RC 22201 (Log W0110-039) 10/12/2001
Computer Science

IBM Research Report

An Analysis of the Prevalence of Cookies on the World Wide Web

Josh Hailpern
John Jay High School
Route 121
Cross River, NY 10518

Charles C. Palmer
IBM Research Division
Thomas. J. Watson Research Center
PO Box 704
Yorktown Heights, NY 10598

IBM Research Division
Almaden - Austin - Beijing - Delhi - Haifa - T.J. Watson - Tokyo - Zurich

Introduction

The communications protocol that runs the World Wide Web is the Hypertext Transfer Protocol, or HTTP [1]. A key feature of HTTP is that it is stateless. This means that a web server can treat each request independently, without saving any information from the previous request. As a result of this optimization, many desirable capabilities were difficult, if not impossible. These include personalisation, e-commerce, and web-based e-mail systems, among others. In order to provide these services, a scheme was devised by Netscape [2] that provided a scheme wherein a web server could store some persistent information on a user's system that could be later retrieved and sent back to the web site. These packets of information were called "cookies". Information like account numbers, passwords, or shopping cart status could then be stored in the cookie for upload back to the web server the next time the user visited that web site. Since potentially sensitive information might reside in a cookie, web browser software was extended to give the user some control over these cookies. In the paper *A Design of a Portable System for the Secure, Fine-grained Access Control of Hypertext Transmission Protocol Cookies*, Hailpern[3] showed that all of the currently available web browser and cookie management software that has come available since that time, fail to offer sufficient control over the use of this potentially private, personal information.

We wanted to develop new cookie management software that would meet all of the requirements put forth in [3], but we wanted to make sure that the cookie "problem" was widespread enough to merit our attention. The current paper presents our assessment of the prevalence of cookie use. We performed this assessment by means of two different sampling techniques: First, we describe the collection of a random sampling of web sites around the world, including an explanation of the techniques involved. Second, we used a select group of the most

popular web sites to see how cookie usage was different there. Then we offer a discussion of our findings. Finally, we discuss the future work remaining in this area.

Background

Prior Research

In [3], Hailpern presented his initial research results that showed how current cookie management software, both in browsers and in standalone applications, provided insufficient cookie management capabilities. In showing this, he established the need for a more complete software solution for cookie management. This resulted in the research described therein to design new cookie management software. As noted earlier, the prevalence of cookie use across the web was not addressed by this prior work, therefore the actual threat posed by cookies was unclear. In order to actually justify the need for a new software solution, an assessment of the prevalence of cookie use and misuse was needed.

Hypothesis

We hypothesized that at least 50% of all web sites and 80% of the most popular web sites employ cookies at the top level (i.e., “home page”) of their web site to store some kind of persistent user data. This hypothesis was based upon the personal experience of the authors and that of their peers.

Methodology

We set out to construct software that would emulate a web browser by taking a web site, or URL, as input, determining whether there was a live web server there, sending that site an

HTTP request for a web page, and then noting the cookies are attempted to be set by that web server.

For the first phase of this study, the authors had planned on obtaining a complete list of the currently-defined domain or site names from the InterNIC[4] as one of the authors had done in the past. Unfortunately, the domain lists were no longer freely available[1], even to academics. As a result, the authors had to design another piece of software that would produce a random sampling of the more than 3.3 million [4] domains on the Web.

An Internet address currently[2] consists of four numbers, each ranging from 0 to 254, and typically written as four numbers separated by three periods, e.g., 9.8.2.3. Each of the numbers is referred to as an "octet", since each is represented as eight bits. Due to the evolution of how IP addresses were assigned, a strictly random generation of four integers used together as an IP address would not provide a reasonably random sample. Instead, the addresses were generated using a Perl program written by the authors with rules designed to produce addresses that are more likely to be valid, live Internet addresses [6]. These rules are:

- Octet 1: A random number between 1 and 9 or 11 and 223, inclusive. The random selection was biased towards numbers greater than 127. This was done because the majority of today's web sites are at Internet addresses with a first octet greater than 127.
- Octet 2: A random number between 1 and 253, inclusive, such that (a) if Octet 1 is 172 then Octet 2 cannot be 16, and (b) if Octet 1 is 198 then Octet 2 cannot be 162.
- Octets 3 and 4: A random number between 1 and 253.

