Title
Internet Data Analysis for the Undergraduate Statistics Curriculum

Authors
Juana Sanchez
Yan He

Publication Date
2011-10-25

Internet Data Analysis for the Undergraduate Statitics Curriculum

Juana Sanchez*
UCLA Statistics

Yan He
UCLA Statistics

December 14, 2003

KEY WORDS: Inverse Gaussian; maximum likelihood; web server log data; internet survey data; internet traffic

Abstract

Statistics textbooks for undergraduates have not caught up with the enormous amount of analysis of Internet data that is taking place these days. Case studies that use Web server log data, Internet survey data or Internet network traffic data are rare in undergraduate Statistics education. This paper summarizes the results of research in three areas of Internet data analysis: users' web browing behavior, user demographics, and network performance. We present some of the main questions analyzed in the literature, some unsolved problems, and some typical data analysis methods used. We illustrate the questions and the methods with large data sets. The data sets were obtained from the publicly available pool of data. Those data sets had to be processed and transformed to make them available for classroom exercises. The processed data sets as well as more material for classes, are available at a web site with address that can be obtained from the main author.

1. Introduction

Statistics textbooks for undergraduates have not caught up with the enormous amount of analysis of Internet data that is taking place these days. Case studies that use Web server log data, Internet survey data or Internet network traffic data are rare in undergraduate Statistics education. We had to conduct a large amount of research and computer work to be able to assess the key areas of current

*Please, address all correspondence and comments to this author

Internet data analysis that are more likely to have an impact on business and the population of users in general. The three data sets and the story around each of them that we present below summarize that research.

One common thread in the three stories told below is that there is an increasing demand for Internet data analysis. Computer Scientists are being called more and more frequently to provide computer log data that can be used to understand users' web browsing behavior, to make web pages more responsive to users' needs. In addition to that, the number of Internet user surveys is growing at an amazing speed for the same reasons. And communication networks are providing data that can help understand how the network itself responds to users, to make the quality of the network better for users.

For the teaching of Statistics, there are some things in common in the three stories. Because the data and graphs that appear in Internet data analysis behave sometimes differently from what we teach students in the descriptive data analysis module, we can use these stories to reinforce what students already know, by contrast. In addition to that, the data sets are huge, much bigger than the ones we usually use in our teaching, presenting the student with the mystery of dealing with such monsters. Finally, but not the least, there are no definite answers yet, so the students are really being exposed to the ongoing search for new paradigms in the engineering, computer science and statistics community.

2. Web browsing behavior at the user level

Once a user enters a web site how many pages or links within the site does that user visit? The answer to this question may suggest actions to improve the site. If similar distributions for the number of pages visited per user are observed at different web sites, then maybe some laws can be established for all sites. Research efforts in this area are directed at finding these laws. Examples of these efforts are Hansen and Sen (2003), Cadez et al. (2003) and Huberman et al. (1998), each analyzing a different data set.

Some of the analysis done in the literature to answer that question can be illustrated with data published in the UCI KDD Archive (Heckerman, 2003). We processed these data to obtain observations on the number of different pages visited by users who entered the msnbc.com page on September 28, 1999 and other information. A random sample of this data set was used by Cadez et al.(2003) to do cluster analysis and visualization of the patterns (order) of visits followed by users, i.e., to see the frequency of whole sequences. This is a very important question, too. But we don't look into it in this paper.

The original data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hours period. Since there are 989818 users, there

