Towards an Empirical Model of WWW Site Response Times

Thomas A. Wagner
Department of Computer Science
University of Massachusetts at Amhest

UMASS Computer Science Technical Report 96-19
March 11, 1996

Abstract

Intuitively we believe that different WWW sites have different response time characteristics and that the characteristics change during the course of the day. These differences are important in the Cooperative Information Gathering paradigm because agents search and retrieve information on the WWW with real-time constraints. Thus agents must often plan which sites to search and choose between multiple possible sites -- response time is one factor used in making this choice. In this empirical study we show that different WWW sites have different response time characteristics and that for many sites response times are cyclical within a 24 hour period. We derive regression models for the sites that can be used in Cooperative Information Gathering and propose future research directions.

1. Introduction

World-Wide-Web (WWW) server response times appear to vary throughout the day and throughout the week. Queries submitted to URL search engines early Saturday morning seem to take much less time than queries submitted at noon Wednesday. Intuitively this is the result of the heavy load incurred by business type traffic, 8 am to 6 pm, in comparison to the relatively light level of activity that occurs on the average American's "sleep, family, and work around the house" day.

1.1 Rationale: Cooperative Information Gathering

From a research perspective, we are interested in this intuitive perception of varying response times because it relates to the Cooperative-Information-Gathering[1] (CIG) model. In the CIG paradigm, multiple agents work cooperatively to locate and retrieve information on the WWW to satisfy a user's informational needs. While the model involves many areas of research in computer science and AI, such as inference driven search, text extraction and satisficing control, the Design-To-Time[2] scheduling aspect of the paradigm is the motivation for modeling WWW site response times. In the CIG model, agents plan their actions to maximize solution quality while meeting real-time deadlines. This means that agents must often select between multiple methods for goal resolution, i.e., agents must plan what sites to search and often must choose one or few sites to search from many possibilities. Response times are important in this context because a site that takes a long time to search may be a less appealing alternative to one that takes a very short time to search, provided the two contain information of similar quality. In economic terms, the limited time resource mandates an opportunity cost perspective to WWW information gathering. Figure 1 illustrates the issue.

In this simplified example the high level goal or task is to *find-information-on-The-Simpsons*, a prime-time FOX cartoon sitcom and truly an interesting piece of popular culture. The task has three subtasks, one or more of which may be executed, and a response time distribution (artificial data) is associated with each subtask. The response time distributions are in keeping with our intuitive notion of sites having different response times at different points during the day. According to the response

time distributions, the FOX site is slowest at noon but the delta between peak and off peak is less pronounced than it is with The Simpsons Fan Club site, which also peaks at noon. In contrast, the Night Owl site is slowest during the late evening/early morning hours and fast during the afternoon.

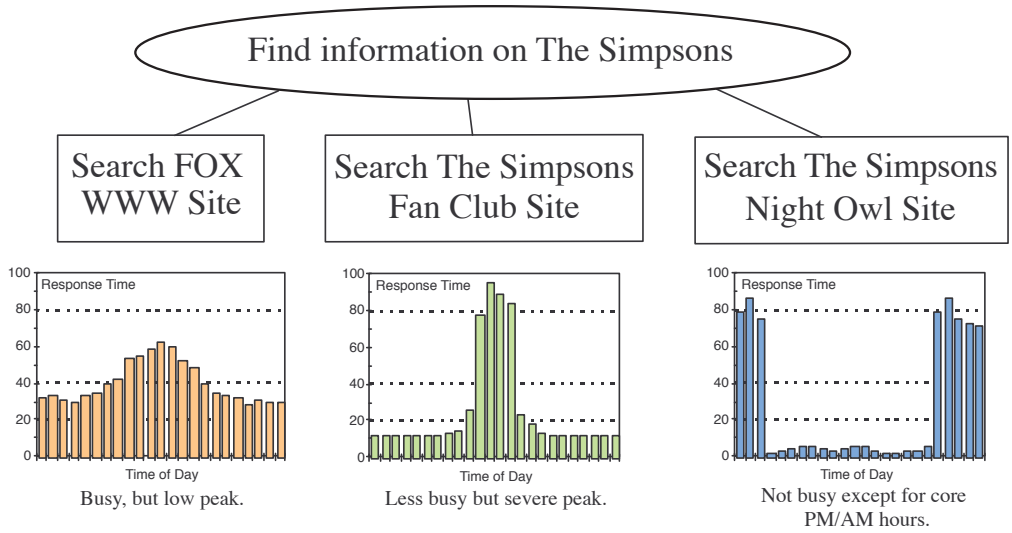


Figure 1 - A Simple CIG-like Task Structure

To illustrate the need for response time models, consider a case where the task is submitted at noon and the agent is given one minute to gather information. The Simpsons Fan Club site is not a viable search candidate because the expected response time is greater than the 60 seconds allocated to the search. Thus the agent may elect to search the FOX site once, because its expected response time is 60 seconds, or it may elect to search the Night Owl site many times. The choice in this case is driven by a distribution expressing expected quality or simply a quality-factor similar to Mycin's[3] certainty factors. If the agent is given 95 seconds instead of 60 seconds, the choices are to search the Fan Club site once, search the FOX site once and the Night Owl Site a couple of times, or to search the Night Owl Site many times.

In addition to basic WWW information servers, such as the Apple WWW site, agents may elect to search using URL search engines such as Yahoo, Infoseek, and Lycos. Response times are equally important in this context for similar reasons -- limited resources lead to an opportunity cost view. Clearly response time characteristics are an important enabling technology for real-time information gathering in the WWW domain.

1.2 Goals of the Study

The goals of this study are twofold. First, we must ascertain if different WWW sites have different response time characteristics. The only related work that even addresses time in the WWW domain is SavvySearch[4], a meta-search tool, that simply posts queries to other index-based URL search engines such as Lycos and Yahoo. SavvySearch uses a generic, static network load profile to determine the number of search engines used in a query. During heavy network load times, e.g., around noon, a smaller number of search engines are used than during off peak hours. SavvySearch does not search any other type of site nor is real-time an explicit issue. Additionally, the same network profile is used for all sites regardless of locale, distance between sites, or site performance characteristics. Our suspicion is that different sites exhibit different response time characteristics that change through each 24 hour period and that the differences are great enough to warrant individual/customized consideration of each site. Additionally, there is no empirical evidence to support the notion that network traffic load is highly related to WWW site response time. Thus a supplementary goal is to determine what measurable factors are related to response time and the

degree of the relation. Our final objective is to build predictive models of the sites that can be used in the CIG project.

2. Experimental Plan

The experimental plans that follow are second generation in that they reflect refinement based on information obtained from an initial two day pilot study. The experiments described below were run during a 10 day period from Wednesday, November 22, to Saturday, December 2nd. Thus the period includes one full work-week and the long Thanksgiving day weekend. The experiments are more observational in nature than manipulative as direction manipulation in this domain is very limited.

Prior to discussing the experimental plans, some definitions are necessary. We will use the term *search engine* to describe WWW URL search database engines such as Lycos, Yahoo, and Infoseek. We will use the term *information server* to denote a non-search engine web site, such as Microsoft's homepage and the term *site* when differentiation is not important. *Response time* is context dependent. If *response time* is used in the context of a *search engine*, it denotes the time required to submit a query to the search engine, for the search engine to process the request, and for the results to be transmitted back to the requester. In the context of an information server, response time denotes the time required to submit an http get request[1] and to obtain the results. The primary difference is that an http get request involves pulling the requested object off the disk and sending it to the requester while the search engine query requires additional processing by the database application.

The information servers sampled in this study are: Netscape, Microsoft, Apple, Playboy, Prodigy and MTV. The reason the entertainment sites such as Playboy, Prodigy, and MTV are included in this study is to see if they have response time characteristics that are different from those of more business and computer related sites such as Netscape, Microsoft and Apple, and to avoid a sampling bias. The search engines observed in this study are: Lycos, Yahoo and Infoseek.

2.1 Experimental Plan for Information Servers

The experimental plan for WWW information servers is as follows:

- Sample the information server response times by submitting a batch of three http get requests. The reason for the batching approach is that floor effects were observed for single http get request submissions during the pilot, particularly during the low-load periods. While the batch contains three different get requests, the same batch is used for each submission to enable pairwise comparisons and to eliminate variance between the number of bytes received in response to each batch request.

- Record response time per batch along with the number of bytes received in response to the corpus of requests. By nature, each http get request returns information of a different size. For example, a document at the Apple site is unlikely to be the same size as a document at the Microsoft site. The size of the documents is thus important so we can normalize response times across the sites if necessary.

