Getting to Know You (Intimately): Surreptitious Privacy Invasion on the E-Commerce Web

Matt Curtin Paul Graves Shaun Rowland
Interhack Corporation
http://www.interhack.net/

Date: 2000/07/31 22:18:26

Abstract

Coremetrics has taken a bold step forward for marketers, going well beyond the typical use of Internet technology to observe and to record the online habits of unsuspecting Web users. While engaged in normal "electronic commerce" activity on various well-known Internet Web sites, we observed a user's name, postal mailing address, telephone number, electronic mail address, and many other personally-identifiable details being reported to Coremetrics. These data—we discovered 71 different variables—are formatted for easy entry into a database system that will build an extensive database of customers of sites that use Coremetrics' service. Personally-identifiable information sent to Coremetrics by one site can trivially be used to link the activity of that user on any site that uses Coremetrics' service. Thus, anything that any Coremetrics site learns about a user can be added to a single database that contains the sum of all information collected by all Coremetrics clients' sites. The more sites that use Coremetrics, the more extensive such a database will be on each user and the more users that would be in such a database.

1 Introduction

Web sites that use Coremetrics' service for tracking site usage are leaking extensive amounts of data to Coremetrics. As a user gives information to a site, typically during the process of making a purchase, that information is being sent not only to the vendor who needs it to complete the order, but Coremetrics, whom the vendor has introduced into the transaction, typically without the knowledge of the user.

This information is not the sort of "anonymous" and "aggregate" information that marketers claim (falsely, in our opinion) is harmless to consumers, but is very detailed, specific information like what's being purchased, the user's name, his mailing address, his email address, his phone number, etc.

This is implemented by embedding JavaScript code in vendor's web site that will cause a connection to be made to Coremetrics with a specially crafted query string that will report such details in a standardized format for easy entry into a database. This fetch is implemented as a request for an image, a 1 pixel blank GIF, what is commonly known as a "web bug".

Where we could get ahold of the JavaScript code in question, it was clearly obfuscated in an attempt to prevent a human from being able to read it. A cleaned-up version of one of these programs can be found in Appendix **??**.

Additionally, because the vendor's pages that include this JavaScript are often encrypted, our normal mechanism for learning what is happening on the network (a packet sniffer) was unable to determine many details.

To find the details of what information was being reported to Coremetrics, we identified the host to which all of these data are being reported and implemented a duplicate of it in our laboratory. Once our duplicate was running, we proceeded to surf Coremetrics client sites. Thus, all information that these sites were sending about us to Coremetrics wound up in our database, not that of Coremetrics.

2 Sites Investigated

During the end of May and beginning of June 2000, our Coremetrics decoy picked up our test network's traffic traffic referred to the decoy from eight Web sites, listed in Table **??**. Note that of these eight, only two leaked personally-identifiable information like a name, address, telephone number, or email address. All sent information about the activity of the user to Coremetrics, including such information as what products the user was browsing, the price of these products, etc. Importantly, every time that such data were reported, they contained the Coremetrics cookie, which would allow all activity on all Coremetrics clients' sites to be correlated into a single user profile. Thus, even though only two sites sent Coremetrics the user's name, once the name was associated with the profile, the activity of that user on all eight sites could be linked with his name. (It is also noteworthy that from our perspective—as an outsider who can see only what Coremetrics is collecting from us—we have no way to know whether such correlation is taking place. Coremetrics claims not to do this and we have no evidence to suggest that it does. The possibility, nevertheless, is there, and we have no way to tell if such correlation has begun.)

Site	*Leaking Name, Address, Phone Number, or Email?*
www.toysrus.com	√
www.ashford.com	
www.fusion.com	√
www.dxcart.com	
www.exofficio.com	
www.getplugged.com	
www.inchant.com	
www.lucy.com	√

Table 1: Sites Sending Personal Information to Coremetrics

Coremetrics clients' sites commonly also make use of banner advertising networks, which can leak information about the user's activity to still other third parties. Table **??** shows a list of other third-party sites cited, whether those sites can identify a user from request to request through the use of cookies, and the sites that introduce these third parties into the transactions.

Such prolific leakage seems to suggest that very little, if any, concern for the privacy of the sites' visitors exists.

3 On the Topic of Data Collection

Such detailed data gathering raises several questions, the central themes of which are "how much is too much?" and "who decides?" We will address these briefly. We want to emphasize that we are not accusing Coremetrics of malice to the Internet population. However, we believe the risks presented to be serious—much more serious than Coremetrics itself seems to believe—and worth the attention of everyone who uses the Web.

As for *why* this is happening at all, it's best to consider Coremetrics' business. Coremetrics provides a means for Web site operators to outsource the job of collecting and analyzing web site data. Instead of providing a *product* that will serve this function, Coremetrics provides a *service*. Thus, there is a need for Coremetrics to be introduced into the conversation between the client and the server.