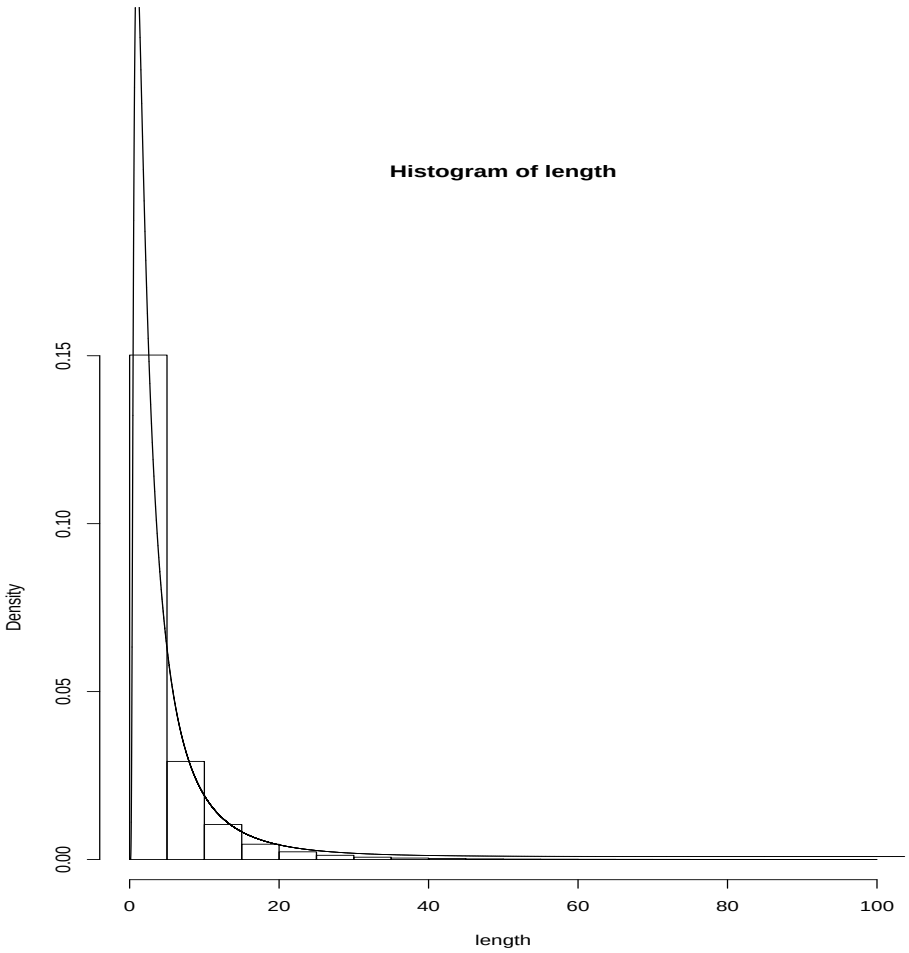

Figure 1: Histogram of the length of visits. Inverse gaussian of same mean and variance overimposed

are 989818 sequences. Each event in a sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail–that is, not at the level of URL, but rather, they are recorded at the levels of page category (as determined by a site administrator). The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs (bulletin board service)", "travel", "msn-news", and "msn-sports". As an example, we write below the sequence for the first three users in the data set (one line per user):

User 1: frontpage, frontpage
User 2: news
User 3: tech,news,news,local,news,news,news,tech,tech

We processed the original data set to obtain the variable "length", which represents the actual total number of links visited by each user. For example, user

one has length=2, user two has length=1, and user three has length=9. The average number of pages visited is 4.747, the median is 2 pages, the minimum is 1 and the maximum is 14800 pages. The histogram, in Figure 1, is very skewed.

Notice that the histogram contains only values of length less than or equal to one hundred, excluding those users that visited more than 100 pages. The longest visits are probably crawlers or maybe different people logged in the same IP address. One of the problems with web server log data is precisely what to do with these crowlers. Should they be included, should they not? Although we did not include them all in the graphs, all the numbers were used for the computations of the statistics. An important fact to observe is that most users visit few pages, but the tails are very long, indicating that some users visit a lot of pages.

What model should we use for this behavior? Huberman et al. (1998) and other authors, recommended an inverse gaussian distribution for the variable length (L). This distribution has two parameters and is described by the formula

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} \exp\left[\frac{-\lambda(L-\mu)^2}{2\mu^2 L}\right]$$

The mean $E(L) = \mu$ and variance $Var(L) = \mu^3/\lambda$, where λ is a scale parameter. This distribution "has a very long tail, which extends much farther than that of a normal distribution with comparable mean and variance. This implies a finite probability for events that would be unlikely if described by a normal distribution. Consequently, large deviations from the average number of user-clicks computed at a site will be observed" (Huberman et al., 1998). Another property is that "because of the asymmetry of the distribution function, the typical behavior of users will not be the same as their average behavior. Thus, because the mode is lower than the mean, care must be exercised with available data on the average number of clicks, as this average overestimates the typical depth being surfed."

It can be shown that the cumulative distribution function of the inverse Gaussian distribution is

$$F(L,\mu,\lambda) = \Phi\left[\sqrt{\frac{\lambda}{L}}\left(\frac{L}{\mu}-1\right)\right] + e^{2\lambda/\mu}\Phi\left[\sqrt{\frac{\lambda}{L}}\left(\frac{L}{\mu}+1\right)\right]$$

where $\Phi[\ ]$ is the standard normal distribution function.

Is the inverse Gaussian really a good model for the data we have? It is instructive to follow the guidelines given in the references mentioned above to answer this question.

Theoretically, by maximizing the likelihood function, the equations for the maximum likelihood estimators (MLE) of λ and μ in the inverse Gaussian distribution given above can be found to be

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} L_i$$

and

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} \frac{(L_i/\hat{\mu} - 1)^2}{L_i}}$$

For the msnbc.com data, we find:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} L_i = \frac{4698794}{989818} = 4.747129$$

$$\hat{\lambda} = \frac{989818}{321169.6} = 3.081917$$

The inverse gaussian with these estimates is fitted to the histogram in Figure 1. Visually, it is a good fit, but we don't show the tails, so we can not conclude from this plot that the fit is good over the whole distribution.

Commands in R to do these computations and the ones that follow can be found in Appendix A.

