into categories and usage types. WebSense automatically categorized the web hits, which eliminated some subjectivity; however, the usage types created in this study (good, bad, distracting, and neutral) were subjective as it is impossible to tell exactly what the user's intent was when requesting a site. These findings represent the best educated guess as to what the user's true intent was when accessing a website.

The choice to use SAS software to run the analysis for this experiment ultimately proved to be a wise one, as the number of raw web hits downloaded off the servers was quite large - 16 million - and a flexible means of manipulating and graphing the data was required. However, there were some drawbacks to the method as well. In the 'Tile Cloud' chart, writing to an HTML page prevented the use of official SAS/GRAPH® procedures, namely, PROC GTILE. Additionally, the limitations of HTML and SAS prevented a word cloud with both vertical and horizontal elements to be created inside each tile. Additionally, only one week was analyzed throughout this study. With different amounts of

workloads throughout the school year as well as different internet trends throughout the year to consider, the week in question may not have been representative of all internet use in a school year.

In contrast to the Pew Research Center's previous research, this experiment found that overall, the majority of Cary Academy student internet use is centered on Society and Lifestyles (17.2875%) and Productivity (17.0145%). This depends, however, on if one is to view internet usage as a whole or just within or out of school hours. During school, Cary Academy students' primary category of usage is Productivity (20.2222%), followed by Information Technology (15.4814%). It would seem that the majority of academic-related activities fall into those two usage categories. Outside school hours, we see a significant shift in priorities - from education to entertainment. The top category is Society and Lifestyles (21.2173%), which includes social networking, followed by Bandwidth-intensive usage such as video streaming (17.2340%), a new trend in internet use which has not been previously documented. Overall, internet usage among Cary Academy students followed a pattern which showed that while tablet computers are used as sources of distraction, they are consistently used for education-related activities as well.

VISUALIZATIONS

DEALING WITH SIGNIFICANT VARIATION

Initial attempts to analyze the raw data were thwarted by large amounts of variability in the instantaneous amounts of hits per second. Figure 2 shows a sample of raw data as recorded by the server:

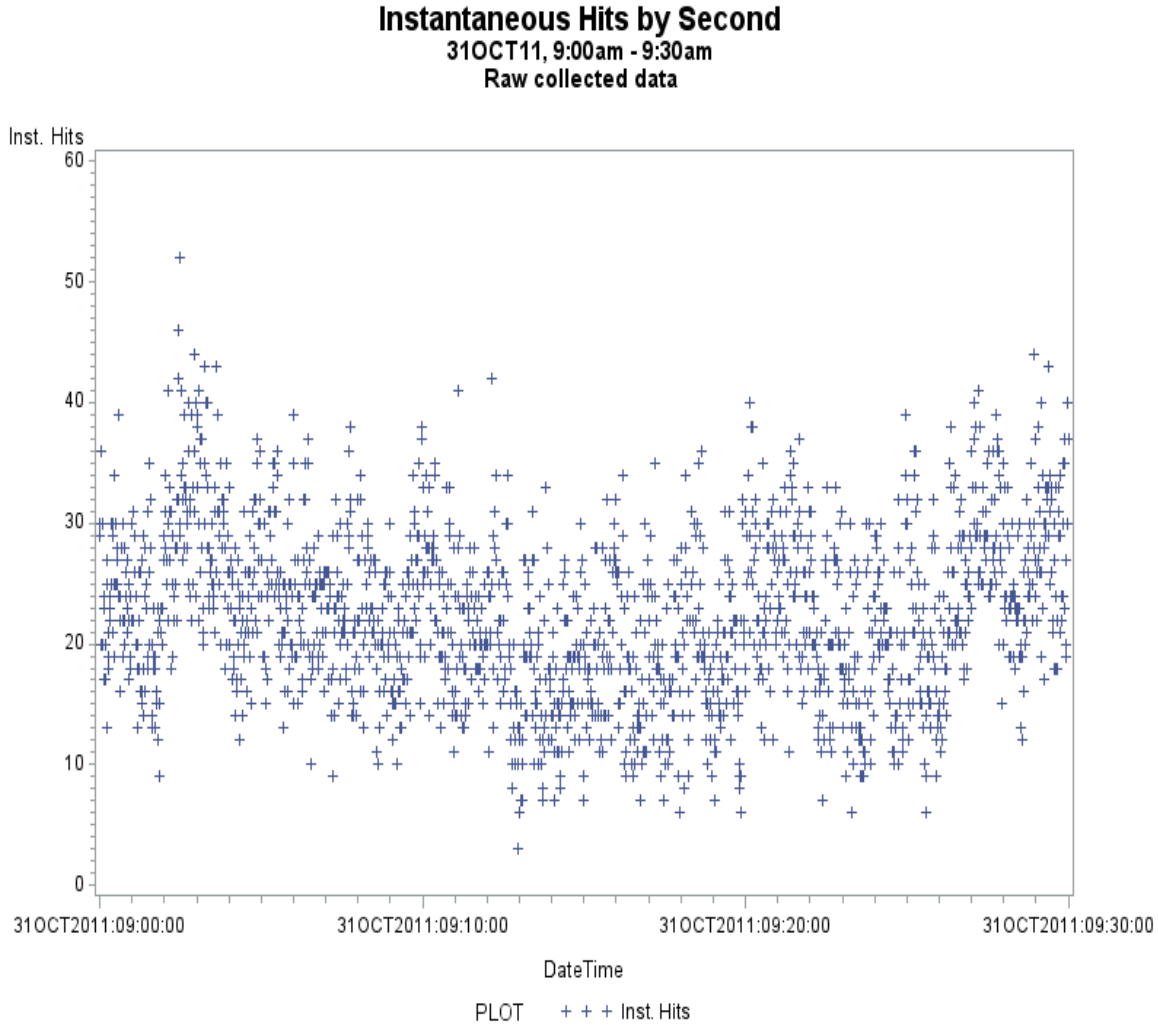

Figure 2: Instantaneous Hits by Second, 31OCT11, 9:00-9:30am, raw data

To deal with this problem, the first step in most analyses was to apply a 600 second (10 minute) moving average, as demonstrated in the following sample DATA step:

```
%LET n=600;
DATA out_set;
       retain s;
       set in_set;
       s=sum(s, Num_Hits, -lag&n(Num_Hits));
       Moving_Average = s / &n;
       drop s;
run;
```

Various durations for the moving average were tested, and the 600 second moving average was found to be the smallest possible number that would still produce a smooth shape when graphed at an hour or day level. The resulting dataset provided a much better indicator of the usage trend over a period of time. Figure 3 shows the same time period after averaging:

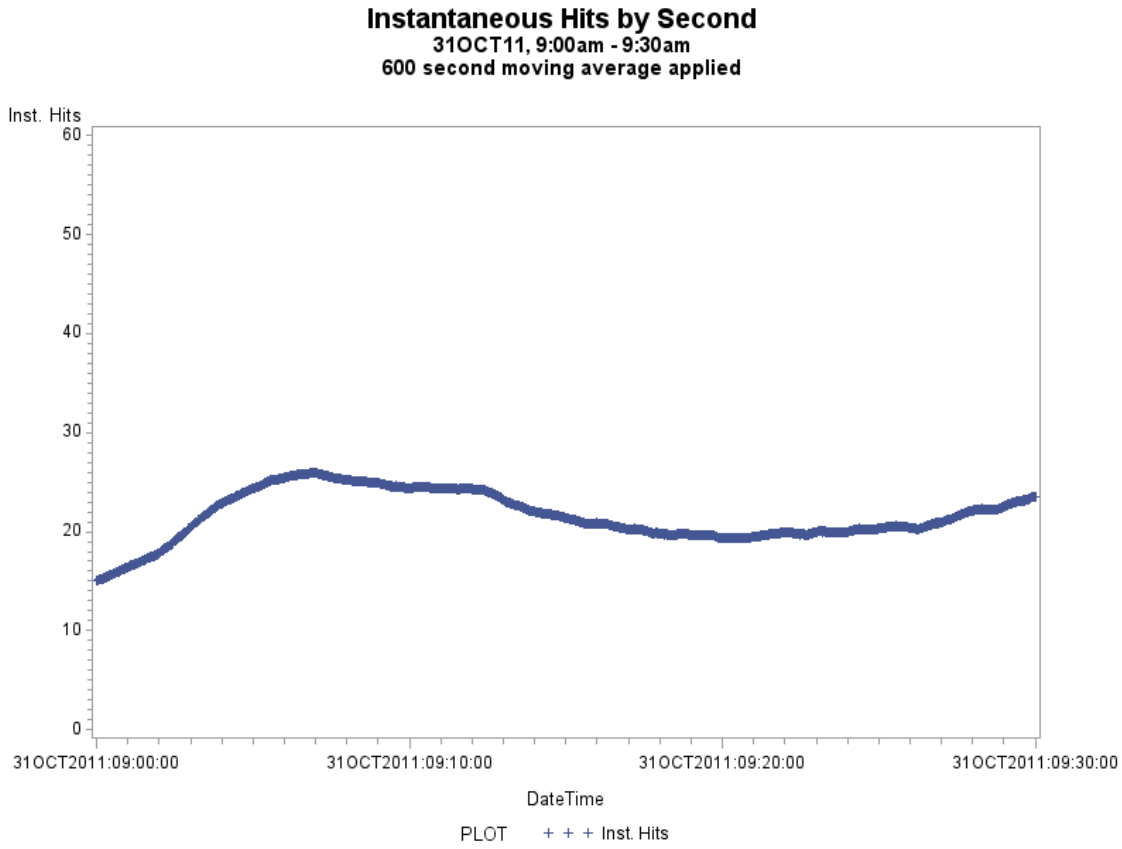

Figure 3: Instantaneous Hits by Second, 31OCT11, 9:00-9:30am, 600 second moving average