A test message was then sent to the newly-generated Internet address. If it responded, then the host name assigned to that address was obtained and the first portion of the name replaced by

---

[1] Apparently some Internet users were downloading the list and using it inappropriately.
[2] The new IPv6 standard will provide two additional octets for addressing.

"`www.`" Then another message requesting a web page was sent to that updated host name. For example, if the Internet address `92.32.15.200` was found to be assigned to `mail.whopper.com`, then a test message would be sent to `www.whopper.com`. If that host name responded, then the host name is added to the random list of web sites.

The authors used this technique to generate over 6500 unique valid web sites, or about 0.2% of the estimated total number worldwide, which were then tested for cookie usage, keeping statistics like which web sites create cookies from their home page and how many.

The second phase of this work involved using lists of popular web sites. These lists are produced by various organizations on the Internet such as ZDNet [5] and others. The authors used these lists to construct a list of the 500 unique "most popular" web sites. The complete list of the popular web sites is included in appendix 1. Once again, the potential web sites on the list were then tested for cookie usage and statistics were gathered as before.

Results

Random Sampling of Web Sites Around the World

Many of these randomly-selected web sites were found to be unavailable. This might have been due to a variety of reasons: server was off-line, server link to the Internet was down, server link to the Internet was too slow, server was password protected, and others.

The authors tested a total of 6,432 randomly selected web sites. From this test, the results shown in Table 1 were obtained. As part of the HTTP standard, all responses from a web server include a return code that is sent as part of the response. These codes tell the browser whether or not the web page request was successful. For example, a return code of 200 means that

everything worked just fine, whereas a return code of 400 means there is no such web page. These return codes are defined by the HTTP standard specification [1].

#	Result
489	Attempted to store one or more cookies
4,679	Did not attempt to store cookies
5	Returned empty pages (nothing came back from the server)
715	Address not running a web server
1	rc=400: No default web page defined[3]
13	rc=401: Unauthorized to access the web page
64	rc=403: Forbidden access to the web page
63	rc=404: Web page not found
8	rc=500: Internal web server error
395	time out, caused when there is no response after some specified time period

Time outs might be caused by a slow server, high hit rate, or repair work on the page. As a result, the server can not be reached in a reasonable time frame and that situation is described as a "time out". Null pages can be caused by a closed web page or one that no longer exists. Overall, a total of 971 cookies were returned by 489 of the web sites in the random sample.

Popular Web Sites

The results from the random test seemed to show that only about 10% of the web sites tried to store a cookie. Since the author's experiences implied a much larger proportion, they decided to perform another test. This time, the authors tested 479 web sites that were identified as the most popular in [5]. From this test, the following results were obtained:

189	Attempted to store one or more cookies

[3] Some web servers fail to correctly implement the documented return codes. In this case, the web server should have returned rc=404.

249	Did not attempt to store a cookie
1	Returned empty pages (nothing came back from the server)
27	Address not running a web server
2	rc=500; Internal web server error
11	Time out, caused when there is no response after some specified time period

Overall, a total of 431 cookies were returned by 189 sites of the web sites in the popular sample.

Discussion

In general, these results imply that the more popular a web site is, the more likely it is to attempt to store one or more cookies on the user's computer. This intuitive result follows the argument that the more popular web sites offer some service such as e-mail, merchandise, or other services, that would benefit from the addition of state to the user's interactions with that web site.

Random Sampling of Web Sites Around the World

From these results, we determined that 489, or 7.6%, of the valid randomly selected web sites returned a total of 971 cookies, or about 2 cookies per web site. Further analysis of the data showed how many cookies were returned by each web site. From those results, the distribution of cookies set per web site was determined.