To see how good is this model, Huberman et al.(1998) and Sen and Hansen(2003) compared the cumulative distribution function implied by the model to the empirical cumulative distribution function derived from the data. Then they use a p-p plot against the fitted distribution. We do the same with the length variable; the plots can be seen in Figure 2. The p-p plot reveals a misfit of the inverse gaussian model to our data. Hansen and Sen (2003) got similar results with the bell-labs.com data set they used.
Another way of investigating whether the inverse gaussian is a good model, is based on the following fact: If you take logs on both sides of the inverse gaussian formula you obtain

$$\log P(L) = -\frac{3}{2}\log L - \frac{\lambda(L-\mu)^2}{2\mu^2 L} + \log\left(\sqrt{\frac{\lambda}{2\pi}}\right)$$

Thus a plot of $\log(L)$ vs $\log(\textit{frequency})$ should show a straigt line whose slope approximates 3/2 for small values of L and large values of the variance. A plot of

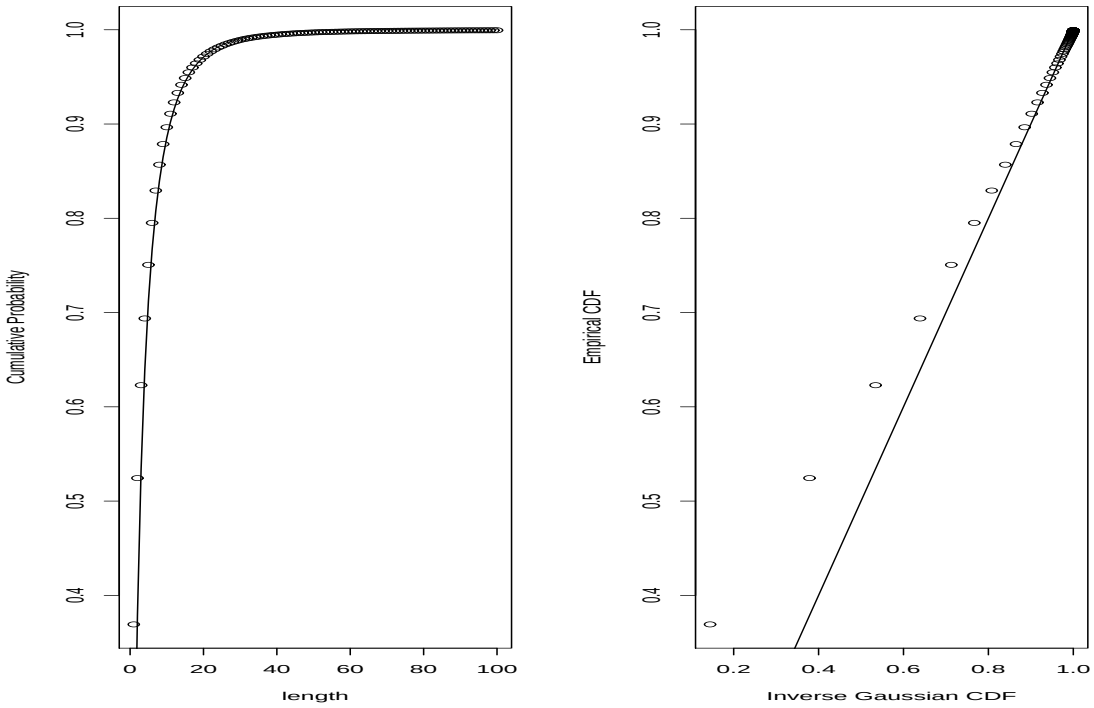

Figure 2: Left:Empirical (o o o) and inverse gaussian (—-) cumulative distributions. Right: p-p plot

the frequency distribution of surfing clicks on the log-log scale can be seen in Figure 3. Acording to this plot, the theoretical result holds. The regression line for the whole range of the data has a slope of -1.9
The log-log plot also helps notice that, up to a constant given by the third term, the probability of finding a group surfing at a given level scales inversely in proportion to its length, $P(L) \propto L^{-\frac{3}{2}}$. This is a characteristic that appears in a lot of Internet data sets. We don't pursue it further here, but it is at the heart of the debate about the nature of the data and the best possible model.

Based on the results above, would we recommend the inverse Gaussian model for the length of visits (or number of links that a user visits) in the msnbc.com data set? This is one of the questions still unanswered and in need of more research.

We shall not dwell further into Web browsing behavior research. But before we conclude this section, we would like to point out that the above is just the tip of the iceberg. Once the distribution of "length" is settled, the next step for researchers is to model the sequence of requests by users. Huberman et al (1998) model them using a simple first order Markov model. Sen and Hansen (2003) try a first and second-order Markov model, a finite mixture of first-order models, and a Bayesian approach. Cadez et al. (2003) investigate simple markov models for different clusters of users. The objective of these modeling attempts is to determine the best model to predict a user's next page request. Pages with higher

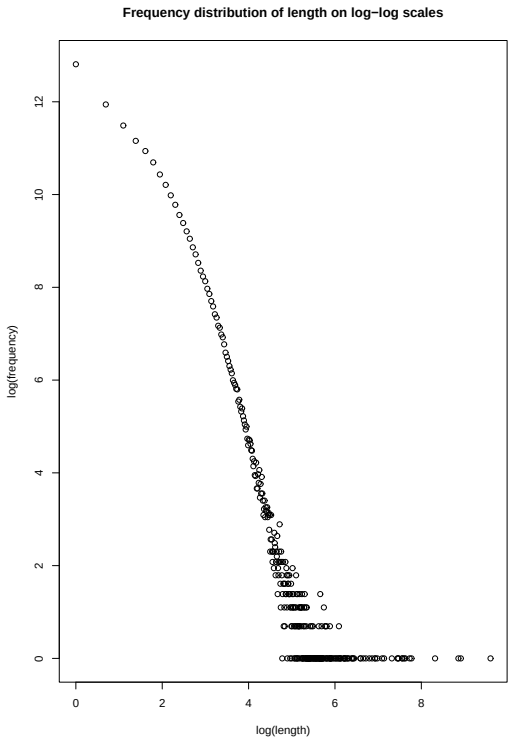

Figure 3: Plot of frequency distribution of length in log-log scale

probability of being requested can then be made more accessible.