- Record a network load factor for the local domain and the local-to-remote path. The load factor is obtained by bouncing five 8k packets on a circular path to and from a local machine and to and from the remote machine. The packets travel according to a non-ignorable TCP/IP feature, the UNIX ping facilitates this process. Researchers at Boston University have also experimented with this approach [5].

---

[1] An *http get request* is the normal order of business for WWW browsers. When a user clicks a "hotlink" the browser sends an http get request to the http server, i.e., it asks for a particular document.

- Control the local CPU load by running the sampling processes at a priority level equal with the root processes. Lowering the root processes below the sampling processes is not an option because the root processes manage the network activity. Record the size of the local run queue, indicative of CPU load, using five, ten, and fifteen minute moving averages.

- Submit the http get request batches every three minutes with a minimum delay between submissions of two minutes. This somewhat counter-intuitive sampling process is a derivative of the pilot study. Sampling at a regular interval results in fewer data points when response times are slow and more data points when response times are fast. Thus the non-uniform delay periods help to more evenly distribute the sample points. We found the delta in most cases to around 60 seconds, hence the one minute buffer zone.

The experimental plan is designed to address several primary issues. First and foremost, the sampling process must not add significantly to the information server's load. Obviously, affecting the load will result in skewed or inaccurate data. The objective is to have the sampling activity account for less than one percent of the information server's load. Most commercial sites publish a hit frequency of 500,000 to 1,000,000 hits per day, which factors to 350/700 hits per minute. The corpus of three http get requests resides within our one percent limit. We should note that exact hit statistics are not available for all sites otherwise we could custom tailor the sampling delay for each site.

Delays between http get requests are equally important. As caching both at the disk level and at the http sever level are likely, identical requests cannot occur one after the other or again the data will be skewed. This is more likely to occur during off peak hours and the resulting error would exaggerate the difference between off peak and on peak hours -- the off peak response times would appear to be faster than they are in actuality. Hence the corpus of three get requests contain requests for different documents and the minimum two minute delay between submissions.

In summary, the experimental variables are:

- Batch size and type of http get requests contained in the batch - manipulated
- Interval between batch submissions - manipulated
- CPU run queue size/load - observed and partially controlled
- Response time and bytes received - observed
- Local and remote network load factor - observed
- Time of day - observed

2.2 Experimental Plan for Search Engines

The experimental plan for the search engines is similar to that used for the information servers. Key differences center around possible sampling bias with respect to the database application. In other words, for a particular set of keywords, it is possible for one URL search engine to dramatically outperform another while they generally perform the same in everyday use. We are interested in general response times, therefore careful consideration of the keywords used is necessary. The experimental plan is as follows:

- Sample the search engine response times by submitting a batch of three queries. The reason for the batching approach is the same as in the information server experimental plan -- floor effects were observed for single queries in the pilot. The terms submitted are "asdfjkl," "intelligence," and "data parallel program visualization and animation." Thus the keywords include on obscure reference or nonsensical query, one very general query likely to result in a large number of hits, and one multi-term query with opportunities for stemming and insignificant term removal, e.g, "and."

- Record response time per batch along with the number of bytes received in response to the corpus of requests. As with the information severs, search engines contain different information and therefore may return very different amounts of information in response to each query. However, as most search engines provide some contextual verbiage for a small subset of the matched documents, i.e., typically they return the best 10 or 20 matches, the size difference is much less pronounced than it is with the information servers. Thus the number of bytes received is less important in this case.

- Record a network load factor for the local domain and the local-to-remote path.

- Control the local CPU load by running the sampling processes at a priority level equal with the root processes.

- Submit the query batches every three minutes with a minimum delay between submissions of two minutes.

Similar to the variables for the information servers, the experimental variables for the search engines are:

- Batch size and type of query terms contained in the batch - manipulated
- Interval between batch submissions - manipulated
- CPU run queue size/load - observed and partially controlled
- Response time and bytes received - observed
- Local and remote network load factor - observed
- Time of day - observed

3. Results: Response Time Overview

3.1 Information Servers

The response times for the information severs is shown in Figures 2, 3, 4, 5, and 6. The Y axis indicates response time in seconds and the X axis is determined by a function of date and time. The time series show the response times for the week of Monday, November 27 through Saturday, December 2. The leading days are omitted for size considerations and because the more interesting cyclical behavior appears during the business week. All time series have been smoothed once with a 4253H median smoother.

