Current Internet Technology and Statistics — Blessing or Curse?

Jürgen Symanzik
George Mason University
Center for Computational Statistics 4A7
Fairfax, Virginia 22030, USA
e–mail: symanzik@galaxy.gmu.edu
WWW: http://www.galaxy.gmu.edu/~symanzik

Abstract

In the first part of this paper, we provide a general overview on the history of the Internet and some of its main features. In the second part, we introduce existing Internet technology that effects todays education and research in Statistics. Examples are electronic journals, statistical software packages, teaching software accessible through the WWW, and on–line access to data, organizations, and individuals — to mention only a few. We finish with a critical discussion on advantages, disadvantages, and even dangers of the current Internet technology.

1 Introduction

During the last few years, Internet technology has become an important factor in our professional and private lifes. There are probably not many among us statisticians that can claim that they are not affected at all by current Internet technology. But — what do we understand by the term *Internet Technology*? As we will see later on in this paper, the Internet is basically a chaotic environment that supports a variety of features, formats, does not belong to any particular person, organization, or country, but, surprisingly, works anyway.

In Section 2 of this paper, we will provide a short summary on the history and main features of the Internet. In Section 3, we will see how closely Statistics and the Internet are related today. In Section 4, we will discuss the advantages, disadvantages, and dangers of the current Internet technology (and possible future developments) with respect to our professional and private lifes.

This paper cites a large number of references on the World–Wide Web. These references are mostly used as examples to highlight particular topics. Within the scope of this paper it is not possible to provide additional links, i. e., pointers, to persons, organizations, or companies. However, with an Internet search engine such as *Yahoo!* (URL:Yahoo) or *excite* (URL:Excite), it should be easy to acquire additional information based on the keywords used throughout this paper.

2 History and Features of the Internet

Not surprisingly, the history of the Internet can be accessed most easily through the Internet itself. When searching for "Internet and History", entire presentations (URL:Neosoft) and comprehensive historical information can be found on somewhat independent sites such as *Internet Valley* (URL:InternetValley1, URL:InternetValley2) as well as on commercial sites that belong to *MCI* (URL:MCI1, URL:MCI2), a telephone company that is also closely related to new Internet developments.

Commissioned by the *Department of Defense* (DoD) (URL:DefenseLINK), the *Defense Advanced Research Projects Agency* (DARPA) (URL:DARPA) began to focus on computer networking and communications technology during the 1960ies. Universities, funded through this initiative, laid the foundations for what later would become the *Advanced Research Projects Agency NETwork* (ARPANET). The first communication between two different sites took place in 1969 when researchers from the *University of California, Los Angeles* (UCLA) (URL:UCLA) tried to log onto the *Stanford Research Institute* (SRI) (URL:SRI) computer. Even though only three characters, **L O G**, had been typed before a system crash occurred, this communication marks the birth of the Internet and also was the starting point of an electronic revolution.

By the end of 1969, two more nodes were developed, one at the *University of Utah* (URL:UUtah) and

the other at the *University of California, Santa Barbara* (UCSB) (URL:UCSB). In 1973, global networking became reality through the addition of the first international connections to the ARPANET: the *University College of London*, England (UCL) (URL:UCL) and the *Royal Radar Establishment*, Norway. While growing steadily throughout the 1970ies and 1980ies, an explosion started in the late 1980ies with the number of hosts (nodes) almost doubled every year since then. In January 1998, the estimated number of hosts on the Internet was about 30 million (URL:InternetValley1).

The Internet provides a variety of different services to its users. Best known are *electronic mail* (e–mail), introduced in 1971, which allows a user to send an electronic message to one or more other users, *telnet* (since 1972) that allows to log onto another computer through a network of multiple intermediate computers, the *file transfer protocol* (ftp), which, since 1973, allows to exchange files between different computers that even support different internal formats and numeric representations, and since 1979 *news groups / discussion groups*, that allow an arbitrary number of users to participate in a discussion on a particular topic.

Released in 1991 by Lindner and McCahill from the *University of Minnesota*, *gopher* (URL:Gopher) became a widely used, text–based, menu–driven interface to access Internet resources. Also in 1991, perhaps the most important Internet development to date was the release of the *World–Wide Web* (WWW, often simply called Web) at the *European Laboratory for Particle Physics* (typically abbreviated as CERN which is the French acronym for *Centre Européenne pour la Recherche Nucléaire*) near Geneva, Switzerland (URL:CERN). Initially, the WWW was a non–graphic (!) distributed hypermedia system based on the *hypertext transfer protocol* (http). A standard naming convention, the *Uniform Resource Locator* (URL), allows to locate individual pieces of information that reside anywhere on the network. URLs typically start with the document type they are pointing to, such as http, ftp, gopher, mail, etc.

It took two more years after its introduction before the WWW, as we know it today, became reality: In 1993, the *National Center for Supercomputing Applications* at the *Univerity of Illinois at Urbana–Champaign* introduced its *NCSA Mosaic* (URL:NCSAMosaic) for the XWindowSystem, Microsoft Windows, and the Macintosh — the first Web browser as we know them today.