Interested readers can experiment in class with many of these questions, either using the raw data, or three different processed data sets that we extracted from this raw data set, and that are available at a web site that we will be glad to provide upon request. Perhaps the reader can obtain Web server log data from the school where this material will be taught. In the latter case, be aware that raw log server data with URLs and detailed computer information needs to be converted to something like the raw data of Heckerman (2003) using Perl or similar programs. After that, you can process it further to use it for data analysis.

3. Demographics of Internet Users

Who is using the Internet? What is the effect of the Internet on users? These are very important questions that are receiving a lot of attention in the research community (Wellman et al.(2002)). At the end of 2001 the number of Internet users in the world was more than 500 million (up from 16 million in 1996). The Internet has quickly become part of our lives and numerous research efforts have been dedicated to trying to understand who is using it and how it is being used. To illustrate some of questions being asked, we use data from a survey conducted by the Graphics and Visualization Unit at Georgia Tech from October 10 to

November 16, 1997. The full details of the survey are available at
http://www.gvu.gatech.edu/user_surveys/survey-1997-10/graphs/#general
The particular subset of the survey provided here is the "general demographics" of internet users. The number of users participating in the survey is 10108.

Thi survey did not use random sampling. How to make these Internet surveys more random is an open question still unresolved. The methodology aspects of the survey are discussed at
http://www.gvu.gatech.edu/user_surveys/survey-1997-10/#methodology

A selective summary of the survey, results and trend analysis appear in the web page provided above, thus, it would be repetitive to reproduce all of them here. For teaching purposes, one can do other analyses. For example, one can look at the relation between age and number of years on the internet, or between marital status and years on the Internet.

Key
frequency
row percentage
column percentage
cell percentage

	Years on Internet					
ageclass	< 1/2	1/2-1	1-3	4-6	> 7	Total
5-10	52	29	10	1	3	95
	54.74	30.53	10.53	1.05	3.16	100.00
	2.97	1.62	0.26	0.06	0.44	0.97
	0.53	0.33	0.10	0.01	0.03	0.97
10-20	194	146	406	107	15	868
	22.35	16.82	46.77	12.33	1.73	100.00
	11.09	8.18	10.76	5.92	2.19	8.86
	1.98	1.49	4.14	1.09	0.15	8.86
20-30	334	352	1,227	778	218	2,909
	11.48	12.10	42.18	26.74	7.49	100.00
	19.10	19.72	32.51	43.05	31.87	29.69
	3.41	3.59	12.52	7.94	2.22	29.69
30-40	375	393	833	435	223	2,259
	16.60	17.40	36.87	19.26	9.87	100.00
	21.44	22.02	22.07	24.07	32.60	23.05
	3.83	4.01	8.50	4.44	2.28	23.05
40-50	385	408	753	296	154	1,996
	19.29	20.44	37.73	14.83	7.72	100.00
	22.01	22.86	19.95	16.38	22.51	20.37
	3.93	4.16	7.68	3.02	1.57	20.37
50-60	251	261	392	156	65	1,125
	22.31	23.20	34.84	13.87	5.78	100.00
	14.35	14.62	10.39	8.63	9.50	11.48
	2.56	2.66	4.00	1.59	0.66	11.48
60-70	112	134	125	24	6	401
	27.93	33.42	31.17	5.99	1.50	100.00
	6.40	7.51	3.31	1.33	0.88	4.09
	1.14	1.37	1.28	0.24	0.06	4.09
>70	46	62	28	10	0	146
	31.51	42.47	19.18	6.85	0.00	100.00
	2.63	3.47	0.74	0.55	0.00	1.49
	0.47	0.63	0.29	0.10	0.00	1.49
Total	1,749	1,785	3,774	1,807	684	9,799
	17.85	18.22	38.51	18.44	6.98	100.00
	100.00	100.00	100.00	100.00	100.00	100.00
	17.85	18.22	38.51	18.44	6.98	100.00