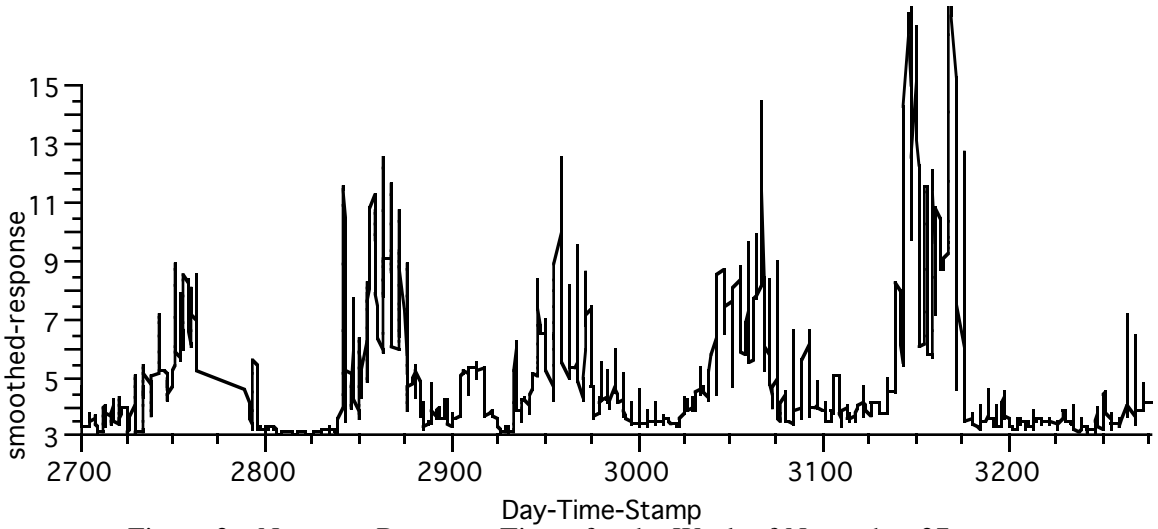


Figure 2 - Netscape Response Times for the Week of November 27

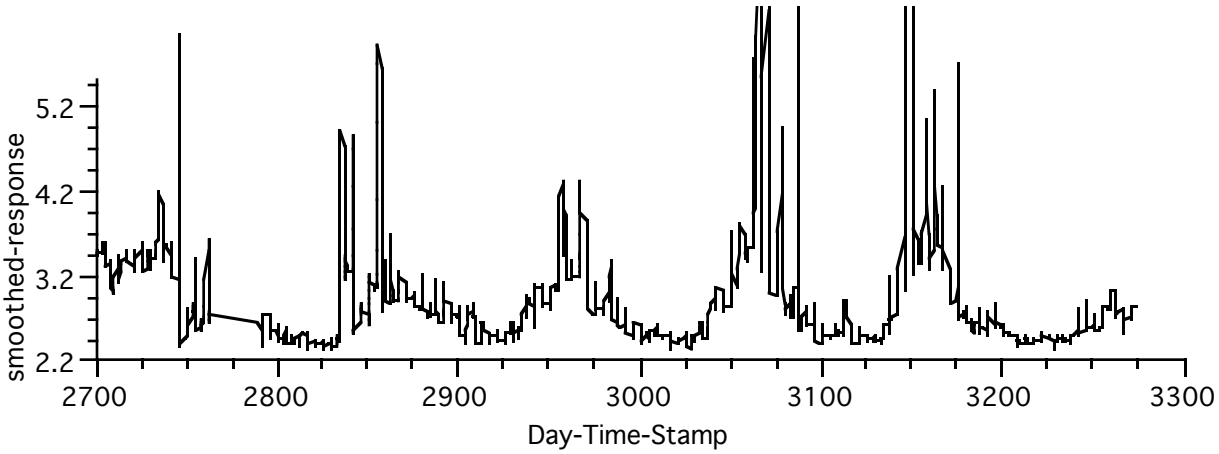


Figure 3 - MTV Response Times

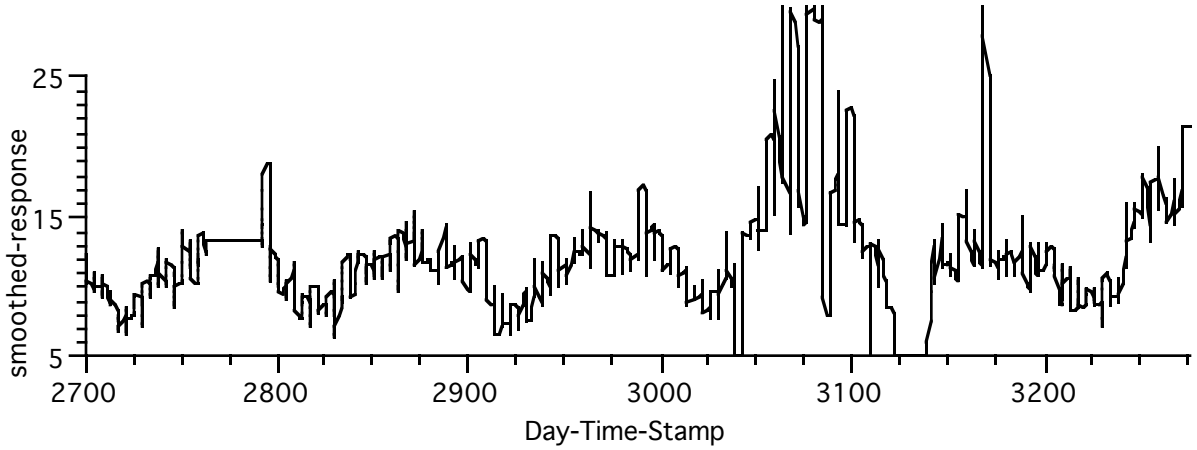


Figure 4 - Playboy Response Times

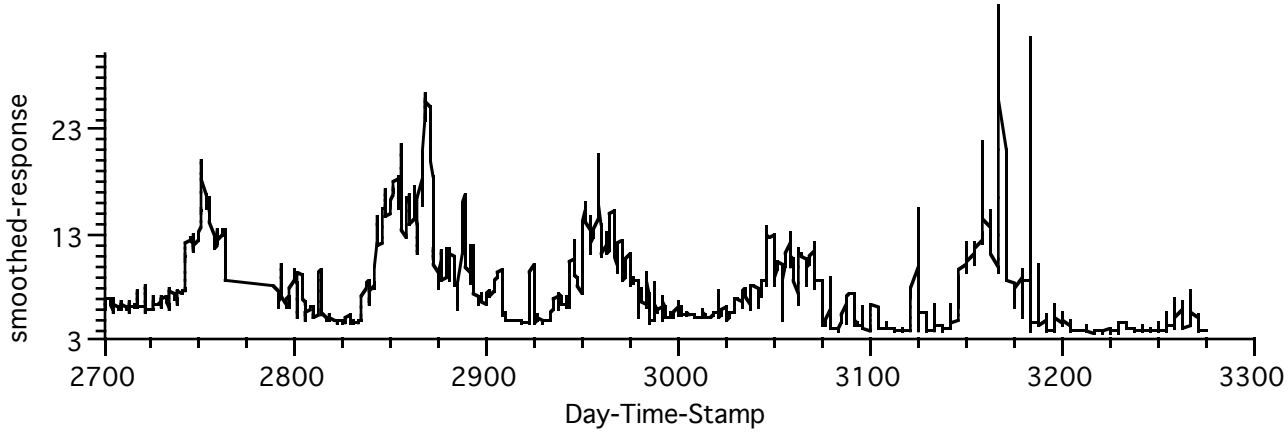


Figure 5 - Microsoft Response Times

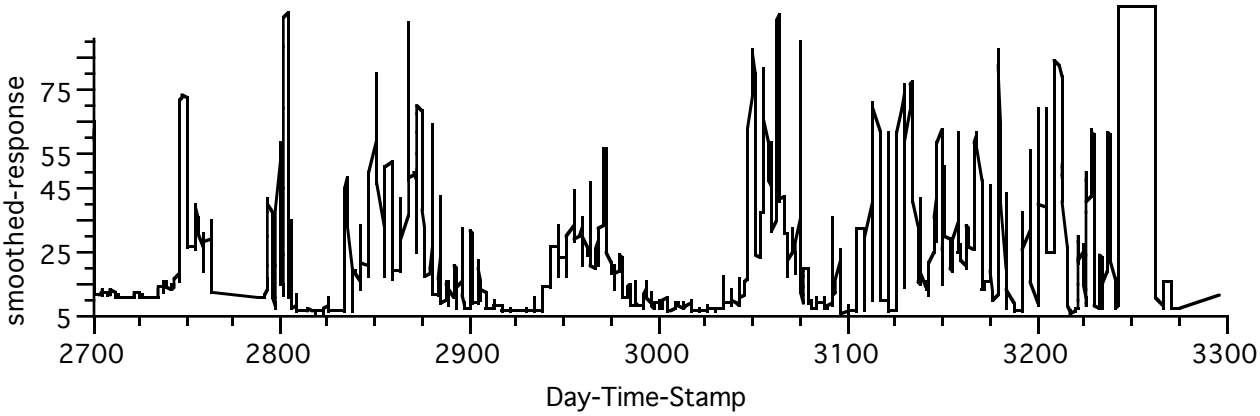


Figure 6 - Apple Response Times

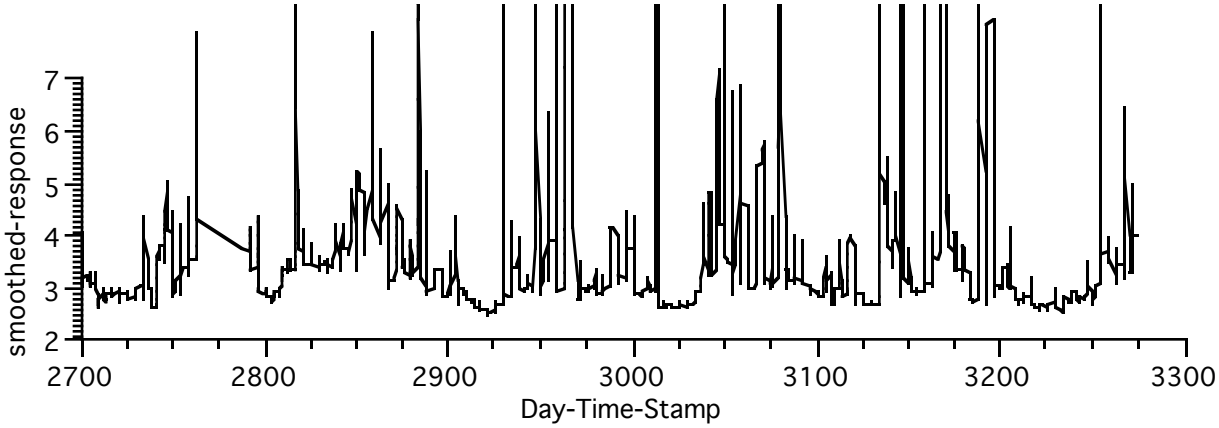


Figure 7 - Prodigy Response Times

The leftmost position of Figure 2, where the X axis label is 2700, indicates the response time at November 27, 0 hundred hours or 12 am. Similarly, 2800 is November 28 at 12 am. Midway between 2700 and 2800 is 12 pm Eastern Standard Time (EST) -- each axis tick is worth six hours. The labels 3100 and 3200 are in actuality December 1st and 2nd respectively, the labeling keeps the dates in sequence without adding the month to the date and time label.

Clearly cyclical behavior abounds. According to the Netscape response times shown in Figure 2, the Netscape information server is slower between business hours than it is during the evenings. The response time trend starts increasing between 8 and 10 am EST and increases generally until sometime between 12 pm and 3 pm EST. Intuitively the increase is the result of the business day beginning on the East coast, followed by the Central and Western time zones.

The anomaly that appears on Monday the 27th between 12 pm and 6 pm EST is the result of the data gathering workstation going down. On some graphs it appears horizontal and others it appears diagonal, but it always takes the form of a straight line. The difference in the slope is the result of a combination of smoothing and where the first datum fell when the machine resumed sampling.

It is interesting to note that the leading and trailing slopes vary from site to site. Netscape has fairly steep edges in contrast to the Playboy site, Figure 4, which has more gradual increases. Comparing Netscape to MTV and Playboy, it appears that the entertainment oriented sites have different response time characteristics. The MTV site exhibits slower response times until later in the day, often only falling off around midnight. Similarly, the Playboy site's response times do not fall until the interval

between 3 am and 6am. If we equate slow response times with user induced load, which is logically appealing even though it cannot be verified, the difference in the response times is attributable to the fact that the sites contain inherently different information. The Netscape information server contains corporate and product information. The MTV site contains information on MTV's cartoon series such as Beavis and Butthead, sound bytes, and snap-shots as well as program schedules, music news and the like. The Playboy site contains subscription information, but also biographical information on the hottest new playmates and several unpublished nude photos of each playmate. The data suggests that playmate information is in demand for more hours of the day -- perhaps being accessed more by home users during the non-business hours or by more European users.