Articles such as Baker (1994) were first read with some disbelief but could easily be verified by downloading the freely available *NCSA Mosaic* executables from NCSA's ftp server. Then, *NCSA Mosaic* took the Internet by storm. Not surprisingly, in 1995 the WWW surpassed ftp as the Internet service with the largest traffic on packet count as well as on byte count. *gopher* was superseded by a much more powerful tool within a few months. However, it took less than four years before the development and support of *NCSA Mosaic* itself was terminated in January 1997. The competition between *Netscape* (URL:Netscape) which went public in 1995 and *Microsoft* with its *Internet Explorer* (URL:InternetExplorer), introduced in 1996, often referred to as the "WWW browser war", resulted in improved Web browser releases with many additional features every few months, but also pushed smaller competitors out of business.

Web browsers plus additional plug–ins can display all types of data and information as different as text, a large variety of image formats, data in different formats, sound, videos, and real–audio and real–video signals. Most commonly used formats and languages, respectively, are the *hypertext markup language* (html), *JAVA*, and *CGI* (mostly *PERL*) scripts. These are the building–blocks for most Web pages. New WWW technology is introduced every few months, but, as history has shown, only very little of the current Internet technology will still have some impact in 5 or 10 years.

3 Internet Technology and Statistics

3.1 Statistics on the Web

Inspired by Taerum & Nelson (1997), we used a WWW search engine, *excite* (URL:Excite), on 5/4/98 to report the number of hits on particular keywords. The result was 3,558,830 hits for "Internet", 692,240 hits for "Statistics", and 10,930 hits for "Internet and Statistics", but only 562,870 hits for "Sex". So, is "Statistics" more popular than "Sex" among WWW users? Obviously not, as some statistics about the most frequently used words in search engines indicate. The top 10 from the *Top 100 Words on the Web* (URL:Pretext) are sex, chat, Playboy, Netscape software, nude, porno, games, porn, weather, and Penthouse. Almost identical results can be found at *Yahoo! Top 200 Search Words* (URL:Eyescream).

So, based on these statistics, the Web seems to be a place where we can obtain information on diverse, but certainly not very professional topics. — Well, as we will see later on in this section, this is not quite true. In addition to other Internet services such as e–mail, ftp, and telnet, the WWW offers a wide variety of features useful for professional statisticians.

3.2 Statistical Summary Pages

What can we do not to get lost in the 692,240 WWW pages related to "Statistics"? Fortunately, statistical summary pages have been created at many places. The *StatLib* (URL:StatLib) Web page mirrors (i. e., provides access to local copies) a large number of statistical software packages, data, and summarizes information on many statistically related questions. In addition, it also provides links to other Web pages with statistical content. Many other places (URL:Lievens/Verstraete, URL:Varnum/Weise, URL:Helberg) provide collections of statistical links but typically do not mirror any material. Of particular interest for the academic reader is the Web page maintained at the *University of Florida Department of Statistics* (URL:UFloridaStats) that provides an overview on statistics departments at universities world–wide.

3.3 Statistical Journals

Statistical journals are presented on the WWW at three different levels. It can be stated that almost every statistical journal has at least a homepage that provides information on recent or upcoming issues and lists editors and associate editors. Journals of this type are, for example, *Computational Statistics* (URL:ComSt) and *Computational Statistics & Data Analysis* (CSDA) (URL:CSDA).

At the second level, there are journals that have homepages and provide additional material on the Web. The *Statistical Computing and Statistical Graphics Newsletter* (URL:SCSG) has on–line versions of its articles in ps and pdf format that date back to April 1993. The *Journal of Computational and Graphical Statistics* (JCGS) (URL:JCGS) allows to place additional material on the Web that goes beyond the basic material presented in the printed article. Updates on the authors' contact addresses or recent developments after the publication of the printed article can also be placed on the Web in this journal. It appears that the combination of printed and electronic material will be the ideal form for statistical publications in the near future.

There are major concerns whether full electronic journals — the third level of statistical journals that are only available on the WWW — will gain more importance within the next few years. Currently, the *Journal of Statistical Education* (JSE) (URL:JSE) and the *Journal of Statistical Software* (JSS) (URL:JSS) are the only two full electronic journals that focus on statistics. JSE has been founded and is on–line since July 1993 and basically provides html documents. JSS exists since 1996 and provides access to ps and pdf files as well as downloadable software and data. We will see in Section 4.2 that currently the disadvantages outweight the advantages (availabilty for a larger readership, less time between acceptance of a manuscript and its publication) of full electronic journals.

3.4 Statistical Software

In 1986, Nash (1986) provided a summary on then current methods for publishing statistical software. Traditionally, listing of programs, mostly FORTRAN or BASIC, were provided in books and journal articles. Subroutines from scientific software libraries such as NAG had frequently been used. Also, according to Nash, "authors of technical reports and journal articles may offer to make their programs available privately". Alternative publishing mechanisms were shareware or freeware, computer bulletin boards, and the use of electronic mail (strangely abbreviated as elmail by Nash). Nash was probably surprised how soon (and in what manner) his prediction "it seems likely that statistical software will be published mostly by electronic means at some time in the not so distinct future" became true. Fortunately, Nash's "figure of the order of 25 cents/1000 characters is probably reasonable at the present time for the communications costs" does not hold any longer — otherwise the author of this present paper would have spent a few million US–$ on downloading software during the last months just to write this paper.