Cookies Set	Sites
1	357
2	93
3	29
4	18
5	10
6	11
7	4
8	4
9	4
10	2
11	1

12	1
13	0
14	1

It was determined that 73% of the web sites that returned cookies set 1 cookie, and 19% set 2 cookies. Yet, most significant is that for the 5168 valid sites, 971 cookies were returned. One can view this as saying that a user visiting random sites would typically receive 1 cookie for every 5 different sites visited.

Popular Web Sites

From the popular web site results, it can be determined that 185, or 42.0%, of the 438 valid web sites returned a total of 431 cookies. That represents an average of 2.3 cookies per web site. Following the same analysis procedure as before, the number of cookies set by each web site was determined. From those results, the distribution of cookies set per web site was determined.

Cookies Set	Sites
1	77
2	52
3	26
4	18
5	4
6	0
7	2
8	1
9	1
10	1
11	0
12	0
13	0
14	1

It was determined that 42% of the web sites that returned cookies set 1 cookie, 28% set 2 cookies, and 30% set 3 or more cookies. Although a greater percent of the randomly generated

sites that returned cookies set 1 cookie, a greater percent of sites that set 2 or 3 cookies were found in the results from the popular sties. Therefore, the chance of getting more than one cookie from a popular site is increased and these results would be more likely to be representative of the typical user experience. Following the same analysis as before, we found that of the 438 valid popular sites, 189 set a total of 431 cookies. One can view this as saying that a user visiting popular web sites would receive at least one cookie 43% of the time. For those sites that do use cookies, an average of between 2 and 3 cookies were actually sent.

One noteworthy observation made about both sets of data is that some web sites appeared to set unusually large numbers of cookies. Some of these, such as www.carparts.com, actually did set that many cookies in preparation for online purchases. Others, such as www.amazon.com, had higher apparent counts due to their partnering with other websites. For instance, `www.toysrus.com` and `www.borders.com` are both actually hosted by `www.amazon.com` Therefore, when a user visits `www.toysrus.com`, the user's browser is redirected to `www.amazon.com`, where cookies are then set by `amazon.com`. Similar instances occur with sites being hosted at `www.icl.fi` and `www.internetconnect.net`. As a result, some of the large number of cookie returns by singular web sites (anything above 5 cookies), is due in part to multiple sites being hosted by a single web site.

Conclusion/Future Work

From the analysis of the results of these tests, one can conclude that cookies are largely prevalent online. Therefore the potential and real abuses described by Hailpern [3] and others [8][9][10] are indeed a concern and that should be addressed. Without tools for the proper management of cookies, an individual's private and personal information could be misused. The

authors plan to continue this research into the final phase in which they plan to build and test a cookie management system that will provide a complete solution to address the various risks of cookie abuse.

Appendix 1: Popular Sites (page 1)

www.123greetings.com	www.autovantage.com
www.1800flowers.com	www.autoweb.com
www.1bookstreet.com	www.babygear.com
www.1stopauto.com	www.backup.com
www.2for1inkjet.com	www.bargaindog.com
www.4adodge.com	www.barnesandnoble.com
www.800.com	www.beachcamera.com
www.abcdistributing.com	www.beatnik.com
www.abercrombie.com	www.beckett.com
www.about.com	www.bestbuy.com
www.acura.com	www.beyond.com
www.adatom.com	www.bid4vacations.com
www.adobe.com	www.biddersedge.com
www.albertsons.com	www.bidtripper.com
www.all-yours.net	www.bidz.com
www.allcommunity.com	www.bigstar.com
www.allmusic.com	www.bigwords.com
www.allposters.com	www.bizbuyer.com
www.amazing-bargains.com	www.bizrate.com
www.amazon.co.uk	www.biztravel.com
www.amazon.com	www.blockbuster.com
www.americangreetings.com	www.bluefly.com
www.andale.com	www.bluelight.com
www.andysauctions.com	www.bluemountain.com
www.andysgarage.com	www.bmg.com
www.anyday.com	www.bmgmusicservice.com
www.aol.com	www.bn.com
www.aolshopdirect.com	www.bonustree.com
www.apcc.com	www.books.com
www.apple.com	www.booksamillion.com
www.art.com	www.booksonline.com
www.artistdirect.com	www.borders.com
www.ashford.com	www.bottomdollar.com
www.ask.com	www.bowcreek.com
www.astralspace.com	www.brandsforless.com
www.atnocost.com	www.buick.com
www.atyouroffice.com	www.buy.com
www.auction-sales.com	www.buybidwin.com
www.auctionrover.com	www.buyerzone.com
Www.auctions.com	www.cabelas.com
www.auctionwatch.com	www.callwave.com
www.audible.com	www.campmor.com
www.audiobookclub.com	www.carclub.com
www.audiofind.com	www.cardmaster.com
www.audiogalaxy.com	www.carfax.com
www.audiohighway.com	www.carparts.com
www.auktionsindex.de	www.carprices.com
www.autobytel.com	www.cars-on-line.com
www.autosite.com	www.cars.com
www.autotrader.com	www.carscene.com