The Apple and Prodigy information servers both exhibit more variance than the other sites during the same period. We cannot account for this behavior, but it leads to an important point. Understanding what is happening at a remote site is inherently difficult. We cannot ascertain if response time changes are the result of backup schedules, changes in hardware, or user incurred load. By the same token, response time is a good item to measure because it lumps these factors together in a single number. Basically, response time is "all that we can get" but it is also sufficient for the CIG planning needs and *the* relevant measure of performance.

3.2 Interesting Behavior

Economists, who work frequently with time series, often account for major holidays in their models. For example, General Motors does not produce automobiles during the interval from December 23 through January 1. Therefore a model of automobile production must reflect this behavior as General Motors is the largest automobile manufacturer in the world by a considerable margin.

The response times for many of the sites exhibited interesting behavior on Thanksgiving day. The MTV site response times for Thursday, November 30 and Thanksgiving day are shown in Figures 8 and 9 (spline smoothed with the line accounting for 80% of the variance). Unlike the response times for the 30th, a business day, on Thanksgiving day response times peak around 10 am and fall off again around 1 pm. The second peak for the day occurs after 6 pm. While strong claims are certainly not justified, the difference in behavior suggests people leaving the computer to eat Thanksgiving dinner and returning once the relatives have left. Netscape, Figures 10 and 11, exhibits similar behavior. These observations are not central to this study, but are interesting just the same. Perhaps our future models should also include special cases for holidays and other important events.

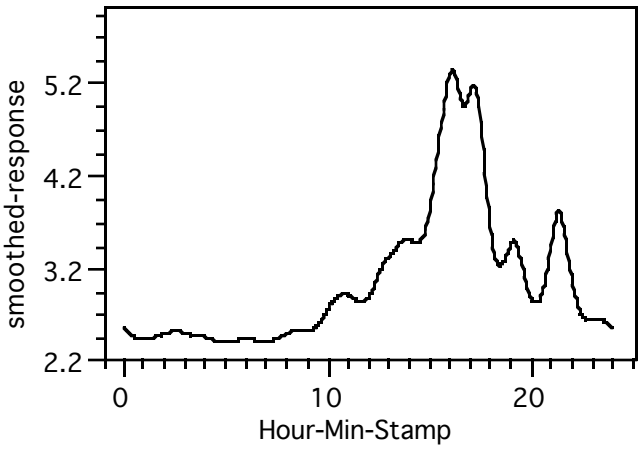


Figure 8 - MTV Response Times on the 30th

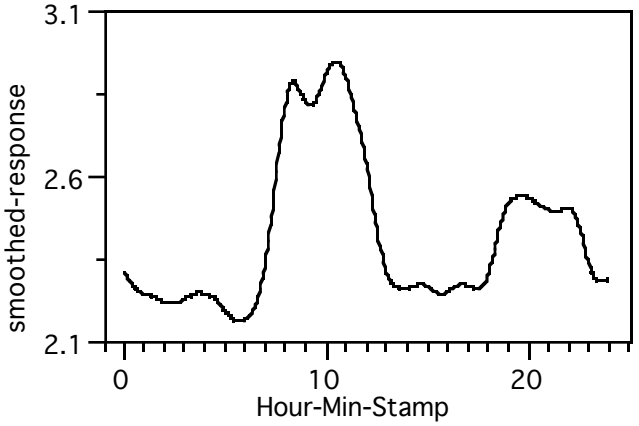


Figure 9 - MTV Response Times on Thanksgiving

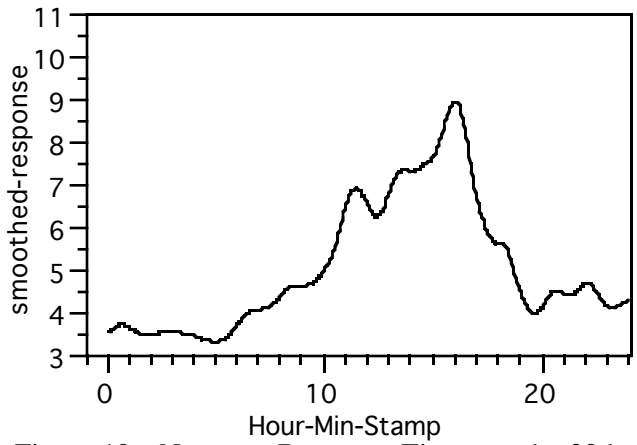


Figure 10 - Netscape Response Times on the 30th

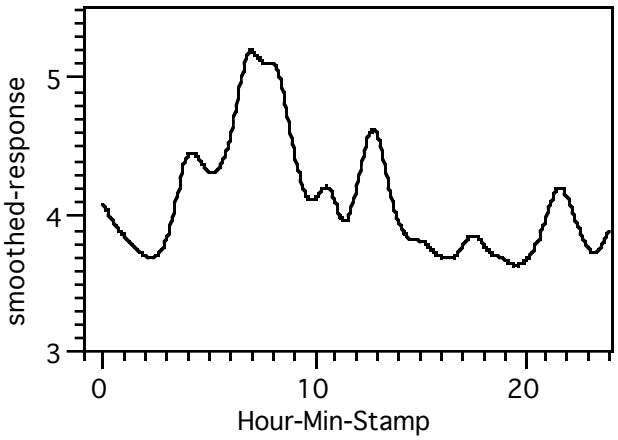


Figure 11 - Netscape Response Times on Thanksgiving

3.3 Search Engines

The response times for Infoseek and Yahoo are shown in Figures 12 and 13. The Lycos data is incomplete and hence omitted. Infoseek exhibits a cyclical response time behavior similar to that observed of the information servers. Generally, Infoseek is slow during business hours. However, the difference between "slow" and "fast" is much more pronounced than at the other sites. As the version of Infoseek sampled during this study is the "free" or non-subscription version, it is the default search engine used by all Netscape browsers, i.e., the *Netsearch* button an all Netscape brand browsers points to the free Infoseek engine. Since Netscape has a lion's share of the browser market, it is reasonable to assume that many people rely primarily on the free Infoseek search engine to locate URLs. The bottom line is that when Infoseek is slow, *it's really slow*.