Today, basically every statistical software is available in electronic form — on CD, floppy disk, through the WWW or ftp sites, or combinations of these. Often, commercial software vendors deliver a preliminary version of the software on CD or floppy disk to the customer. Updates, bug fixes, or new releases are later made available through the WWW. It is technically possible to restrict the access to Web pages so that only registered customers get access to this additional material. Homepages of software companies typically contain additional documentation on the software, answers to frequently asked questions (FAQs), and information how to get additional help on the product. Examples of homepages for commercial software are the Web pages of the *SAS Institute*'s product palette *SAS* (URL:SAS), *JMP* (URL:JMP), and *StatView* (URL:StatView) or *MathSoft*'s *S–Plus* (URL:SPlus) Web page.

Freely distributed software is widely available on the WWW today. User–developed routines for particular statistical packages are available from statistical summary pages, e. g., a large number of *S–Plus* routines from *StatLib* (URL:StatLib), or directly from the software authors' Web pages. Software (and data) is also published as an electronic appendix to articles in electronic journals such as JSS (URL:JSS). Entire packages such as *XGobi*

(URL:XGobi) and *R* (URL:R) are freely available on the Web, but, in addition, articles such as Swayne, Cook & Buja (1998) and Gentleman & Ihaka (1997) are published to make readers of classical paper journals and conference proceedings aware of these products. Mixtures between commercial and freely distributed software also exist. As an example, *XploRe* (URL:XploRe) executables can be downloaded for a free two month evaluation. While the UNIX version of *XploRe* is generally free, the interested user can purchase the PC version after two months (or simply download another copy for another two months).

Many JAVA–based statistical routines and packages have been developed recently. So far, these are mostly non–commercial packages that have been developed at universities or research institutes. No commercial JAVA–based statistical package exists so far. However, it seems to be only a question of months, rather than years, when a Web browser will be used to access commercial statistical JAVA–based software. The user most likely will indicate the URL that contains the data to be analyzed and the methods that should be applied. The results will be displayed on the user's terminal. The user will be charged for this service, based on the CPU time needed on the provider's site to conduct the statistical analysis. Both batch mode (where the calculations are made on a supercomputer at the site of the service provider and only results are displayed on the client's computer) and interactive mode (where calculations are made on the client's computer) are features likely to become common in the near future.

Currently, WebStat (URL:WebStat) and the JAVA version of XploRe (URL:XploRe-Java) allow the input of user data through the indication of a URL. Unfortunately, this works only with the latest Web browsers and JAVA versions. It might take a few more months before such browsers become standard. At the current point of time, most users of these packages have to copy and paste the data from another window into the data window of the JAVA package. Additional information on these two packages can be found in West & Ogden (1997), West & Ogden (1998), and Schmelzer, Kötter, Klinke & Härdle (1996). Unfortunately, disadvanategs similar to those mentioned earlier for full electronic journals exist for statistical software that is maintained on a single site on the WWW.

The *Globally Accessible Statistical Procedures* (GASP) Initiative (URL:GASP) is the first approach to create a statistical summary page that consists entirely of links to JAVA–based applications useful for introductory statistical classes, particular research topics, and pointers to JAVA–based statistical packages. More on GASP can be found in West & Piegorsch (1997), where an example in environmental biology has been highlighted.

The *Graphics Production Library* (GPL) (URL:GPL) is a tool that allows to create graphics in the style of Cleveland (1993), Cleveland (1994), and Carr (1994) by modifying an html document and calling a JAVA applet that evaluates the information from the html page. More details on the GPL can be found in Carr, Valliant & Rope (1996). Other JAVA–based software packages are the *S/JAVA* interface (URL:SJava) and *Mondrian* (URL:Mondrian).

3.5 Teaching Statistics

The idea to incorporate software, in particular teaching software (often called teachware), at an early stage of the statistical education of students has become very popular during the last few years. Lock (1997) provides a general overview on Internet resources for teaching statistics, Rossini & Rosenberger (1994) and Rossini & Rosenberger (1996) discuss the merits of the WWW as an additional teaching assistant, and Currall (1997) and Roenz (1997) discuss general ideas of computer–aided teaching in statistics. Two teachware products that can be downloaded at no charge through the WWW are the STEPS (*STatistical Education through Problem Solving*) (URL:STEPS) and *Quercus* (URL:Quercus) software packages. STEPS, described in Redfern & Bedford (1994), Bowman (1997), and Bowman, Currall & Lyall (1997), focuses on statistical applications in biology, business, geography, and psychology. *Quercus* is aimed at tutoring bioscience students in the basic techniques of statistical analyses. Harner & Wojciechowski (1998) and Wojciechowski & Harner (1998) present Web–based teachware for introductory statistical courses. Their software consists of JAVA applets and links to Lisp–Stat constructor functions.