Appendix 1: Popular Sites

www.carsdirect.com	www.deja.com
www.carsmart.com	www.delias.com
www.cartrackers.com	www.dell.com
www.cashmap.com	www.dillards.com
www.castlemountains.com	www.directcoupons.com
www.catalog.com	www.discounttiredirect.com
www.catalogcity.com	www.discovery.com
www.cc-inc.com	www.dobhran.com
www.cdnow.com	www.dooyoo.de
www.cdrom.com	www.doubleclick.com
www.cduniverse.com	www.doubledaybookclub.com
www.cdw.com	www.drugstore.com
www.cdworld.com	www.dsports.com
www.checkout.com	www.dvdexpress.com
www.chevrolet.com	www.e-cards.com
www.chipshot.com	www.e4me.com
www.choicemall.com	www.eangler.com
www.chryslercars.com	www.eastbay.com
www.chumbo.com	www.easyeinkauf.de
www.cityauction.com	www.ebags.com
www.classifieds2000.com	www.ebates.com
www.clinique.com	www.ebay.com
www.clubphoto.com	www.ebaymotors.com
www.cnet.com	www.ebworld.com
www.cnetgameshopper.com	www.ecampus.com
www.cnn.com	www.ecards.com
www.collectibles.com	www.ecircles.com
www.columbiahouse.com	www.ecost.com
www.compUSA.com	www.eddiebauer.com
www.compaq.com	www.edeal.com
www.compusa.com	www.edmunds.com
www.compusastores.com	www.edu.com
www.computers4sure.com	www.efollett.com
www.consumerreview.com	www.egghead.com
www.cooking.com	www.egreetings.com
www.coolsavings.com	www.egroups.com
www.costco.com	www.ehobbies.com
www.coupons.com	www.emusic.com
www.creative.com	www.enews.com
www.crucial.com	www.enutrition.com
www.crutchfield.com	www.ep.com
www.cvs.com	www.eroo.com
www.cybergold.com	www.esprit.com
www.cybermoo.com	www.estyle.com
www.damark.com	www.etown.com
www.danschocolates.com	www.etracks.com
www.dash.com	www.eve.com
www.dealdeal.com	www.eversave.com
www.dealofday.com	www.ewanted.com
www.dealtime.com	www.excitestores.com