GRAPHING COMPLEXITIES

The next obstacle to overcome was the sheer complexity of the data. WebSense-assigned categories occasionally had sub-categories, and each 'hit' was linked with an anonymous user ID as well as a grade level number (6-12). To cope with these limitations, a graph was produced based on the one published in a July 31st article by the New York Times detailing the ways that different subgroups of Americans spend their day (2). This technique layers each category on top of the other categories, producing a 'sandwich' effect, where at each point, the sum of all the categories is equal to the total usage or percentage. In SAS, this was accomplished in a DATA step. First, the total usage value was multiplied by negative one, essentially flipping it over the x axis, and creating the distinctive "envelope" shape when graphed with the original value. The first category (lowest on the graph) was multiplied by two and added to the 'lower' value. Additional categories were then multiplied by two and added to the previous category, to create the 'stacked' look for each category. The doubling was necessary because total usage was effectively

doubled when flipped over the x axis. Using PROC GPLOT, the new category variables were drawn onto an overlay plot, and the AREAS= option was used to create different shaded areas for each category. Figure 4 demonstrates the new technique with the categories originally assigned by the WebSense software. A 600 second moving average was applied to each category's instantaneous hit count to create a smooth trend line. The full code may be found in the appendix.

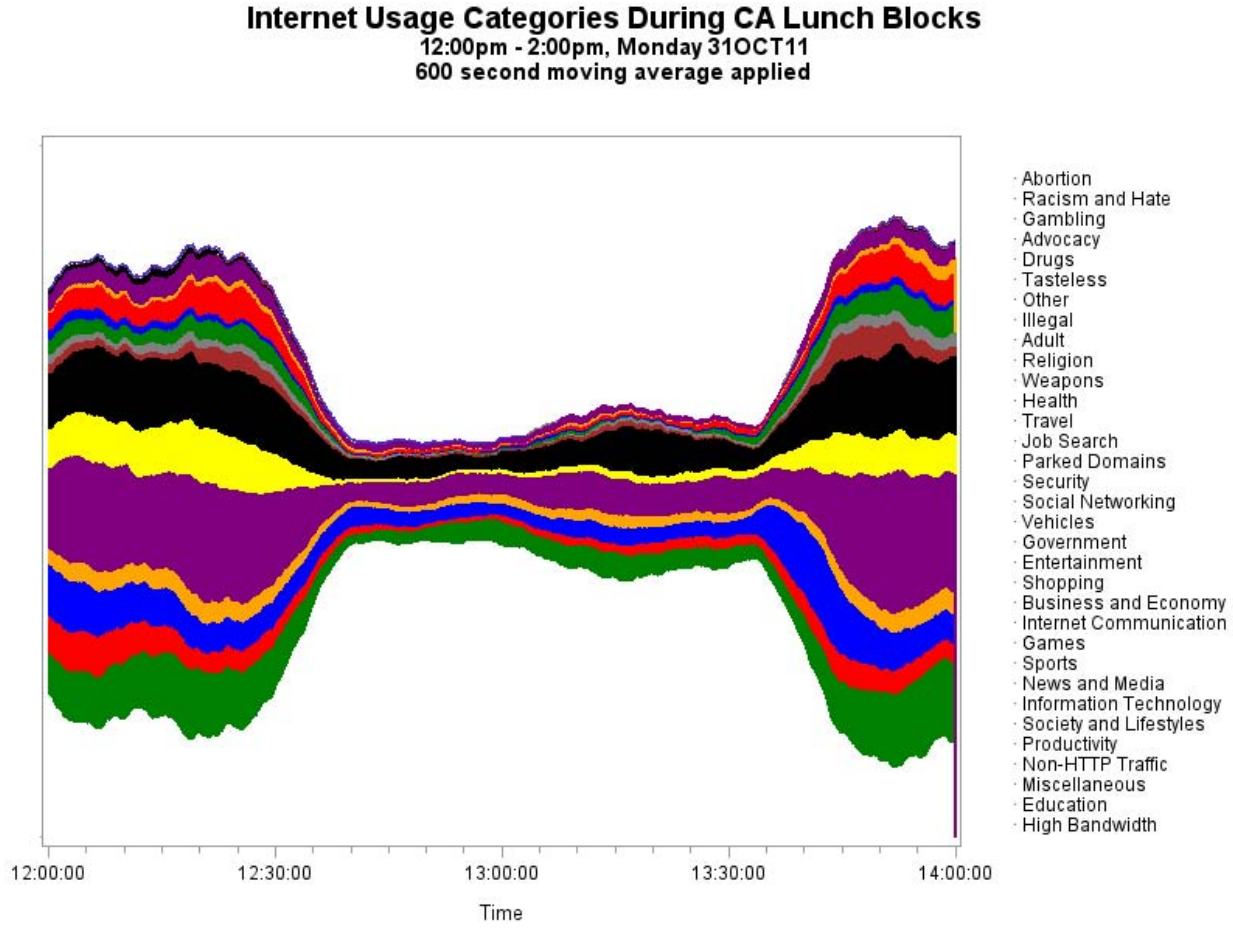

Figure 4: Internet Usage Categories around CA Lunch Blocks, new graphing technique

Ultimately, however, the original categories proved too small and too numerous to effectively graph, and the above technique of graphing with categories was thought to contain too much visual clutter. A new variable was created to solve this problem, by combining the existing categories into 'Types' of usage. The categories and subsequent types are displayed in figure 5:

Usage Type:	**Good**	**Neutral**	**Distracting**	**Bad**
Categories:	Education	Business and Economy	Abortion	Adult Material
	Government	Extended Protection (A WebSense feature)	Advocacy Groups	Drugs
	Information Technology	Health	Bandwidth	Gambling
	Job Search	News and Media	Games	Illegal or Questionable
	Productivity	Non-HTTP	Internet Communication	Military or Extremist
		Parked Domains	Religion	Racism and Hate
		Security	Shopping	Tasteless
			Social Organizations	Violence
			Society and Lifestyles	Weapons
			Special Events	
			Sports	
			Travel	
			Vehicles	

Figure 5: Usage types and equivalent categories

The choices of which categories go into which usage types are subjective (for example, both Religion and Abortion are considered Distracting). Despite the controversy surrounding these and other assignments, it is important to remember that the usage types were created as a rough tool to see how usage stacked up, not as a finely-tuned instrument for judging students' behavior based on a set of moral or ethical standards. It was felt that considering the existing inherent unreliability of WebSense's own category-assignment software, the creation of higher-level categories did not add an unacceptable level of uncertainty to the equation. However, it is important that readers keep this uncertainty in mind while viewing the following graphs and charts.

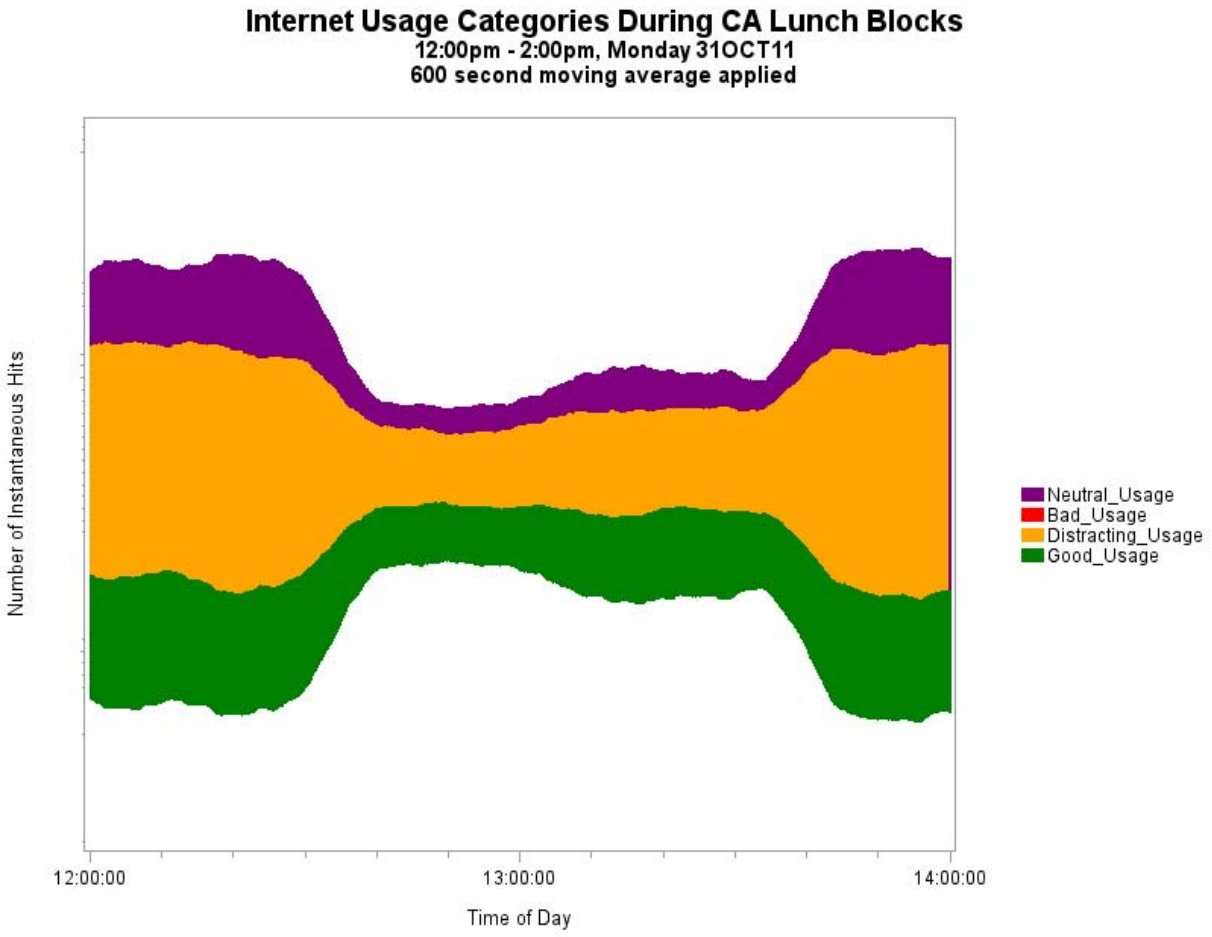

Figure 6: Grouped Internet Usage Types around CA Lunch Blocks, new graphing technique

A DIFFERENT APPROACH

One problem with the first technique for viewing internet usage at Cary Academy was the sheer number of distinct categories - 34 of them - made it almost impossible to discern where the majority of usage was coming from. Normally, a word cloud or pie chart would be useful for this kind of display, but the categories existed inside each "type" of usage and it was felt that to be truly effective, a visualization must be created to show category use and the more general type of usage at the same time. Ideally, the best representation would be some kind of nested pie chart or tag cloud, however, SAS does not currently have a built-in functionality to support this kind of visual combination.