Entire (introductory) statistical textbooks have been designed for the WWW. Examples are David M. Lane's *HyperStat* (URL:HyperStat), Jan de Leeuw's *Statistics: The Study of Stability in Variation* (URL:StudyOfStability), and David W. Stockburger's two books on *Introductory Statistics: Concepts, Models, and Applications* (URL:IntroductoryStatistics) and *Multivariate Statistics: Concepts, Models, and Applications* (URL:MultivariteStatistics).

Kent L. Norman (URL:Norman) has probably conducted the most advanced experiment of using Internet technology for teaching so far. In his Psychology *Psyc 200: Statistical Methods* class, he did not allow any pen or paper at all — everything was in electronic form: textbook, notetaking, homework assignments, and exams. Solutions had to be turned in by e–mail or on Web

pages. e–mail and discussion lists were used for individual questions, to broadcast information to all students, and to allow an ongoing discussion among students and instructor. It is possible to request a guest password at URL:Norman to check the material directly on the Web. Obviously, an electronic classroom has been used during the lectures. See the technical report *Emergent Patterns of Teaching/Learning in Electronic Classrooms* (URL:ElectronicClassrooms) by Ben Shneiderman, Ellen Yu Borkowski, Maryam Alavi, and Kent L. Norman, for a general discussion on the use of electronic classrooms. Kent L. Norman's experiences with on–line classes are available in his on–line book *Teaching in the Switched On Classroom: An Introduction to Electronic Education and HyperCourseware* (URL:ElectronicTeaching).

3.6 On–Line Access

Almost every major organization that is of interest for a statistician, e. g., the *American Statistical Association* (ASA) (URL:ASA), the *Interface Foundation of North America* (IFNA) (URL:IFNA), and the *National Science Foundation* (NSF) (URL:NSF), is presented on the Web. Information on all major statistics departments is available on the WWW as well. Examples are the *Universität Dortmund, Germany* (URL:UDortmund), the *University of Sheffield, England* (URL:USheffield), *Iowa State University, Ames, Iowa* (URL:ISU), and *George Mason University, Fairfax, Virginia* (URL:GMU) — in fact all the statistics departments the author of this article has attended so far.

General information on conferences related to statistics is available through the on–line version of the *Amstat News Dateline* (URL:AmstatDates). Information on particular past or upcoming conferences and workshops such as the *Interface '98* (URL:Interface98), the workshop on *Data Visualization in Statistics* (URL:DataVis), and the workshop on *Statistical Science on the Internet* (URL:StatSciInternet) is available as well.

An invited talk in the session "Impact of the Internet on Education and Research in Statistics" at the *Interface '98* (URL:Interface98) conference was the basis for this paper. The presentation (and this paper as well) is accessible as a Web–browsable document (URL:InternetTalk). It is very likely that the other two speakers in this conference session, *Martin Theus* (URL:Theus) and *Duncan Temple Lang* (URL:Lang), also provide additional material (such as electronic versions of their transparencies or preprints of the proceedings papers) related to their talks through their personal WWW homepages. A large number of individuals in academia (faculty and students) maintain up–to–date personal homepages on the WWW that often contain the person's curriculum vitae and resume, links to electronic versions of technical reports the person authored, preprints of upcoming papers, outlines of recent presentations, and access to interesting data the person analyzed. Personal data and links to non–scientific Web pages can also be found on many personal homepages. Understandably, personal homepages in industry are less frequent and often less detailed.

3.7 Other On–Line Features

Electronic submission of abstracts and registration are possible for major statistical conferences such as the *Joint Statistitical Meetings* and the *Interface* conference. Submission of NSF research proposals through the (electronic) NSF–Fastlane is even required. Recently, JCGS (URL:JCGS) has started to request electronic files (ps or pdf) to be sent by e–mail or made accessible through a Web/ftp side instead of mailing 5 paper copies previously required for a paper submission.

Many universities, but other organizations as well, provide restricted access through the Web (typically accessible only for people within the same university/organization) to their telephone directories, library catalogues, and other material available in local data bases. Many places provide a WWW interface that allows access to a local copy of the *Current Index to Statistics* (CIS) data base (URL:CIS).

Many new statistical applications that use Internet technology are currently under development or will be developed in the future. Kabacoff (1996) provides details on user–friendly WWW interfaces to data bases and statistical programs. Lee, Shing & Chu (1996) discuss the use of the WWW to collect data, and Arnold (1997) discusses how topics such as distance learning, virtual conferences, consulting, and teaching are influenced by Internet technology.

Finally, it is very convenient to browse through electronic bookshelves such as *amazon.com* (URL:Amazon) and order books directly through the Internet. We can also check *Springer* (URL:Springer), *Wiley* (URL:Wiley), *Chapman & Hall* (URL:ChapmanHall), and other publishers for new releases or complete overviews of available statistics books and immediately place an on–line order.