Appendix 1: Popular Sites

www.exoridor.com	www.half.com
www.expedia.com	www.hallmark.com
www.express.com	www.hammacher.com
www.expressit.com	www.hardware.com
www.ezbid.com	www.hardwarestreet.com
www.familywonder.com	www.harley-davidson.com
www.fashionmall.com	www.hickoryfarms.com
www.fatbrain.com	www.hifi.com
www.fingerhut.com	www.hobbyrat.com
www.firstauction.com	www.hollywood.com
www.flooz.com	www.hollywoodvideo.com
www.florist.com	www.home-publishing.com
www.flowersusa.com	www.homedepot.com
www.fogdog.com	www.homegrocer.com
www.ford.com	www.homestead.com
www.fragrancenet.com	www.honda.com
www.franklincovey.com	www.honda2000.com
www.fredericks.com	www.hondamotorcycle.com
www.free2try.com	www.honesty.com
www.freebiestuff.com	www.hotbot.com
www.freebitz.com	www.hp.com
www.freeforum.com	www.hpshopping.com
www.freemousepads.com	www.hsn.com
www.freeride.com	www.hud.org
www.freeshop.com	www.ibaby.com
www.freeshopcoupons.com	www.ibeauty.com
www.ftd.com	www.ibm.com
www.fun2send.com	www.ice.com
www.furniture.com	www.iescrow.com
www.furniturefind.com	www.imandi.com
www.gap.com	www.imix.com
www.gateway.com	www.inshop.com
www.gateway.net	www.insight.com
www.gear.com	www.intellichoice.com
www.get-it-for-free.com	www.interplay.com
www.getmobile.de	www.iprint.com
www.getmusic.com	www.iqvc.com
www.gevalia.com	www.itargetcards.com
www.giantrewards.com	www.iwant.com
www.giftcertificates.com	www.jcpenney.com
www.gifts.com	www.jcrew.com
www.giggo.com	www.jcwhitney.com
www.gmbuypower.com	www.jeepunpaved.com
www.gmc.com	www.kbb.com
www.gocollect.com	www.kbkids.com
www.golfsmith.com	www.kozmo.com
www.google.com	www.ktel.com
www.googlegear.com	www.lancome.com
www.gotsavings.com	www.landsend.com
Www.greatergood.com	www.lillianvernon.com
www.greeting-cards.com	www.literaryguild.com
www.groceryworks.com	www.living.com
Www.grolier.com	www.llbean.com

Appendix 1: Popular Sites

www.lonelyplanet.com	www.nuauction.com
www.lordandtaylor.com	www.officedepot.com
www.loserkids.com	www.officemax.com
www.loveandromance.net	www.ohmygoodness.com
www.ltdcommodities.com	www.oldsmobile.com
www.lycos.com	www.onemade.com
www.magazineoutlet.com	www.onestopwinning.com
www.mall.com	www.onvia.com
www.marlo.com	www.outpost.com
www.marthabymail.com	www.overstock.com
www.marvel.com	www.pacprod.com
www.massmerchandise.com	www.pcconnection.com
www.mazdausa.com	www.penthouse.com
www.mediabay.com	www.personalogic.com
www.mediaplay.com	www.petopia.com
www.mercata.com	www.pets.com
www.messagemates.com	www.petsmart.com
www.micronpc.com	www.planetoutdoors.com
www.milesource.com	www.planetrx.com
www.mobshop.com	www.playboy.com
www.more.com	www.pocketstuffer.com
www.mothernature.com	www.pontiac.com
www.motortrend.com	www.popularlink.com
www.mp3.com	www.postcards.org
www.mp3board.com	www.powells.com
www.msn.com	www.powwow.com
www.musicforce.com	www.preisauskunft.de
www.musiciansfriend.com	www.priceradar.com
www.musicmaker.com	www.pricescan.com
www.mvp.com	www.pricewatch.com
www.mygeek.com	www.productopia.com
www.myhelpdesk.com	www.proflowers.com
www.myitem.com	www.qool.com
www.mypoints.com	www.qpb.com
www.mypostcards.com	www.quantex.com
www.mysimon.com	www.quixtar.com
www.nada.com	www.qvc.com
www.napster.com	www.radioshack.com
www.nbc.com	www.randomhouse.com
www.necx.com	www.rats2u.com
www.neimanmarcus.com	www.recycler.com
www.net2phone.com	www.redenvelope.com
www.netflix.com	www.reel.com
www.netjewels.com	www.refundcents.com
www.netmarket.com	www.regards.com
www.netstoreusa.com	www.rei-outlet.com
www.nextag.com	www.rei.com
www.nextplanetover.com	www.replaytv.com
www.nicecards.com	www.respond.com
www.nissan-usa.com	www.rioport.com
www.nordstrom.com	www.ritzcamera.com
www.northernlight.com	www.rubylane.com
www.nortonweb.com	www.saks5thavenue.com