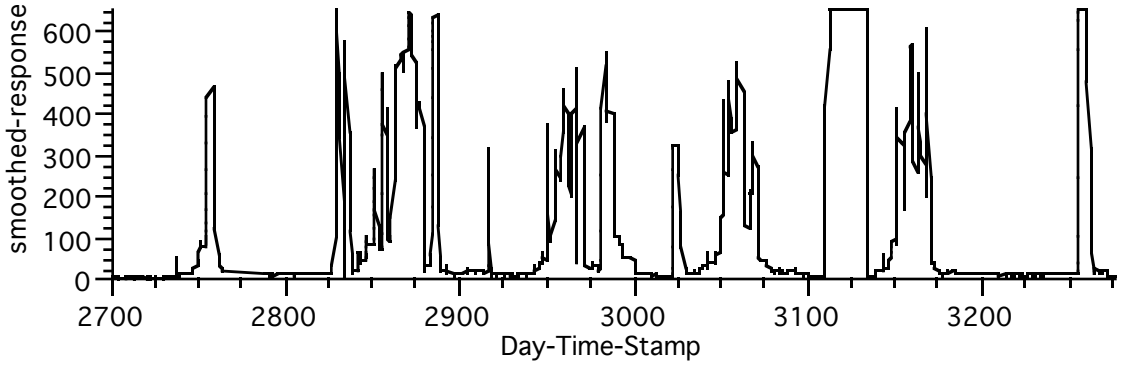


Figure 12 - Infoseek Response Times

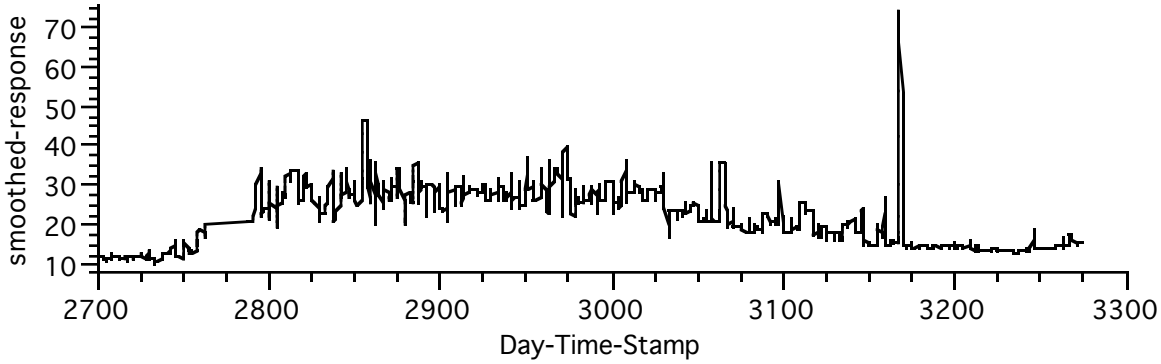


Figure 13 - Yahoo Response Times

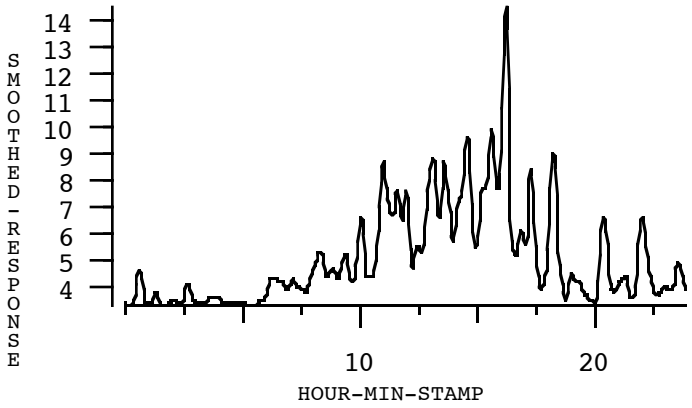
In contrast to the polite cyclical behavior observed elsewhere, Yahoo's response times, Figure 13, do not vary with the 24 hour clock. At first glance, one might assume that Yahoo has either completely ample or completely inadequate resources. However, an alert observer will quickly note that if Yahoo has completely inadequate resources the response times will all be pushed towards a ceiling. As local minima and maxima abound, and are accompanied by a decreasing trend as the week progresses, the proper interpretation is that Yahoo has completely adequate resources to meet consumer demand. Since the same query batch was submitted to all search engines, comparisons between the Y axis are meaningful. In general, during the afternoon hours, Yahoo will give a response time an order of magnitude lower than Infoseek.

3.4 Preliminary Observations

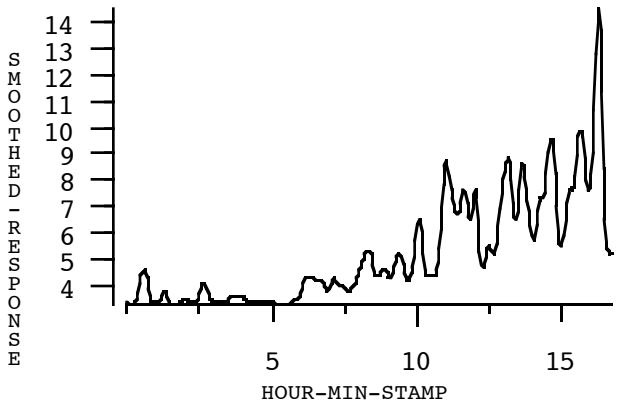
Recall that one of the goals of this study is to determine if different sites have different response time characteristics. Clearly, the data supports this notion. Furthermore, the cyclical behavior exhibited by almost all sites suggests that time of day and response time are highly correlated. The second objective of this study is to attempt to build models of site response times that can be used by information gathering agents to plan and schedule their activities. In order to build predictive models, we must determine the degree of influence that each measured factor has on response time.

4. Netscape: A Closer Look

Because the datasets are numerous and large, it is helpful to examine a small subset of the data to "get a handle" on the behavior and interactions of all the variables. Accordingly, Figure 14 shows the response times for the Netscape information server on Thursday the 30th. As the response time behavior is cyclical rather than linear, we cannot use traditional linear tools, such as linear regression, to evaluate its behavior -- the correlation coefficient is approximately zero. To simplify matters, Figure 15 displays the response times for the 30th lopped off at around 3:45 pm. The trend for this portion is essentially increasing.



Figures 14 - Netscape Response Times for the 30th



Figures 15 - Netscape Response Times for the 30th Lopped off at 3:45 pm

Since the trend is increasing, linear regression is applicable and Figure 16 shows the regression plot of response time with respect to the 24 hour clock. Table 1 contains the fit information. In the case of the raw response time, the R^2 value is 65% thus the regression line accounts for 65% of the variance and is considered a good fit. The P-value from the f-test for this case is significant. Applying a log transform improves the fit driving R^2 up to 75%. Clearly the time of day plays an important part in response time behavior, at least for this small data fragment.

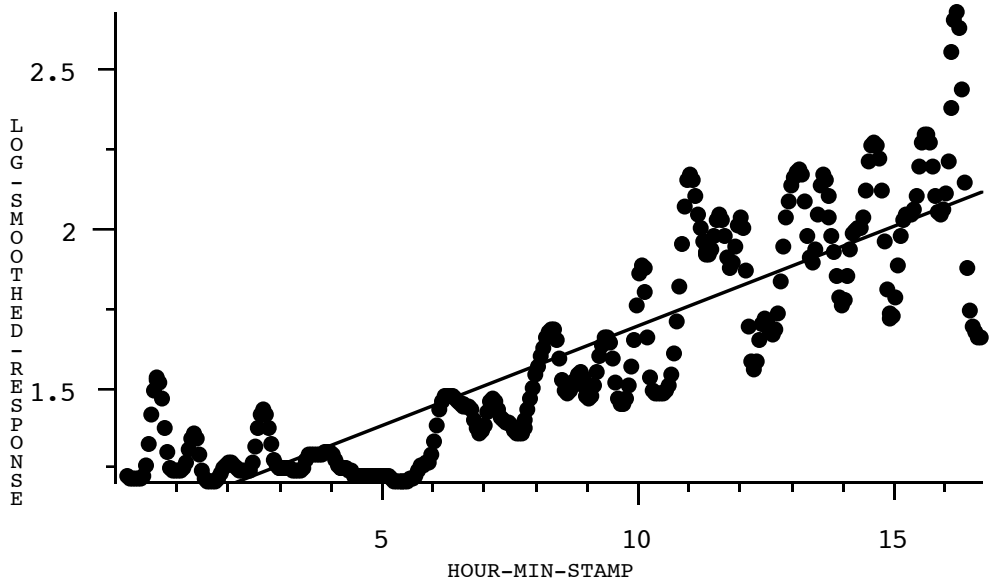


Figure 16 - Regression Plot of Log Transformed Response Time and Time of Day

Variable	Slope	Intercept	R^2	Std Error of Slope	P-Value
Smoothed Response	.35	2.35	**.65**	.014	0.000
Log Smoothed Response	.06	1.07	**.75**	.0020	0.000

Table 1 - Regression Results of Response Time versus Time of Day

Now consider the other recorded variables. The non-local network load factor is displayed in Figures 17 and 18. The Y axis is the average time required to send five 8k packets to the remote host in hundredths of a second. The X axis is the familiar date/time stamp. Figure 17 contains the factor for the week of November 27 and Figure 18 for November 30th only. Cyclical behavior is evident in the full week time series, but less pronounced than with the response times. The cyclical behavior in Figure 18 is less obvious, but this is partly an artifact of scaling -- the peaks above 1.7 seconds skew the scale and obscure the cyclical trend.

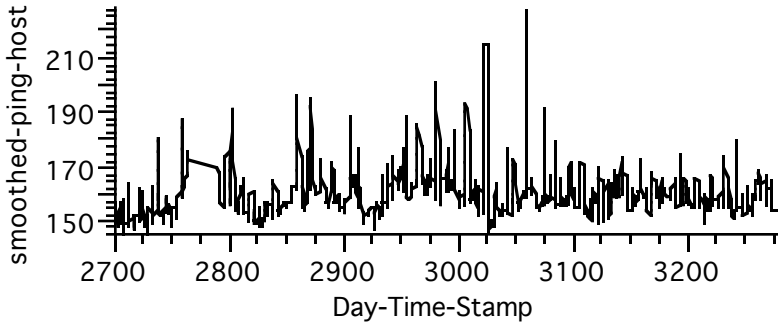


Figure 17 - Non-Local Network Load (To Netscape)

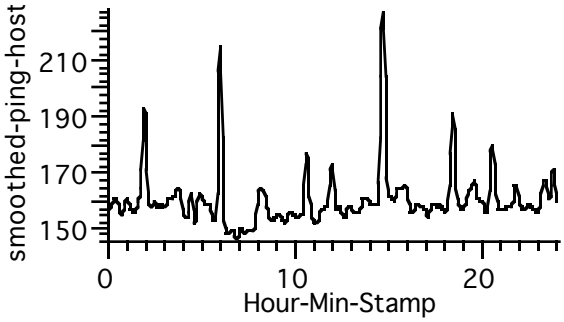


Figure 18 - Non-Local Network Load for the 30th

The local network load factor, Figures 19 and 20, exhibits much less regularity both on the weekly interval and on the 30th. The 1 and 5 minute mean run-queue sizes, the CPU load factor, for November 30 are shown in Figure 21. Generally, in cases where the run-queue size is less than one job the machine is considered idle.