4 Discussion

4.1 Advantages

Obviously, there are numerous reasons why the Internet is so popular and has such a big impact on our everyday life — in private and as a statistician. First of all,

the Internet provides fast access to a diversity of services. We can hardly recall that just a few years ago we had to write letters instead of sending e–mail, submit a paper copy of an abstract or paper in time so it was at its destination before the submission deadline instead of making a last minute electronic submission through a Web side, or actually call a person instead of exchanging notes by e–mail. And then, think of all the services and information accessible through the WWW.

Internet technology is easy to use. Typically, learning how to operate a Web browser with a mouse and drag and click operations is a matter of minutes. In addition, no complicated or cryptic commands have to be memorized and no difficult new tools have to be explored.

The de facto standard of two established Web browsers, *Netscape* (URL:Netscape) and *Internet Explorer* (URL:InternetExplorer), that are nevertheless very similar, and a large number of additional plug–ins available for most hardware platforms make Internet technology available almost everywhere — at home, at work, when traveling, and at many public places such as libraries or so–called "Internet cafes".

Internet technology is cheap — for the consumer and the provider of these services. Typically, Internet access for private use in the United States costs less than the basic telephone line (however, the latter is still required in many cases to access the Internet through a modem from home). At many places, individuals can freely access the Internet while the associated organization (university, company, etc.) covers the costs. Since information is much faster available through Internet technology such as Web sites or e–mail, productivity simply raises by the fact that information, data, or documents are available within a few seconds while it previously took hours, days, or even weeks before printed copies of documents have been made available or data and software has been shipped on tape or floppy disk. The time we have to spend on finding the desired information on the Internet seems to be negligible.

From the provider of Internet services' point of view, there are many factors that help to reduce cost. First of all, information and services can be provided quickly and directly to the client without any intermediate salesperson or going through a time–consuming and expensive production process such as printing. While, previously, a large number of employees was required to answer questions and requests from clients, much of this information is easily accessible through Web browsers today. However, the actual cost reduction still has to be determined, because the creation and maintainance of attractive Web pages definitively represent a considerable cost factor.

Another advantage of Internet technology is the internationalization, discussed in more detail by Taerum & Nelson (1997). Information is made available to the entire world, and the entire world can find almost every bit of information on a Web server anywhere in the world. Summary statistics of accesses to Web servers typically show that people from all over the world use this capability, given they have the required technical equipment.

4.2 Disadvantages

As stated in the previous section, Internet technology is fast — at least when it comes to sending e–mail to any place in the world or accessing Web pages, telnet to another computer, or ftp to another computer on the same continent. However, access to oversea destinations can be very slow. The supporting hardware for Internet technology is not available at the same level all over the world or it is simply restricted, resulting in slower connections and less capacity for the transfer of information. A severe problem often is the inter–continental access. For example, everyone in Europe who tried to access information in the US knows how difficult this can be in the middle of the day, while access is reasonably fast in the morning. Access between other continents is even worse.

What about the reliability and security of Internet technology? Can we safely provide our credit card number when ordering through the WWW? Who can gain access to our account by monitoring our login sequence for a telnet session, who reads our typically unencrypted e–mails? What is the percentage of e–mails that never reach their destination? Unfortunately, no precise answers or numbers can be given at this point. However, we should be aware that these things may happen.

Another question addresses the reliabilty of services available on the Internet. If we buy a book, floppy disk, or CD, we have permanent access to this particular version and information. This is not true for services on the Internet. We always have to worry whether something is still there tomorrow — and, in particular, whether a particular version is still there tomorrow. There are many reasons why this is not guaranteed. What if the person that maintains a Web page changes the job, dies, or simply selects another private Internet provider? What if a Web server goes down permanently, for example when a company goes out of business, a research group is no longer funded, or an entire department is closed?

But even if the information is still accessible, is it still the same information we get the next day? Usually, old versions of text, software, and programs on Web servers are deleted after an update. In many cases, not even a trace of the old version is available. What can we do

if we want to see some older information, e. g., from one week or one year ago? What can we do if our research or work is based on an old software version accessible through the Web but this version is replaced by a new but incompatible version? Which data formats (e. g., ps, pdf, html, gif, jpg) will be still supported in the future, e. g., in one year and in 10 years? Do we have to update information we want to maintain permanently every few years to conform to then current technology or are we in a period where information quickly becomes inaccessible simply through the development of new technology? Some concerns are justified. Who is currently still able to read old 8" or 5 1/4" floppy disks or even UNIX stream tapes? And what about our old *nroff* and *troff* documents that were very popular for text processing on UNIX systems before LaTeX has been developed? Sure, for many old formats there are conversion programs (often freely available on the WWW), but it certainly requires considerable efforts to catch up with new standards.