Appendix 1: Popular Sites

www.salesco.com	www.targetshop.com
www.salesmountain.com	www.tavolo.com
www.samgoody.com	www.tendollars.com
www.sampleville.com	www.textbooks.com
www.saturn.com	www.theglobe.com
www.save.com	www.ticketmaster.com
www.sayit.com	www.tigerdirect.com
www.schnaeppchenjagd.de	www.tirerack.com
www.schwab.com	www.towerrecords.com
www.screensaver.com	www.toyota.com
www.searchiq.com	www.toysrus.com
www.sears.com	www.travelocity.com
www.searsportrait.com	www.tucows.com
www.seeuthere.com	www.tunes.com
www.sega.com	www.tvtoday.de/shopping/info/info.hbs
www.selfcare.com	www.twec.com
www.service911.com	www.ubid.com
www.sfbc.com	www.ubl.com
www.sff.net	www.universalstudios.com
www.shabang.com	www.urbanfetch.com
www.sharperimage.com	www.valpak.com
www.sheetmusicplus.com	www.valueamerica.com
www.shopathome.com	www.valuepage.com
www.shopfinder.de	www.valuepass.com
www.shopintuit.com	www.varsitybooks.com
www.shoplet.com	www.vehix.com
www.shopnow.com	www.victoriassecret.com
www.shoppinglist.com	www.videoflicks.com
www.shopsports.com	www.virtualflorist.com
www.sierratradingpost.com	www.virtualflowers.com
www.skyauction.com	www.virtualpresents.com
www.smarterkids.com	www.vitamins.com
www.smartshop1.com	www.vitaminshoppe.com
www.sonicnet.com	www.volition.com
www.sony.com	www.vw.com
www.sonymusic.com	www.walgreens.com
www.sothebys.com	www.walmart.com
www.southpark.com	www.wantads.com
www.sparks.com	www.warehouse.com
www.spiegel.com	www.warentest.de
www.spinner.com	www.wbstore.com
www.sportsauthority.com	www.webcertificate.com
www.sportsmansguide.com	www.webflyer.com
www.spree.com	www.webshoppersdream.com
www.stamps.com	www.webshots.com
www.staples.com	www.webswap.com
www.startsampling.com	www.weeklyfreebie.com
www.stoneage.com	www.wine.com
www.storerunner.com	www.winfiles.com
www.stromtarife.de	www.x10.com
www.surplusauction.com	www.xingtech.com
www.symantecstore.com	www.yahoo.com
www.target.com	www.yamaha.com

Appendix 1: Popular Sites

www.yimg.com www.ynot.com www.yodlee.com www.yvesrocherusa.com www.zanybrainy.com www.zappos.com www.zdnet.com www.zones.com	

Bibliography

1. HTTP protocol specification (RFC) 2109- http://www.cis.ohio-state.edu/htbin/rfc/rfc2109.html
2. PERSISTENT CLIENT STATE HTTP COOKIES http://home.netscape.com/newsref/std/cookie_spec.html 1999
3. Hailpern, J., "*A Design of a Portable System for the Secure, Fine-grained Access Control of Hypertext Transmission Protocol Cookies*" http://www.watson.ibm.com/leo/ScienceResearch/jjhs/hailpern2001.html
4. Zook, M., Internet Geography Research Project; See http://socrates.berkeley.edu/~zook/domain_names/index.html
5. ZDNet The Top 100 Web Sites- http://www.zdnet.com/products/stories/reviews/0,4161,2394453,00.html
6. Schales, D.L., Personal communication, June, 2001.
7. InterNIC. http://www.internic.org/ 15-Aug-2001
8. Berghel, Hal. "Digital Village: Caustic Cookies." Communications of the ACM Vol.44, No.5 (May '01): p19-22.
9. Fu, Kevin. Emil Sit. "Web Cookies: Not Just a Privacy Risk." Communications of the ACM Vol.44, No.9 (Sept '01): p120.
10. Daudelin, Arte. E-Security Advances for Everyday Banking, Bank Technology News.. Page 1. February 2000