Pearson chi2(28) = 771.6993 Pr = 0.000

```
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+

   Marital |                  Years on Internet
    Status |   < 1/2    1/2-1   1-3     4-6    > 7  |  Total
-----------+-------------------------------------------+--------
  divorced |       209    216    246      87     40 |    798
           |     26.19  27.07  30.83   10.90   5.01 | 100.00
           |     11.51  11.92   6.46    4.80   5.92 |   8.04
           |      2.11   2.18   2.48    0.88   0.40 |   8.04
-----------+-------------------------------------------+-------
live other |       124    130    389     214     83 |    940
           |     13.19  13.83  41.38   22.77   8.83 | 100.00
           |      6.83   7.17  10.22   11.81  12.28 |   9.47
           |      1.25   1.31   3.92    2.16   0.84 |   9.47
-----------+-------------------------------------------+---------
   married |       747    746  1,558     703    309 |  4,063
           |     18.39  18.36  38.35   17.30   7.61 | 100.00
           |     41.13  41.17  40.92   38.80  45.71 |  40.95
           |      7.53   7.52  15.70    7.08   3.11 |  40.95
-----------+-------------------------------------------+---------
 separated |        41     35     40      13     10 |    139
           |     29.50  25.18  28.78    9.35   7.19 | 100.00
           |      2.26   1.93   1.05    0.72   1.48 |   1.40
           |      0.41   0.35   0.40    0.13   0.10 |   1.40
-----------+-------------------------------------------+---------
    single |       640    634  1,543     784    231 |  3,832
           |     16.70  16.54  40.27   20.46   6.03 | 100.00
           |     35.24  34.99  40.53   43.27  34.17 |  38.62
           |      6.45   6.39  15.55    7.90   2.33 |  38.62
-----------+-------------------------------------------+---------
   widowed |        55     51     31      11      3 |    151
           |     36.42  33.77  20.53    7.28   1.99 | 100.00
           |      3.03   2.81   0.81    0.61   0.44 |   1.52
           |      0.55   0.51   0.31    0.11   0.03 |   1.52
-----------+-------------------------------------------+---------
     Total |     1,816  1,812  3,807   1,812    676 |  9,923
           |     18.30  18.26  38.37   18.26   6.81 | 100.00
           |    100.00 100.00 100.00  100.00 100.00 | 100.00
           |     18.30  18.26  38.37   18.26   6.81 | 100.00

         Pearson chi2(20) = 273.9131   Pr = 0.000
```

According to these tables, the majority of Internet users in 1997 were between 20 and 50 years old, and they had been on the Internet less than 7 years for the most part. The majority of users were married or single. Separated, widowed and divorced people were not big users. For most marital status groups, three years or less on the Internet were the most common time.

Many more conclusions can be derived from the results. Data like the one in this survey can really lead to a lot of discussion as well as to a lot of practice with conditional, marginal and joint distributions. We have this data set saved in Stata and in text format. Students can investigate all the relations that interest them in this data set. Alternatively, they could try to obtain more recent surveys, for example the one undertaken by the UCLA Center for Communication Policy at `http://www.stat.ucla.edu/pages/internet-report.asp`. From the same site one can access information on similar surveys around the world.

4. Internet traffic

Can we predict how the Internet network will perform at any time? An answer to this question would have the practical implication of helping optimize service provision to keep all customers and network administrators happy. However the Internet network is so heterogenous and unregulated, and the lack of cooperation

among individual servers is so prevalent that monitoring the network is a very challenging task. To monitor, good models are needed. But before the modeling stage is reached, a good understanding of the data is necessary. There seems to be now a consensus among researchers about the nature of the data but not about the best way to model that data. In this section we only unveil the nature of the data set we analyze and summarize some of the issues. The language is very specialized, and for this reason, we introduce some get-acquainted kind of information below (Gautam, 2003).

Messages that flow from a source to a destination through a network are also known as traffic. This traffic and the network conditions are extremely random in nature.

There are three types of telecommunication networks –telephony (telephone network for voice calls, fax, and also dial up connections), cable-TV networks (cable, web-TV, etc.) and high speed networks such as the Internet. We are concerned with high-speed networks.

Traffic flowing through the networks can be classified into several types. Two of the most common traffic types are ethernet packets/frames and ATM cells. The length or size of an Ethernet packet ranges anywhere from 60 bytes to 1500 bytes and generally follows a bimodal distribution. The length of ATM cells is fixed at 53 bytes. Therefore the network traffic comprises of millions and billions of these little packets or cells. We are concerned with packets traffic.

The packets arise because when a message needs to be sent from a source to a destination, it is broken down into small packets and transported that way from the source to the destination.

The protocols responsible for this transport of packets over networks are user datagram protocol (UDP) and transmission control protocol (TCP). With UDP, the destination does not acknowledge the receipt of packets to the source. TCP is an acknowledgement (ACK) based protocol. In this paper, we are concerned with packets transported by TCP protocol.

There are several network performance measures that contribute to the Quality of Service of a network. Among others, we have: (a) loss probability, or the probability of delivering a message with some data loss; (b) delay or latency, the time lag between the source sending a message and the destination receiving it; (c) delay-jitter or measure of the variation of the delay; (d) bandwidth or rate at which messages are processed. These measures can be used for optimal design and admission control of the networks. Design deals with buffer sizes, link capacities, network parameters, traffic shaping parameters, and other. Admission control involves rejecting or accepting an incoming request for connection. These variables are very hard to measure; some researchers have come up with inference methods to estimate them. These methods are very complicated and difficult to understand

by beginning statistics students. Because the nature of the data is so different from that of data we are more used to in our undergraduate classes, students should be able to understand that scientists are having a very hard time coming up with good models.