Of course, we can think of possible solutions that help to preserve information on the Internet, e. g., Internet archives and distributed digital libraries. However, some questions come up immediately. First, who maintains such archives? Can we assume that libraries in the future will subscribe to journals by mirroring (i. e., providing identical digital copies) material from Web pages and make it available to their local customers? If a sufficiently large number of digital copies exists world–wide, this would certainly overcome some of the previously mentioned problems. In particular, we do not have to worry that information suddenly and permanently disappears from a single Web site. Access to digital libraries located on the same continent should be much more reliable than the previously described problems with inter–continental access. The next question is, how often such archives should be updated. For printed media such as books, it usually takes months or years before a new edition becomes available. An electronic document on the Web can be updated instantaneously, e. g., whenever an error has been detected. This leads to the final question. Which information should be stored? Possibilities are to maintain copies at fixed dates (e. g., every week, month, year) or to store a traceable history that allows to reconstruct the information available at a particular date.

Until such archives and distributed digital libraries exist for software and publications, it is not recommendable to rely on full electronic journals or software that is only accessible through a single Web site. Possible disadvantages listed earlier in this section seem to outweight the advantages (availability for a larger number of people, less time for new releases and updates). But even if we overcome these disadvantages, do we, as humans, really want to read everything from a computer screen? Or would we finally end up printing the information we retrieve from a Web page to have a copy at hand we can annotate individually and even read at places where we do not have a computer available?

Another severe disadvantage comes with the frequent changes of URLs and with the removal of Web pages. There is nothing that prevents authors of Web pages to remove material they do not want to display any longer for whatever reason. Also, the reorganization of material on the Web where files are renamed or the directory structure is modified occurs frequently. This makes it difficult to find a given document specified through a URL in case of a reorganization and impossible if this URL has been deleted. In academia, this can cause problems. Did the author of a paper really find this information on the Web, simply mistype the URL, or has this page been removed or renamed since the author's last access? As a reader/reviewer, we should not be too worried if one or two cited Web pages do not exist even after a short time of the publication date. But what if a relatively large number of URLs does not exist? Did the author do a bad job or is it just bad luck that so many cited URLs do not exist any longer?

Actually, how to correctly cite a URL is still a big open question. So far, we only indicate a URL as a reference and nothing else. For a book or an article, we indicate author, publication year, title, edition, editor, publisher, and place of publication. In addition to a URL, one should consider to indicate author, title of the document, webmaster (i. e., the person that is responsible for the electronic version but not necessarily for the contents), the date of the last update of this page, and the date of the last access to this Web page. Unfortunately, the complete information is not always available. Moreover, everybody who has ever created a Web page knows that the date of the last update often is not updated and therefore is the most unreliable part of a Web page and should be interpreted as "update on this date or later".

4.3 Dangers

One of the remarkable things of the Internet is, that people behave totally different from "real–life" — in particular when it comes to the use of the WWW. This can be small things up to important topics. Assuming that many people check floppy disks for possible viruses, why do not too many people proceed with the same care when it comes to downloading software and documents from a Web page? Do we really believe this software and these

documents are free of any computer virus?

More severely, why do so many people (including the author) voluntarily expose a large part of their private life such as information on leisure activities, friends, relationships, families, etc. on a Web page? Would the same people post the same information on a blackboard in their neighborhood or on a board in the main shopping center? Probably not. But why do we expose our otherwise well–kept privacy on the Web? A question only a psychologist may answer. Is it even advisable to post this material on the Web or can this information be used against us? For example, we might think of employers that carefully check Web pages and search information about a future employee on the Web instead of relying on a resume only.

The problem that e–mail can be read by unauthorized persons already has been mentioned. However, even more dangerous things can happen through fraudulent e–mails. Some mail programs allow to inidicate an arbitrary sender. How many people really check that the e–mail they just received is authentic? Of course, it is possible to learn more about an e–mail by carefully checking its header, but many modern mail programs by default do not display the mail header any more. Encryption also exists but it is not widely used today.

A further danger relates to privacy issues. Web browsers leave a traceable history, the so–called *cache*. There, a copy of the files the user downloaded in a particular time frame, often the last month, is maintained. What kind of information about an employee does this provide to an employer if the percentage of work related pages and leisure pages the person has accessed becomes known. And what power does this give to a system administrator, e. g., at a university, who can check which xxx–related Web pages students downloaded in a public computer lab although strictly prohibited at most universities.

So, what is the final destination of the Internet, and in particular, the WWW? Will it become a network where everyone has immediate access to all knowledge ever accumulated or will it lead to an Orwell–like world where everyone knows everything about anybody else with no space left for privacy? Obviously, Internet technology, like any other technology, is neither blessing nor curse. It is up to us, the humans, to determine whether it will be a blessing or a curse for us as statisticians and for mankind in general. Definitely, we should continue to develop new Internet technology — for statistics as well as for other applications. Nevertheless, we should always recall that there are certain disadvantages and dangers of this technology we better do not simply ignore.

Acknowledgments

Symanzik's work was supported by a National Science Foundation Group Infrastructure Grant DMS–9631351. Thanks are due to Natascha Vukasinovic for her helpful comments.

References

Arnold, J. T. (1997), 'The Mbone, Desktop Conferencing, and the Internet: New Tools for the Statistician', *Computing Science and Statistics* **29**(2), 344–353.

Baker, F. M. (1994), 'Navigating the Network with NCSA Mosaic', *Educom Review* **29**(1), 46–51.