In a paper submitted to SAS Global Forum 2011 titled "What Were We Talking About at Those SAS Conferences, or Let's Make Some Tag Clouds" (1), Chang Chung and John King outlined a technique in which SAS generated the HTML code for a tag cloud and then displayed it using the 'x' command. That technique was modified to create several tag clouds inside appropriately sized rectangle containers - essentially, a tile chart - though not one created through PROC GTILE (the full code may be found in the appendix). Figure 7 shows the result, nicknamed a "tile cloud:"

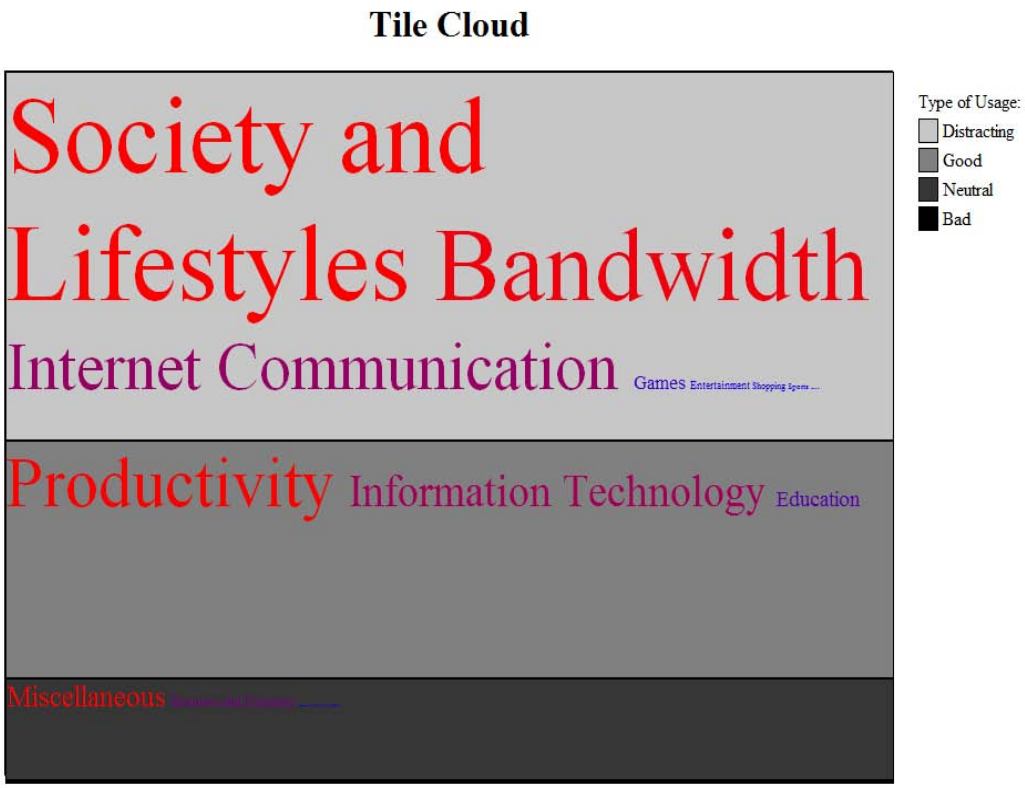

Figure 7: Tile chart of usage types with word cloud of categories interposed

The size of the words in the above visual are both relational to themselves (Society and Lifestyles is proportionally larger than Internet Communication, for example), and to each other (Society and Lifestyles is also proportional to Productivity and Miscellaneous). The colors of the words range from red to blue, in proportion to the percentage of the whole usage type being taken up by each usage category.

COMPARING USAGE DURING AND AFTER SCHOOL

Using the two techniques described above, it was possible to juxtapose visualizations comparing usage before and after school hours. Averaging the instantaneous hits at each time of day during the school week produced a general model of the usage patterns during the week. That dataset was used to create the following visuals:

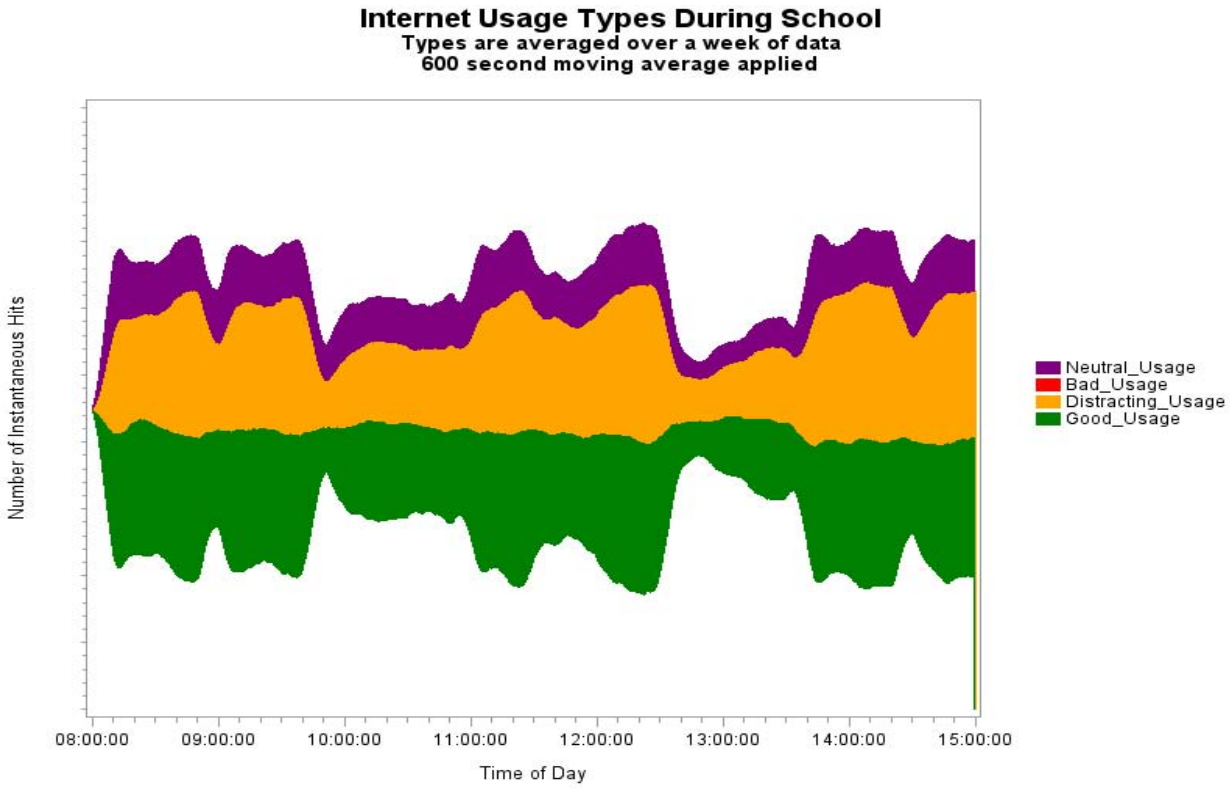

Figure 8: Average week model, school hours, new graphing technique

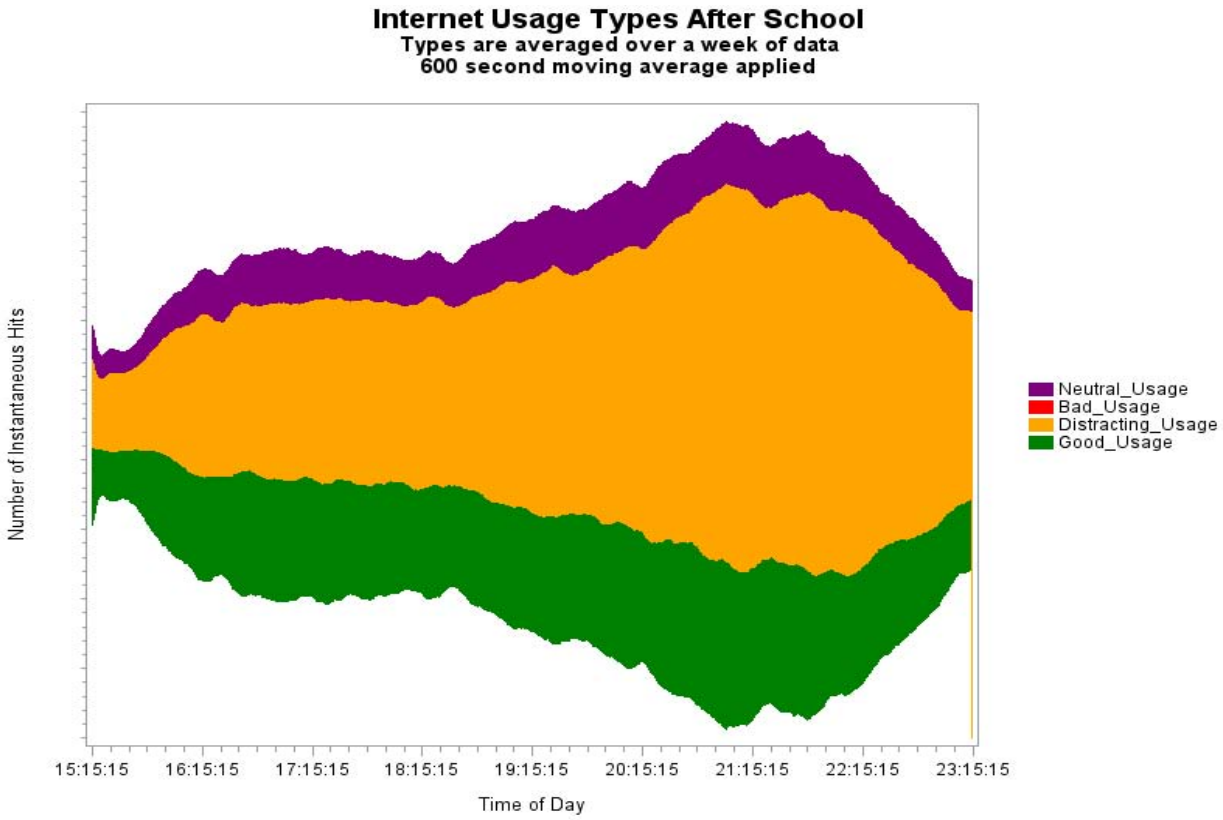

Figure 9: Average week model, after school hours, new graphing technique

Cary Academy uses a block schedule for classes, meaning that students have all their classes in various orders each day, in 45-minute blocks, except for on Wednesdays and Thursday. On those days, Cary Academy follows a double block pattern, with 3, 1.5 hour class blocks. Classes are separated by a 5 minute break, in which students move from classroom to classroom. While use during the school day follows the predictable 'block' pattern synchronized to the Cary Academy schedule, after school the total usage increases to a peak around 9:10pm and then decreases, probably due to students going to bed. The double blocks don't appear to have a huge effect on the graph. Interestingly, throughout the school day and at home, the "Good" type of usage (comprised of websites related to productivity, research, and academics) stays at roughly the same percentage, while the "Distracting" category causes the increase in total use. In other words, this data suggests that Cary Academy students are just about as likely to be doing work at 5:00pm as 11:00pm, but later in the evening, they are much more likely to also be involved in distracting activities - social networking, streaming video, etc. However, what this graph cannot tell us is the detailed categorical breakdown of each of the Type categories in our given time periods, and for that, we turn to the 'tile cloud' method, as shown in figures 10 and 11:

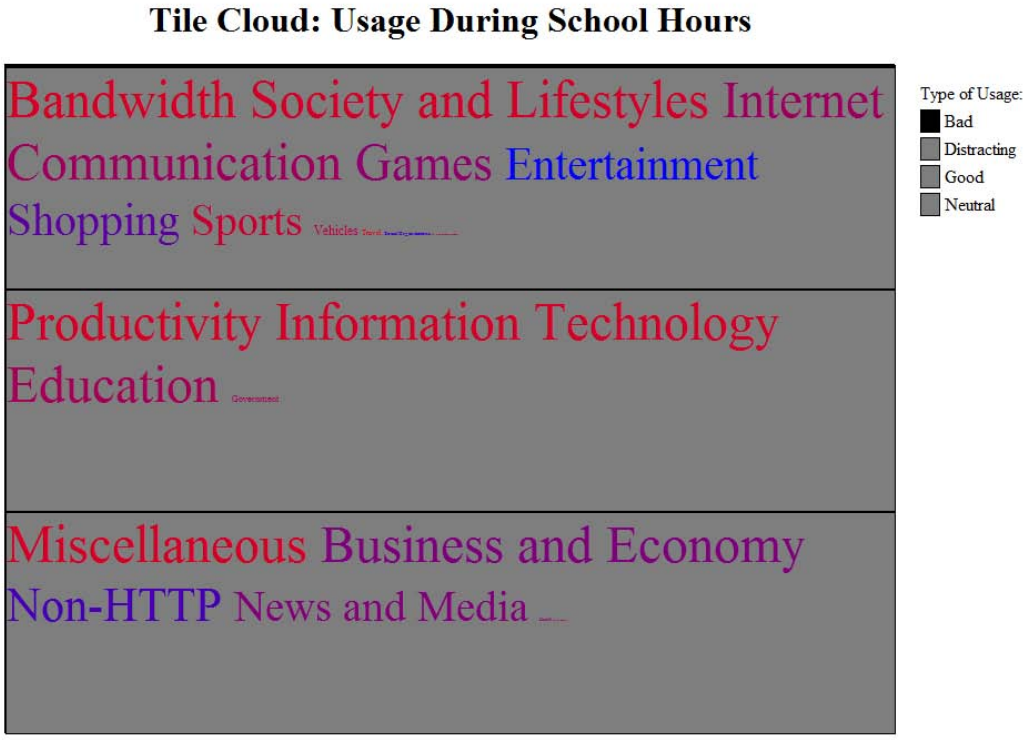

Figure 10: "Tile cloud" of usage during school hours

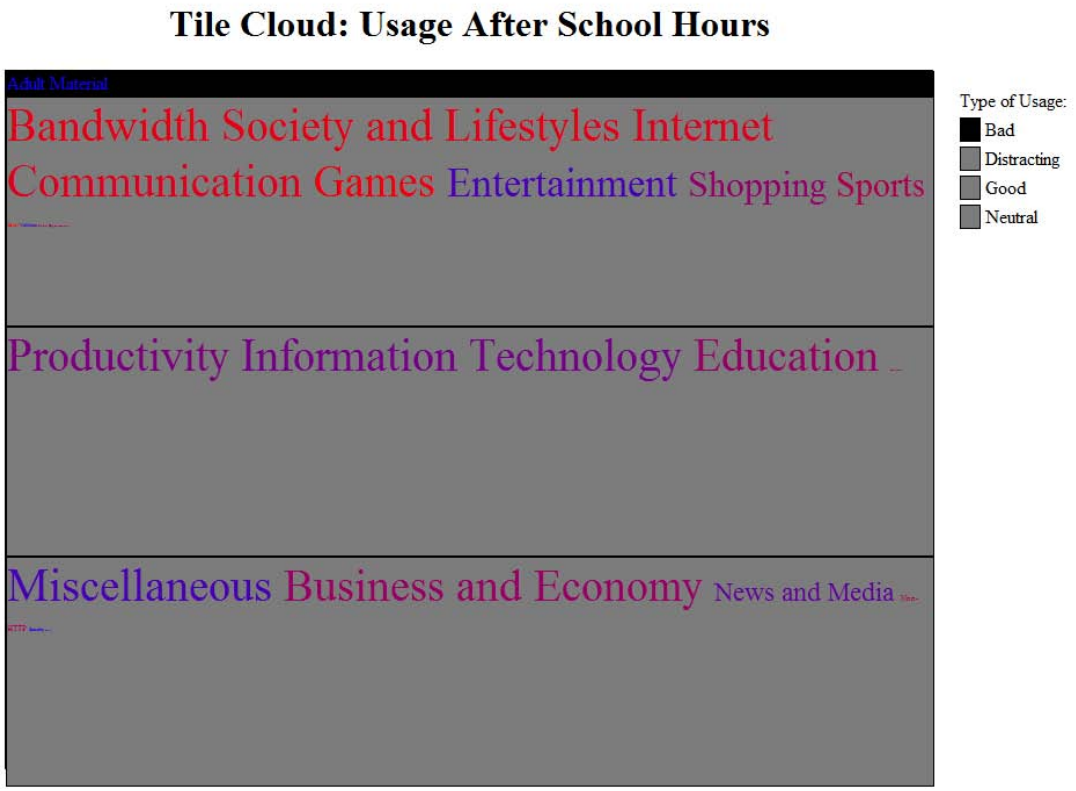

Figure 11: "Tile cloud" of usage after school hours

The "tile clouds" make it possible to compare usage as a function of all aggregated hits in a time period. In Figures 10 and 11, the "tile clouds" show that students use the internet in roughly the same ways in school and out of school, with some minor differences - non-HTTP traffic was significantly reduced at home (it is thought that in-school requests to the email server are categorized as non-HTTP, which would explain this difference). Also, requests for sites categorized as 'adult material' become more significant after school, though not by much. It is interesting, however, that there is a jump in adult material after school, even though students are aware that their web usage is still be monitored.. As the 'Distracting' usage type increases, it would seem that the 'Adult' category would increase with it. Because web requests in the 'Adult' category are blocked, it's possible that students access more entertainment-based websites after school hours, some of which may cause additional hits to advertisement servers which fall into the 'Adult' category.

It is also possible that the insights from this data could be correlated with other results from studies of teenagers - perhaps declining internet usage after 10:00pm is an indicator of times teenagers go to bed, or the slight dip in usage around 6:00pm could be indicative of the average time at which families eat dinner. All would be fascinating topics for further study.

CONCLUSION

Cary Academy's internet monitoring software, and similar software at schools like it, represents a huge wealth of detailed information about the way that modern teenagers live their lives. But with the high demands on Information Services staff, few - if any - schools have the time or motivation to run analysis on their data. Perhaps, as with Cary