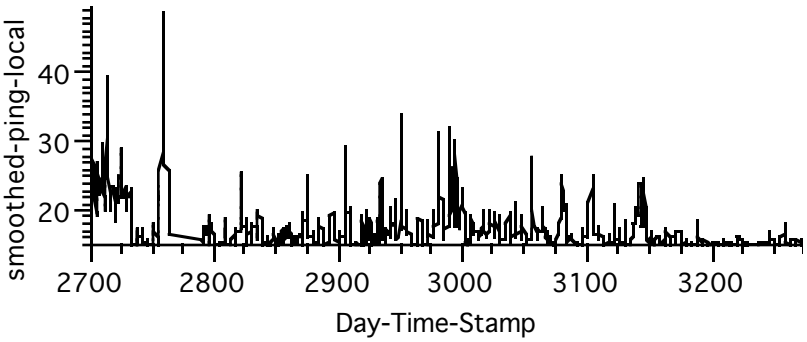


Figure 19 - Local Network Load

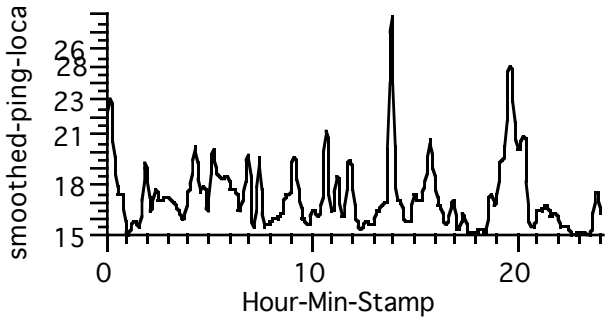


Figure 20 - Local Network Load for the 30th

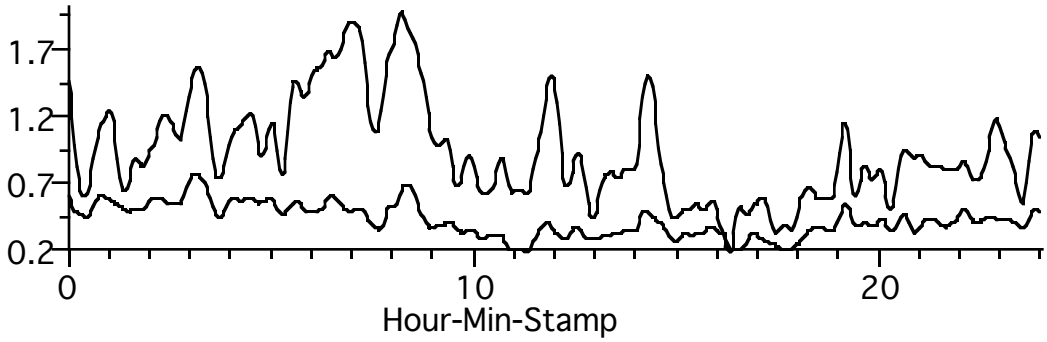


Figure 21 - 1 and 5 Minute Run Queue Sizes for the 30th (5 is below 1)

Regression results of non-local network load, local network load, and run-queue size with respect to response time are shown in Table 2. The R^2 values for the non-local and local network factors are tiny and the P values of the f-tests indicate the results are not significant. It appears for this fragment of data that the network load does not significantly affect response time. Interestingly, the one minute run queue appears to be related to response time, though the R^2 values are very small. Figure 22 shows the regression plot of log transformed response time and log transformed one minute run queue. Table 3 contains the relevant regression statistics. As the run queue size decreases the response time increases, i.e., as the CPU load decreases the response time increases. The semantics of this are frightful and the regression is included only for pedagogical reasons. The lesson is that when the R^2 value is low, the relation should be questioned regardless of the P value. Clearly the relation is an artifact of this data fragment.

Variable	Slope	Intercept	R^2	Std Error of Slope	P-Value
Ping Remote	.0010	5.07	**.006**	.0007	.18
Ping Local	-.02	17.7	**.003**	.02	.32
1 Minute Run Queue	-0.03	1.3	**.16**	.004	0.000
Log 1 Minute Run Queue	-.04	.3	**.24**	.004	0.000

Table 2 - Regression Results of Variable Vs Time (Chopped off at 3:45 or 16.75)

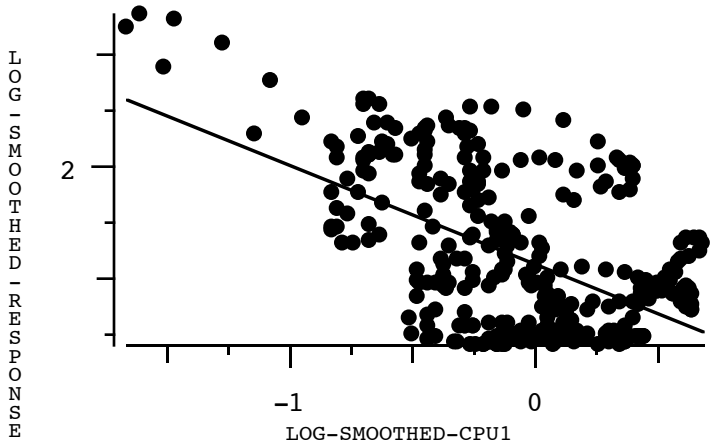


Figure 22- Regression Plot Log Smoothed Response Time Vs Log Smoothed 1 Minute Run Queue

Slope	Intercept	R^2	Std Error of Slope	P-Value	Partial Correlation (t constant)
-.44	1.6	.30	.04	0.000	-.30

Table 3 - Regression of Log Smoothed Response Vs Log Smoothed 1 Minute Run Queue

To complete our micro-evaluation, Figure 23 displays a flat regression model of response time and the other, most highly related variables. Again, the semantics of the relationship between the CPU load statistic and the response time call the value of the normalized coefficient, Beta, into question. Additionally, the cyclical trend of the non-local network load is out of phase with the response time cycle thus diminishing any relationship (of course, this could also be true in all cases also and hence the diminishment consistent and valid) that may exist. Given the high R^2 of time of day, it is unlikely that the non-local network traffic is a major player in the response time -- at least for this particular dataset.

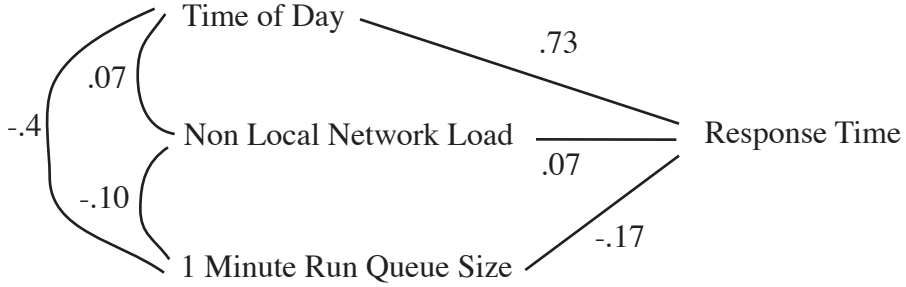


Figure 23 - Flat Regression Model (Selected Multiple-Variate Regression)

5. Dealing with Cycles

5.1 Non-Linear Transform

The major dividend of our micro-examination in Section 4 is that time of day and response time are highly related. Furthermore, response time is cyclical during the 24 hour period and thus defies normal linear regression. Due to the size of the data, the number of days and the number of sites, partitioning the response time data into increasing and decreasing trends on a daily basis is an extremely unpleasant prospect. Accordingly a non-linear transform is needed. Figure 24 illustrates

such a transform, called Paul's Transform, aptly named after the originator, Paul Cohen[6]. At the trailing end of the increasing trend, a local maximum is obtained and all subsequent points are reflected over a horizontal line at that point. Thus the decreasing trend that typically appears during the latter portion of the day is realigned to continue the increasing trend from the first half of the day.