Bowman, A. (1997), 'Computer–based Learning and Statistical Problem–solving', *Computing Science and Statistics* **29**(2), 312–318.

Bowman, A., Currall, J. & Lyall, R. (1997), 'The Birds and the Bees: Interactive Graphics and Problem Solving in the Teaching of Statistics', *Statistics and Computing* **7**(4), 237–246.

Carr, D. B. (1994), Converting Tables to Plots, Technical Report 101, Center for Computational Statistics, George Mason University, Fairfax, VA.

Carr, D. B., Valliant, R. & Rope, D. (1996), 'Plot Interpretation and Information Webs: A Time–Series Example from the Bureau of Labor Statistics', *Statistical Computing and Statistical Graphics Newsletter* **7**(2), 19–26.

Cleveland, W. S. (1993), *Visualizing Data*, Hobart Press, Summit, NJ.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Hobart Press, Summit, NJ.

Currall, J. (1997), 'Computer Aided Statistics Teaching: Real Advance or Technological Fashion?', *Computing Science and Statistics* **29**(2), 321–327.

Gentleman, R. & Ihaka, R. (1997), 'The R Language', *Computing Science and Statistics* **28**, 326–330.

Harner, E. J. & Wojciechowski, W. C. (1998), 'A Web–enhanced Introductory Statistics Class', *Computing Science and Statistics* **29**(1), 316–321.

Kabacoff, R. I. (1996), 'Developing Interfaces between the World Wide Web and Statistical Applications', *Computing Science and Statistics* **27**, 481–484.

Lee, T.-W., Shing, C.-C. & Chu, S.-C. (1996), 'Automating Statistics in WWW', *Computing Science and Statistics* **27**, 485–489.

Lock, R. H. (1997), 'Internet Resources for Teaching Statistics', *Computing Science and Statistics* **29**(2), 339–343.

Nash, J. C. (1986), Publishing Statistical Software, *in* T. J. Boardman, ed., 'Proceedings of the 18th Symposium on the Interface between Computer Science and Statistics', American Statistical Association, Washington, D.C., pp. 244–247.

Redfern, E. J. & Bedford, S. E. (1994), Teaching and Learning through Technology — The Development of Software for Teaching Statistics to Non–Specialist Students, *in* R. Dutter & W. Grossmann, eds, 'COMPSTAT 1994: Proceedings in Computational Statistics', Physica–Verlag, Heidelberg, pp. 409–414.

Roenz, B. (1997), 'Computer–aided Teaching in Statistics — Some Experiences', *Computing Science and Statistics* **29**(2), 319–320b.

Rossini, A. J. & Rosenberger, J. L. (1994), 'Teaching Statistics and Computing via Multimedia through the World Wide Web', *Statistical Computing and Statistical Graphics Newsletter* **5**(3), 1, 10–13.

Rossini, A. J. & Rosenberger, J. L. (1996), 'One more Teaching Assistant: The World–Wide–Web', *Computing Science and Statistics* **27**, 511–514.

Schmelzer, S., Kötter, T., Klinke, S. & Härdle, W. (1996), A New Generation of a Statistical Computing Environment on the Net, *in* A. Prat, ed., 'Compstat – Proceedings in Computational Statistics', Physica–Verlag, Heidelberg, pp. 135–148.

Swayne, D. F., Cook, D. & Buja, A. (1998), 'XGobi: Interactive Dynamic Graphics in the X Window System', *Journal of Computational and Graphical Statistics* **7**(1), 113–130.

Taerum, T. & Nelson, T. (1997), 'Internationalization of Statistics: The Internet', *Computing Science and Statistics* **29**(2), 354–359.

West, R. W. & Ogden, R. T. (1997), 'Statistical Analysis with WebStat, a Java Applet for the World Wide Web', *Journal of Statistical Software*. On–line Journal at `http://www.stat.ucla.edu/journals/jss/v02/i03`.

West, R. W. & Ogden, R. T. (1998), 'WebStat: An Environment for Statistical Analysis on the the World Wide Web', *Computing Science and Statistics* **29**(1), 307–310.

West, R. W. & Piegorsch, W. W. (1997), 'Interactive Statistics on the Internet: Applications in Environmental Biology', *Computing Science and Statistics* **28**, 439–444.

Wojciechowski, W. C. & Harner, E. J. (1998), 'Learning Statistical Concepts Using Web–based Dynamic Graphics', *Computing Science and Statistics* **29**(1), 322–326.