There are two main areas of interest in the research on Internet traffic data. One area is concerned with understanding how traffic data perform within a single route, i.e., in the connection between a pair of nodes in the network, which allows the use of stochastic process modeling (Willinger et al., 1995; Willinger and Paxson, 1998; Paxson and Floyd, 1995). It has been found that the models used for telephone networks are not good for Internet traffic (Willinger et al., 1995). The other area is more focused in modeling the simultaneous activity across all the nodes in the network, i.e., traffic measurements at different nodes based on carefully done sampling at different nodes. This latter approach is known as "network tomography" (Castro et al., 2003) and it is very recent. The goal of both approaches is to predict measures of performance of the networks. The single route approach supporters believe that by careful selection of processes to model the traffic, more theoretical analysis can be done for multiple routes as well. Many attempts are being made by this group at maintaining the old queueing models that work so well with telephone networks, with some adaptations to the different behavior of Internet data (Guerin et al., 2003).

It should be pointed out that because of the privacy of network data researchers are having a very hard time obtaining the data they need. But for the data sets that have been made publicly available, there are standardized ways to reduce data on traffic to statistics-friendly data. For example, Jeff Mogul used the steps described at this site `http://ita.ee.lbl.gov/html/contrib/sanitize-readme.txt` to make the TCP data set we use in this section user friendly.

The data discussed here can be obtained from the Internet Traffic Archive at : `http://ita.ee.lbl.gov/html/contrib/DEC-PKT.html` and is labeled dec-pkt-1.tcp. This data set summarizes traces of one hour's worth of TCP traffic between Digital Equipment Corporation and the rest of the world on March 8th, 1995. Paxson and Floyd (1995) also analyzed this data set. It describes 2,153,462 million packets and contains the following 6 variables.

- *timestamp* of packet arrivals. For the first packet in the trace, this is the raw time. But I removed the integer part.
- *source*: Source host
- *destination*: Destination host
- *sourceport*: source TCP port
- *destport*: Destination TCP port

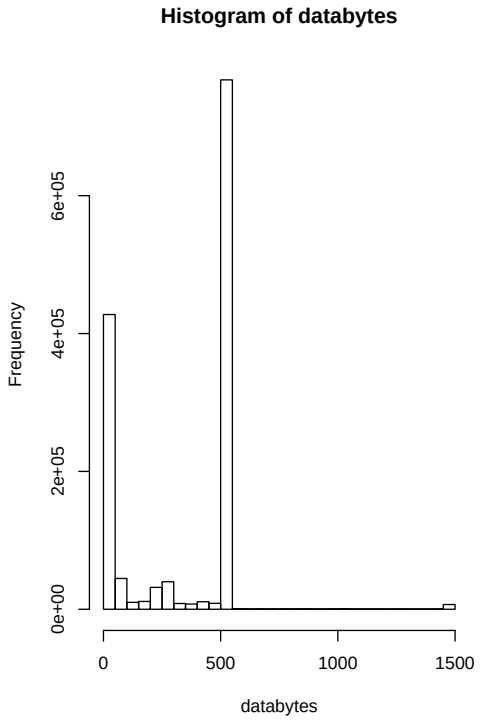

Figure 4: Histogram of the number of data bytes in a packet

- *databytes*: number of data bytes in the packet, or 0 if none (this can happen for packets that only ack data sent by the other side. We remove the 0 packages in all the analysis done below.

The packet size and number of packets per minute are very important variables, and we will analyze them here.

Our analysis in this paper will be limited to some of the preliminary descriptive data analysis usually done. Since the data set we use is only for an hour there are many things that other studies look at that we can not investigate.

A histogram of the number of bytes per packet is shown in Figure 4. We can see that it is bimodal as expected. The minimum value is 1 and the maximum is 1460. The most frequent package size is around 512. Packages of that size arrive uniformly throughout the day, they are not concentrated in any particular hour. The correlation coefficient between the timestamp and the databytes variables is 0.01128, illustrating the lack of relation.

Another question of interest is: do the times between arrivals of two consecutive packets to Digital Equipment follow an exponential distribution? This question is important because if that is the case, there is still the possibility that some variant of Poisson models could be used for modeling the number of packets arriving per

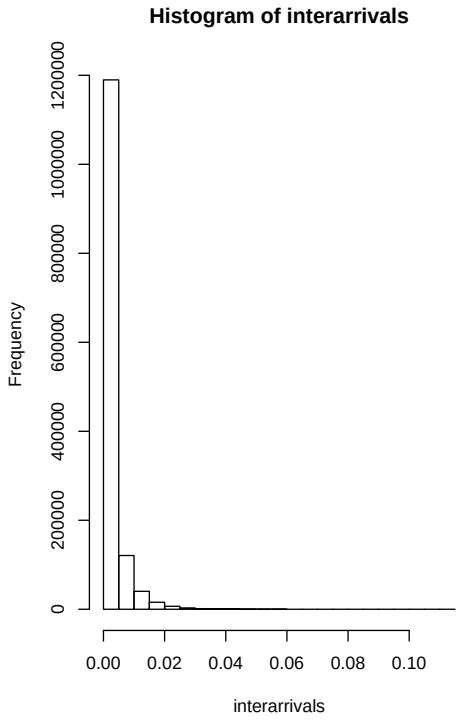

Figure 5: Distribution of interarrival times of packets

unit of time. A histogram of the timestamp between two arrivals reveals that inter-arrival times are exponential. We can see it in Figure 5, where we have an exponential kind of distribution with a short tail. Summary statistics show that on average, packages arrive within 0.000976 seconds of each other, according to the median. The smallest interarrival time is 0, for packages that arrive simultaneously, and the maximum amount is 0.11. The consensus in the literature is that this variable is exponential.