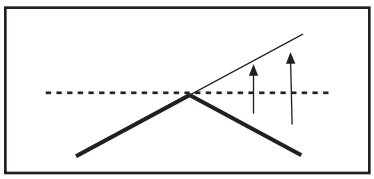


Figure 24 - Paul's Transform

The regression line fitted to the transformed response time for Netscape on the 30th is shown in Figure 24 and the fit data appears in Table 4. The R^2 value of 88% is high and the P value from the f-test is significant at any level.

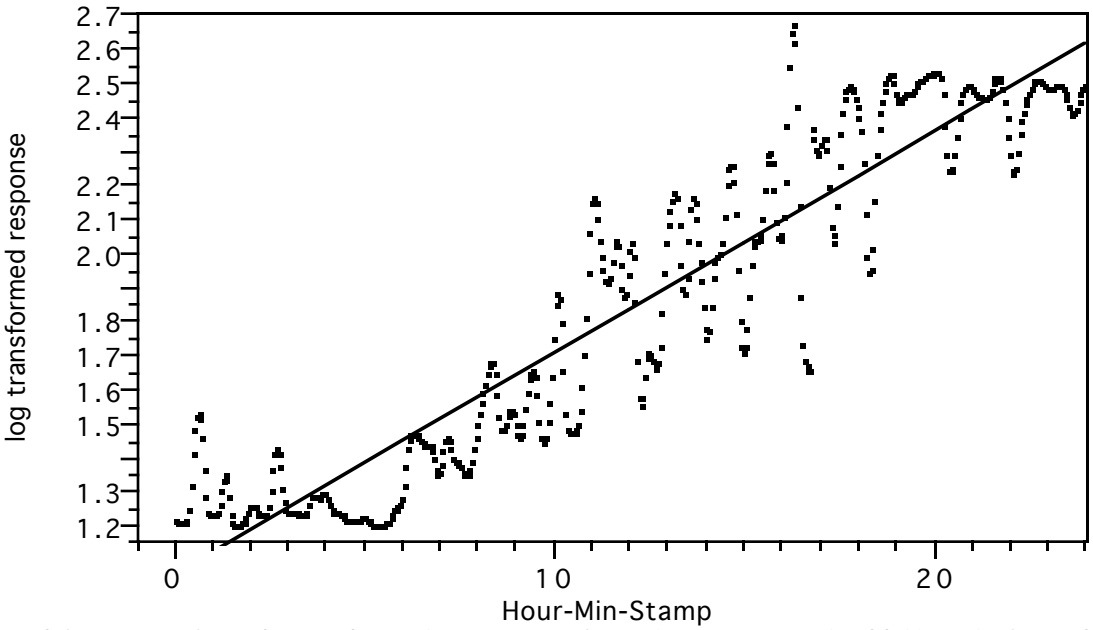


Figure 25 - Regression of Transformed Response Times (Netscape on the 30th) and Time of Day

Summary of Fit

RSquare	0.877652
RSquare Adj	0.877395
Root Mean Square Error	0.168308
Mean of Response	1.837161
Observations (or Sum Wgts)	479

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	96.92837	96.9284	3421.699
Error	477	13.51224	0.0283	Prob>F
C Total	478	110.44061		0.0000

Table 4 - Regression Results

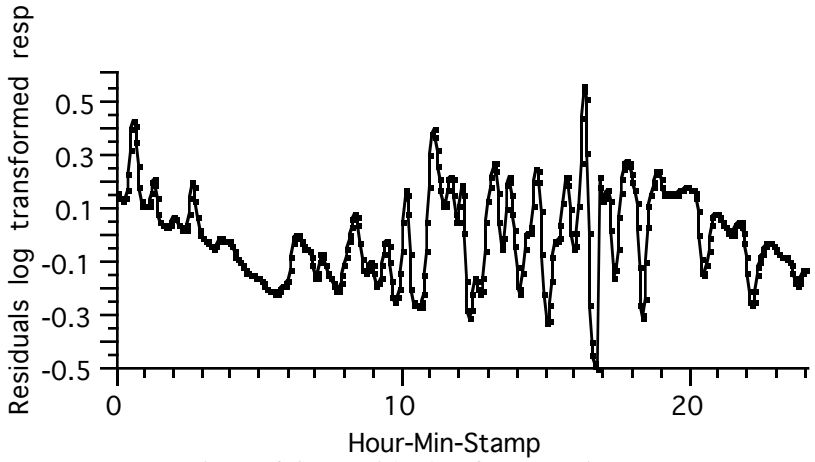


Figure 26 - Residuals of Regression

The residuals of the regression are shown in Figure 26. The envelope in the center of the plot indicates that the regression model is less accurate during the afternoon hours, i.e. the variance unaccounted for by the model is greatest in the afternoon. The same transform, i.e., the transform from the 30th with the local maxima (X,Y) value from the 30th, was applied to all of the business days. The regression plot is shown in Figure 27 and the characteristics of the regression in Table 6. The high R^2 values for all days except the incomplete day of December 2, combined with the significant P values, is extremely encouraging. Not only is response-time very highly correlated with time of day, but the different days of the week exhibit very much the same response time characteristics. To further support this observation, Figure 28 and Table 7 show the results of applying Paul's Transform to the response times for all days, lumping the days together, and regressing with respect to time of day. The R^2 value of 72% is high in keeping with the observation that the days exhibit similar response time cycles. This bodes well for the notion of using regression models in information-gathering planning and scheduling.

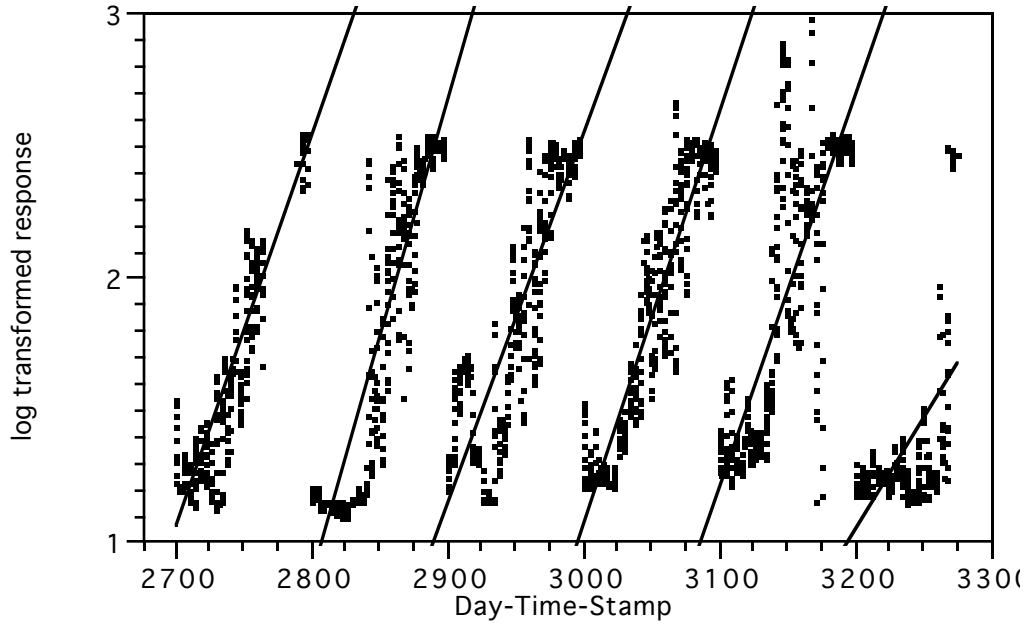


Figure 27 - Paul's Transform for the Entire Week (Netscape Response Times)

Day	R^2	Root Mean Square Error	Analy of Var F	Analy of Var P
27	.85	.17	1909	0.00
28	.85	.22	2527	0.00
29	.73	.25	1306	0.00
30	.88	.16	3406	0.00
1	.65	.32	853	0.00
2	.30	.28	150	0.00