URL References

URL:Amazon:
`http://www.amazon.com`

URL:AmstatDates:
`http://www.amstat.org/publications/amstat_news/Pages/Pages/dateline.html`

URL:ASA:
`http://www.amstat.org/`

URL:CERN:
`http://www.cern.ch/CERN/WorldWideWeb/WWWandCERN.html`

URL:ChapmanHall:
`http://www.chaphall.com/`

URL:CIS:
`http://galton.uchicago.edu/~cis/`

URL:ComSt:
`http://comst.wiwi.hu-berlin.de/`

URL:CSDA:
`http://www.elsevier.com:80/inca/publications/store/5/0/5/5/3/9/`

URL:DARPA:
`http://www.arpa.mil/`

URL:DataVis:
`http://www.research.att.com/~andreas/workshop98.html`

URL:DefenseLINK:
`http://www.defenselink.mil/`

URL:ElectronicClassrooms:
`ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/98-04HTML/98-04.html`

URL:ElectronicTeaching:
`http://www.lap.umd.edu/SOC/sochome.html`

URL:Excite:
`http://excite.com`

URL:Eyescream:
`http://www.eyescream.com/yahootop200.html`

URL:GASP:
`http://www.stat.sc.edu/rsrch/gasp/`

URL:GMU:
`http://www.galaxy.gmu.edu/`

URL:Gopher:
`gopher://gopher.micro.umn.edu/`

URL:GPL:
`http://www.monumental.com/dan_rope/gpl/`

URL:Helberg:
`http://www.execpc.com/~helberg/statistics.html`

URL:HyperStat:
`http://www.ruf.rice.edu/~lane/hyperstat/`

URL:IFNA:
`http://www.galaxy.gmu.edu/stats/IFNA.html`

URL:Interface98:
`http://www.stat.umn.edu/interface98.html`

URL:InternetExplorer:
`http://www.microsoft.com/ie/`

URL:InternetTalk:
`http://www.galaxy.gmu.edu/~symanzik/talks/1998_interface_internet/int98_start.html`

URL:InternetValley1:
`http://www.internetvalley.com/intval.html`

URL:InternetValley2:
`http://www.internetvalley.com/archives/mirrors/davemarsh-timeline-1.htm`

URL:IntroductoryStatistics:
`http://www.psychstat.smsu.edu/sbk00.htm`

URL:ISU:
http://www.public.iastate.edu/~stat/

URL:JCGS:
http://www.amstat.org/publications/jcgs/

URL:JMP:
http://www.jmpdiscovery.com/

URL:JSE:
http://www.stat.ncsu.edu/info/jse/

URL:JSS:
http://www.stat.ucla.edu/journals/jss/

URL:Lang:
http://cm.bell-labs.com/cm/ms/departments/sia/duncan/

URL:Lievens/Verstraete:
http://allserv.rug.ac.be/~flievens/stat.htm

URL:MCI1:
http://www.mci.com/aboutus/company/news/internetpolicy/basics.shtml

URL:MCI2:
http://www.mci.com/aboutus/company/news/internetpolicy/yesterday.shtml

URL:Mondrian:
http://www.research.att.com/~theus/Mondrian/Mondrian.html

URL:MultivariateStatistics:
http://www.psychstat.smsu.edu/MultiBook/mlt00.htm

URL:NCSAMosaic:
http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html

URL:Neosoft:
http://www.neosoft.com/internet/presentation1/index.html

URL:Netscape:
http://www.netscape.com/

URL:Norman:
http://cognitron.umd.edu/cognitron.html

URL:NSF:
http://www.nsf.gov/

URL:Pretext:
http://www.pretext.com/nov97/shorts/short4.htm

URL:Quercus:
http://www.stams.strath.ac.uk/external/quercus/

URL:R:
http://www.stat.auckland.ac.nz/r/r.html

URL:SAS:
http://www.sas.com/

URL:SCSG:
http://cm.bell-labs.com/cm/ms/who/cocteau/newsletter/index.html

URL:SJava:
http://cm.bell-labs.com/cm/ms/departments/sia/project/java/html/index.html

URL:SPlus:
http://www.mathsoft.com/splus/

URL:Springer:
http://www.springer.de/

URL:SRI:
http://www.sri.com/

URL:StatLib:
http://lib.stat.cmu.edu/

URL:StatSciInternet:
http://cm.bell-labs.com/cm/ms/who/cocteau/comsci/index.html

URL:StatView:
http://www.statview.com/

URL:STEPS:
http://www.stats.gla.ac.uk/steps/

URL:StudyOfStability:
http://www.stat.ucla.edu/textbook/

URL:Theus:
http://www.research.att.com/~theus/

URL:UCL:
http://www.ucl.ac.uk/home.html

URL:UCLA:
http://www.ucla.edu/

URL:UCSB:
http://www.ucsb.edu/

URL:UDortmund:
http://www.statistik.uni-dortmund.de/

URL:UFloridaStats:
http://www.stat.ufl.edu/vlib/statistics.html

URL:USheffield:
http://www.shef.ac.uk/uni/academic/I-M/ms/index.html

URL:UUtah:
http://www.utah.edu/

URL:Varnum/Weise:
http://henry.ugl.lib.umich.edu/chdocs/statistics/stat_guide_home.html

URL:WebStat:
http://www.stat.sc.edu/~west/webstat/version1.0/

URL:Wiley:
http://www.wiley.com/

URL:XGobi:
http://www.research.att.com/~andreas/xgobi/index.html

URL:XploRe:
http://www.xplore-stat.de/

URL:XploRe-Java:
http://www.xplore-stat.de/WWWJava/x4java.html

URL:Yahoo:
http://yahoo.com

All URLs have been verified between May and July 1998.