Unlike in telephone networks, where the number of arrivals per unit of time is Poisson, in Internet networks exponential inter-arrival times do not translate into Poisson behavior for the number of packets per unit of time, or for the number of data bytes per unit of time. This has been attibuted by computer scientists and engineers to the burstiness of those variables over time, regardless of the time scales at which we measure them. Many papers have been written showing how this periodic outbursts of activity over time are present in almost all the publicly available data sets. The challenge has been and still is to determine why. Most papers have resorted to asymptotic explanations, and to this day there is no consensus on this matter.

The data set we use in this section is too short to investigate the burstiness of the number of packets per unit of time thoroughly. But it illustrates what this behavior looks like in a plot. We created a data set called "pacpersecond"

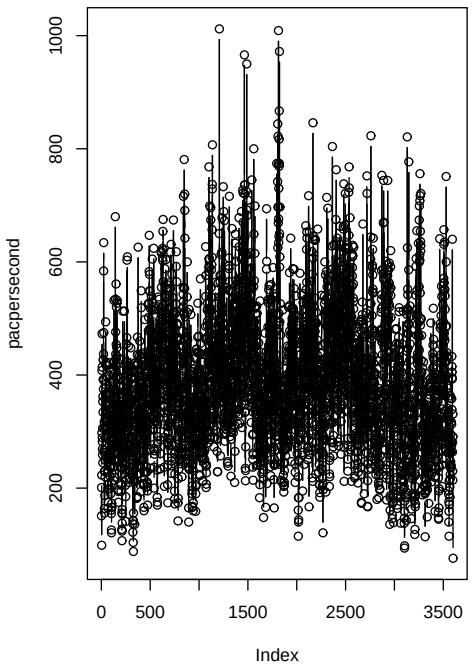

Figure 6: Trace of the number of packets per second

representing the number of packets arriving every second, and plotted it against an index variable representing seconds. Figure 6 shows that the number of packets per second displays burstiness. Changing the scale does not make the burstiness disappear; we created another variable called "pacperminute" that represents the number of packets arriving per minute. Figure 7 shows the trace of this variable against an index variable representing minutes. Again, the burstiness is present. The fact that the burstiness does not disappear whether we measure the number of packets per minute, per second, or millisecond, is not a characteristic of Poisson counts. And this is what is making the modeling so difficult.

The above descriptions of the data we are using in this paper indicate that this data set follows the same behavior as most of the data sets analyzed in the literature on network traffic. With these pieces of information, researchers are trying to determine what kind of model of network traffic will capture these characteristics. There are many other interesting types of analysis that one can do with longer data sets, but they all lead to the same conclusions. These other data sets can be found in the address given earlier in this section.

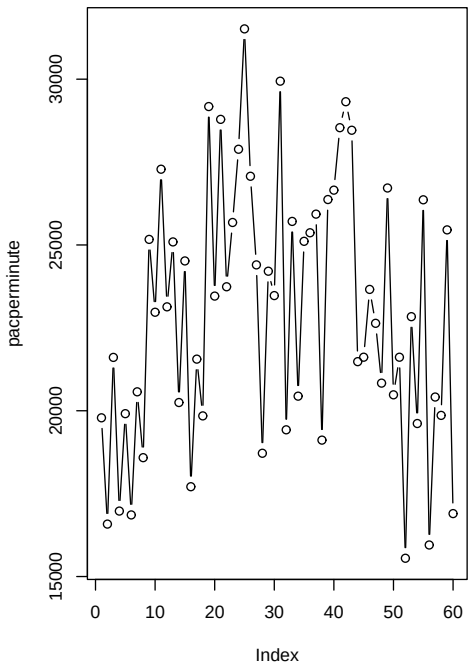

Figure 7: Trace of the number of packets per minute

5. Conclusions

As the reader may be able to ascertain from the above stories, there is plenty of interesting projects for our students in the analysis of Internet data. The data sets that most researchers are using to test their models are publicly available. These data sets, after appropriate processing, can be assigned to students to investigate questions similar to those summarized above or more advanced ones.

What is most relevant is that the analysis of Internet data is at its early stages and therefore there are many unsolved questions and no established paradigm. By involving students in those questions, we are making them active participants in current debates, without leaving the classroom. Interested readers can contact the authors to access exercises and data sets that we have prepared for use in the classroom.