Table 6 - Regression Results by Day

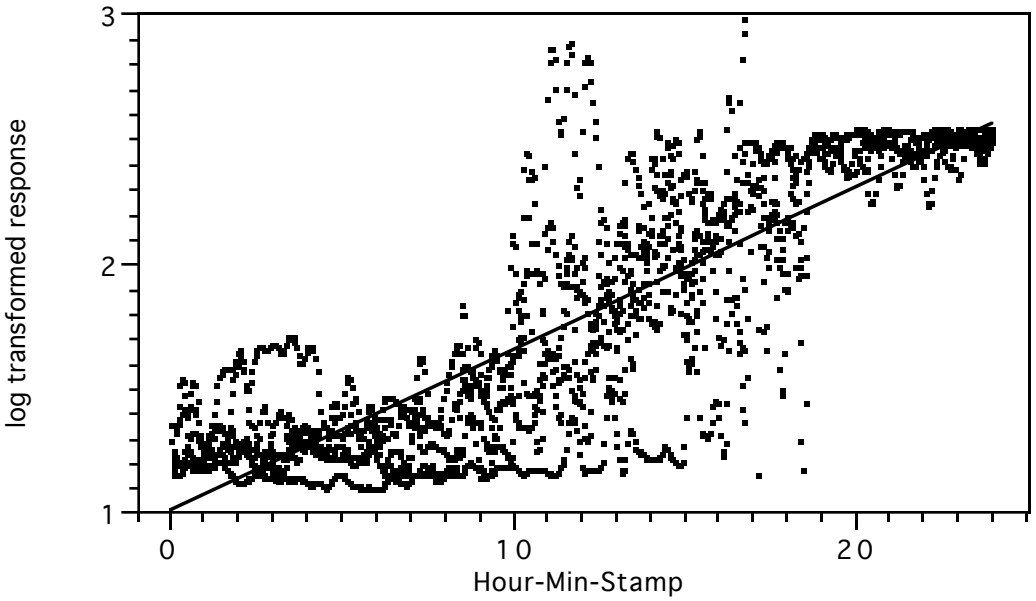


Figure 28 - Paul's Transform All Days at Once

R^2	Root Mean Square Error	Analy of Var F	Analy of Var P
.72	.28	6505	0.00

Table 7 - Paul's Transform All Days at Once

5.2 Polynomial Regression

An alternative to transforming the data is using non-linear regression models. Figure 29 and Table 8 shows the results of fitting a six degree polynomial regression line to the cyclical data of Netscape on the 30th. While the results are adequate, the model accounts for less of the variance than the transform from Section 5.1. Since high degree polynomials are difficult to work with and the fit is not any better, is in fact worse, than the transformation approach, we will focus on the non-linear transform in our future efforts.

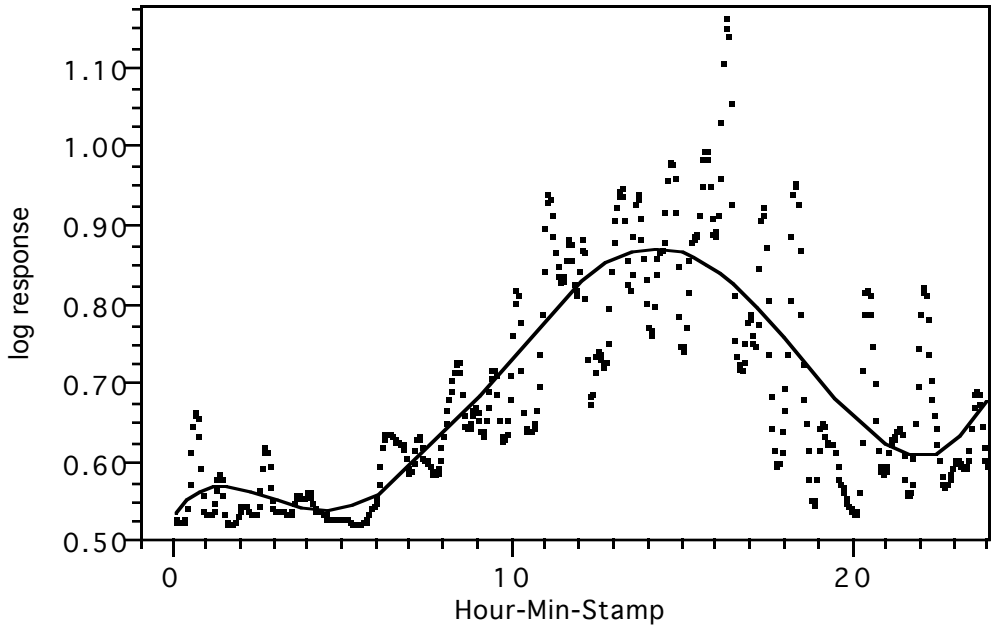


Figure 29 - Polynomial Regression, Degree 6, Netscape Response Times on the 30th

Summary of Fit	
RSquare	0.683344
RSquare Adj	0.679319
Root Mean Square Error	0.07873
Mean of Response	0.683634
Observations (or Sum Wgts)	479

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	6.3135031	1.05225	169.7628
Error	472	2.9256251	0.00620	Prob>F
C Total	478	9.2391282		0.0000

Table 8 - Polynomial Regression Results

6. Conclusion and Future Directions

The goals of this study are to determine if WWW sites have different response time characteristics and to attempt to model the response times if indeed they are different. Clearly, the sites have different response times and cyclical behavior based on the 24 hour clock predominates. Using Paul's Transform we have built a simple linear regression model of the Netscape information server that accounts for 72% of the variance observed during the 5 day business week of November 27.

While the statistical literature offers many advanced tools for cyclical time series analysis, such as autoregression[7][8] and the ARIMA[9][10] model, the simplicity of our non-linear transform and its good performance precludes further exploration unless warranted by the data. Our future work is clearly defined. The non-linear transform must be applied to the data for all WWW sites and if the results are positive, R^2 values are high and fit statistically significant, we will have good regression models for the observed data. If the results are not on the same tier as the Netscape results, then alternate tools such as the ARIMA model must be considered.

Once models are built we must test the models to determine their predictive value. While a testing procedure has not yet been defined, the newly sampled data should also be recorded in order to further refine the models if performance is lacking. An interesting characteristic of the WWW domain is its ever changing and dynamic usage patterns 11]. Growth of the WWW is tremendous and it is highly likely that sites will frequently change their hardware capacities to keep pace with the increasing number of "webbers." It seems probable that sites will thus exhibit periods of gradually increasing response times followed by abrupt improvement as capacities are increased. Regardless of the form of the change, the models used by information gathering agents must adapt and evolve as the environment changes. In short, the agents must improve the models based on the observed performance, but the updating or learning algorithm(s) will require further thought.

In summary, the preliminary results of this study are promising. Different sites exhibit different response times at different times during the day thus response time must be considered during planning. Furthermore, it appears that a simple one variable linear regression model will yield good predictive value and can thus satisfy our planning needs at this stage. The model must be extended to more sites and then tested. Model updating is a less near-term, but still important, issue.

7. References

[1] Tim Oates, M. V. Nagendra Prasad and Victor R. Lesser "Cooperative Information Gathering: A Distributed Problem Solving Approach," *UMASS Technical Report 94-66*, 1994.

[2] Alan Garvey and Victor Lesser, "Design-to-time Real-Time Scheduling," *IEEE Transactions on Systems, Man and Cybernetics - Issue on Planning, Scheduling and Control*, Vol. 23, No. 6, 1993.

[3] E.H. Shortliffe, *Computer Based Medical Consultations: MYCIN*, New York, American Elsevier.

[4] Daniel Dreilinger and Adele Howe, "An Information Gathering Agent for Querying Web Search Engines," *Department of Computer Science Unpublished Technical Document,* Colorado State University, September 1995.

[5] Mark Crovella and Robert Carter, "*Dynamic Sever Selection in the Internet,*" *Department of Computer Science Unpublished Technical Document,* Boston University, August 1995.

[6] Paul R. Cohen, *Empirical Methods for Artificial Intelligence,* Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1995.

[7] Wayne A. Fuller, *Introduction to Statistical Time Series*, John Wiley & Sons, Incorporated, 1976, pp. 18 - 92.

[8] R.J. Bhansali, "Autoregressive Estimation of the Prediction Mean Squared Error and an R^2 Measure: An Application", *The IMA Volumnes in Mathematics and its Applications,* Volume 45, Springer-Verlag, 1991, pp. 7 - 24.

[9] Gene V. Glass, Victor L. Wilson and John M. Gottman, *Design and Analysis of Time-Series Experiments*, Colorado Associated University Press, Boulder, Colorado, 1975, pp. 82 - 119.

[10] Eivind Damsleth, "Forecasting the Production of Pigs -- The Birds and the Bees Reconfirmed," *Time Series Analysis: Theory and Practice,* O.D. Anderson, Editor, Elsevier Science Publishers B.V., 1984, pp. 159-174.

[11] Paul D. Wagner and Thomas A. Wagner, "The WWW in the Classroom: Access without Adult Material," *Under Review at The Technology Teacher,* March, 1996.