Academy, other schools could learn more about their usage habits and patterns, and by publishing those findings, belay parents' fears that their children are wasting time at school. Or, perhaps schools could use analytics to identify and track 'problem usage.' It is the authors' hope that this paper has provided a few more tools for schools to glean important, pertinent information about student internet use habits from their existing tracking and monitoring tools.

REFERENCES

- (1) Chang, Chung Y. 2011. "What Were We Talking About at Those SAS Conferences, or Let's Make Some Tag Clouds." *Proceedings of the SAS Global Forum 2011 Conference*. Available at http://support.sas.com/resources/papers/proceedings11/253-2011.pdf.

- (2) Cox, Amanda, Shan Carter, Kevin Quealy, and Amy Schoenfeld. "For the Unemployed, the Day Stacks Up Differently." New York Times. 31 JUL 2009, New York Edition BU5. Web. 13 Mar. 2012. Available at http://www.nytimes.com/2009/08/02/business/02metrics.html.

- (3) Lenhart, Amanda. 2009 Parent-Teen Cell Phone Survey. Pew Internet & American Life Project, May 6, 2007, accessed March 4, 2012. Available at http://www.pewinternet.org/Shared-Content/Data-Sets/2009/September-2009-Teens-and-Mobile.aspx.

ACKNOWLEDGMENTS

The authors would like to thank Peter Todd, Dmitry Manakhov, and the Cary Academy Information Services Department for their generous help in locating and downloading the raw data. We would also like to extend thanks to SAS Institute for making this paper submission possible.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Cary Academy
1500 N. Harrison Avenue
Cary NC 27513
Phone: (919) 677-3873
Fax: (919) 677-4002
http://www.caryacademy.org

Aaron Daniels
(919) 815-2132
aarondaniels43@gmail.com

Simon King
(919) 362-9776
sking72@gmail.com

Jacob Warwick
(919) 923-4351
jacobw125@gmail.com

APPENDIX

Additional code, such as the code to create the "envelope" graph and the "tile cloud" chart, may be found at http://www.jacobwdesigns.com/gf2012.