Appendix A

We present below the R program used to obtain the results and graphs of Section 2 above.

```
# Read data, do histogram and superimpose the inverse Gaussian
```

```
length=read.table(‘‘msnbclength.txt’’,header=T)
length=length$length



length=length[length<=100]

library(SuppDists)

postscript(‘‘histogram.eps’’,horizontal=FALSE)

par(mar=par(‘‘mar’’)+c(0,0,15,0))

hist(length,prob=T)

pts=seq(from=par(‘‘usr’’)[1],to=par(‘‘usr’’)[2],len=length(length))

lines(pts,dinvGauss(pts,4.747129,3.081917),xpd=T)

<
dev.off()




# Do the plot of log length against log frequency and
# find regression estimates

freqtable = read.table(‘‘frequencytable’’,header=TRUE)
length=freqtable$length
frequency=freqtable$frequency
plot(log(length),log(frequency))
lm(log(frequency)~log(length))



library(SuppDists)
length=read.table(‘‘msnbclength.txt’’,header=TRUE)
length=length$length

n=length(length)
lengthsum=sum(length)
lengthtable=table(length)
lengthfrequency=table(length)/n
cumlength=cumsum(lengthfrequency)
```

```
mu=4.747129
lambda=3.081917

l=seq(1,100)
prob=pinvGauss(l,mu,lambda)


postscript(‘‘cdf.eps’’,horizontal=FALSE)
par(mfrow=c(1,2))
plot(l,cumlength[1:100],type=’’p’’,xlab=’’length’’,ylab=’’Cumulative
probability’’)
lines(l,prob)
plot(prob,cumlength[1:100],xlab=’’Inverse
Gaussian CDF’’,ylab=’’Empirical CDF’’)
lines(l/100,l/100)
dev.off()
```

Appendix B

We present here the R program used to obtain the results and graphs of Section 4.

```
dec1 = read.table("dec-pkt-1.tcp",header=TRUE)

timestamp = dec1$timestamp[dec1$databytes>0]
databytes = dec1$databytes[dec1$databytes>0]
source=dec1$source[dec1$databytes>0]
destination=dec1$destination[dec1$databytes>0]
sourceport=dec1$sourceport[dec1$databytes>0]
destport=dec1$destport[dec1$databytes>0]

hist(databytes)
summary(databytes)

plot(timestamp,databytes)

arrivals = matrix(timestamp)
interarrivals = matrix(rep(0,1378970),ncol=1)
for(i in 1:1378970) {
  interarrivals[i]= arrivals[i+1]-arrivals[i]
}
hist(interarrivals)
summary(interarrivals)
```

```
y =rexp(1378970,383)
hist(y)

pacpersecond = matrix(c(rep(0,3600)))

for(i in 1:3600) {
pacpersecond[i]=length(timestamp[(floor(timestamp))==i])
}
plot(pacpersecond,type=''b'')

pacperminute = matrix(c(rep(0,60)))
pacperminute[1]= sum(pacpersecond[1:60])
for(i in 2:60) {
j=(i-1)*60+1
k=i*60
pacperminute[i]=sum(pacpersecond[j:k])
}
plot(pacperminute,type=''b'')
```

Acknowledgements

The research in this paper was funded by the Office of Instructional Development, under Instructional Improvement Grant OID IIP 03-20 to the main author, to whom all correspondence should be addressed. The second author is a graduate student in Statistics who collaborated in the material of Sections 2 and 3. The contents of the paper are a small part of a larger project intended to create activities and data sets for undergraduate students in Computer Science taking their first course in Statistics. We would like to thank Walter Rosenkrantz, Jeff Mogul, Zhiyi Chi, and Mark Hansesn for helpful suggestions or comments.

References

[1] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S. (2003). Model-based clustering and visualization of navigation patterns on a Web site. Accepted for publication on Journal of Data Mining and Knowledge Discovery, 7(4).

[2] Castro,R., Coates M., Liang, G., Nowak R., Yu, B. (2003). Network Tomography: recent developments.

[3] Digital Equipment Corporation. The traces were made by Jeff Mogul (mogul@pa.dec.com) of Digital's Western Research Lab (WRL). The trace correspond to DEC-WRL-1

[4] Gautam, N. (2003) Stochastic Models in Telecommunications for Optimal Design, Control and Performance Evaluation. *Handbook of Statistics, Vol. 21* D.N. Shanbhag and C.R. Rao, eds. Elsevier Science B.V., 2003.

[5] Hansen, M.H. and Sen, R.(2003). Predicting Web User's next access based on log data. Journal of Computational and Graphical Statistics, Vol 12, No. 1, March, p. 143-155.

[6] Heckerman, D. The UCI KDD Archive (`http://kdd.ics.uci.edu`) Irvine, CA: University of California, Department of Information and Computer Science. The URL for the data used in this paper is `http://kdd.ics.uci.edu/databases/msnbc/msnbc.data.html`

[7] Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong Regularities in World Wide Web Surfing. Science, Vol. 280, 3 April.

[8] Paxson,V., Floyd S.(1995). Wide-Area Traffic: The Failure of Poisson Modeling,IEEE/ACM Transactions on Networking, 3(3), pp. 226-244, June 1995.

[9] Wellman, B. and Haythornthwaite edts.(2002). The Internet in Everyday Life. Blackwell Publishing.

[10] Willinger, W and Paxson, V. (1998). Where Mathematics meets the Internet. Notices of the AMS, September 1998. p 961-970.

[11] Willinger, W., Taqqu, M.S., Leland, W.E. and Wilson, D.V. (1995). Self-similarity in High-Speed PAcket Traffic: Analysis and Modeling of Ethernet Traffic Measurements.Statistical Science, Vol 10, No. 1, p 67-85.

Juana Sanchez
UCLA Department of Statistics
8130 Math Sciences Building
Box 951554
Los Angeles, CA 90095-1554
jsanchez@stat.ucla.edu

Yan He
UCLA Department of Statistics
8130 Math Sciences Building
Box 951554
Los Angeles, CA 90095-1554
yanhe@stat.ucla.edu