3.1 What Coremetrics Is Doing Right

When we began our investigation, we started with the Coremetrics web site and the descriptions of its technology and service. Although we find their claims of "security" grandiose and somewhat tiresome, we were pleasantly surprised to see how up-front Coremetrics is about exactly what they're doing and how they do it. Additionally, Coremetrics provides two levels of opting out of their system. There is a complete opt-out, where the Coremetrics cookie, instead of having a user-specific identifier, contains a token that

Third-Party Site	*Cookies?*	*Referred By*
switch.avenuea.com	√	www.toysrus.com www.lucy.com
a1896.g.akamaitech.net		www.toysrus.com www.lucy.com www.getplugged.com
a1428.g.akamai.net		www.getplugged.com
medals.bizrate.com	√	www.fusion.com
view.accendo.com	√	www.getplugged.com
ads.admonitor.net	√	?
ad.doubleclick.net	√	www.petstore.com
207.178.130.149	√	www.getplugged.com (?)
partners.quokka.com		www.fusion.com
service.bfast.com	√	www.fusion.com www.lucy.com
209.24.233.190		www.fusion.com
216.35.185.221		?
ad.linksynergy.com		www.fusion.com
www.dxcart.com	√	www.inchant.com

Table 2: Non-Coremetrics Third Party Sites

Coremetrics servers should ignore. The second is a partial opt-out, where the data are collected as usual, but those fields they list as having "personally identifiable information" will not be recorded.

3.2 Why What Coremetrics Is Doing Right Isn't Enough

Although Coremetrics does clearly explain what it's doing—once you know that information is being collected by a third party and you know which third party to investigate, there are several problems with the Coremetrics system with regard to the privacy of individual Web users. We again wish to stress that we are discussing what is *technically possible* and what *can happen*. We simply cannot be sure of what actually is happening.

3.2.1 Collects Everything By Default

Information is being collected about surfers of Coremetrics' clients' sites without their knowledge and before they have the opportunity to decide whether they want to be tracked. Additionally, since they have not seen any descriptions of what's happening, they don't have the opportunity to see to what degree they're being tracked. Even those who don't mind banner advertising networks are likely to find the collection of their name, phone number, and email address to be extremely invasive. The fact that Coremetrics returns an invisible web bug instead of a visible image, perhaps one that would take the user to a description of Coremetrics and a list of exactly what data were collected in the transaction is strong evidence to support the assertion that this system was designed to work without the users' knowledge. Indeed, were the system drawing too much attention to itself, it would be a nuisance and could make the system unusable. So how quiet is quiet enough to avoid being a nuisance and how quiet is an attempt to avoid detection?

3.2.2 Implicitly Places Trust In Coremetrics

This approach relies on Coremetrics to continue to do the Right Thing with regard to its opt-out records and its handling of data internally. The nature of the HTTP cookies requires that if the system is enabled at all, the cookie in question will be sent along with the request to Coremetrics. Coremetrics cannot know whether it is to save the information sent until after it receives the information and looks at the value of the cookie. As a result, even those who have completely opted out of the system have their data reported; it's

up to Coremetrics to honor the web surfer's request to have the information ignored. If Coremetrics changes its policy and begins to read data marked for "opt out", there's no way for anyone to tell.

3.2.3 Opt-Out Mechanisms Fail

Failures happen in software [**?**], people change browsers [**?**], and people will sometimes use other computers. In any of these cases, the opt out mechanism is defeated. As a result, those who have explicitly opted out are once again being tracked with extreme detail, often without their knowledge. We believe that reliance upon such an unpredictable mechanism is unworkable.

3.2.4 No Guarantee Against Future Misuse

There is no guarantee about what will happen with the database that is built. What mechanisms are there in place to prevent the data from falling into the wrong hands—perhaps those who would like to use the data for blackmail—or being used in ways completely unrelated to the original purpose—perhaps being subpoenaed by a court that wants to know what a given user's activity on the web has been recently? Even if Coremetrics does take reasonable precautions to ensure the safety of the data and would fight such subpoenas, what happens if Coremetrics is bought by another company without the same convictions? Readers who consider such scenarios to be unbelievable might be interested to learn that Netscape Communications Corporation has still-unaddressed privacy problems with its Smart Browsing feature. These problems were first raised in 1998 [**?**]. Netscape has since been bought by America Online.

The primary difference between Coremetrics' possession of this information and some other apparently related situations is that Coremetrics, as a service provider to the vendor, does not own the data. Coremetrics doesn't have the option of selling that which is not theirs. Nevertheless, the potential for mishandling is present, as is the possibility of the data being stolen despite taking every reasonable precaution.

This is the tricky part about privacy: once private information is disclosed, there's no going back. There is no remedy against exposure of private information.

3.2.5 Data Collected Falsely Considered Reliable

The fifth problem is that data in the Coremetrics database is likely to be taken as reliable. What is to prevent someone from creating a simple program that will constantly feed bogus information to the Coremetrics data collector? We estimate that building such a program from scratch would take a web programmer no more than a few hours to create and to debug. Individuals could have a similar effect by adding a few lines of HTML on their web sites. For the purposes stated by Coremetrics, this margin of uncertainty isn't highly important. But considering this in the context of unintended use of the database, there could be very serious consequences.

3.2.6 Coremetrics Architecture Unnecessarily Enables Tracking Across Sites

Descriptions of the Coremetrics service seem to indicate that the tool is used only to identify visitors of a particular site, that demographic data provided to a given site operator is only data collected from his site. If this is true, there is only one imaginable reason for the Coremetrics architecture to work as it does, that is, by having *everything* from *every* Coremetrics-enabled site being reported to a single source. That reason is the ability for a global opt-out of all Coremetrics tracking, irrespective of the site that calls the Coremetrics code.

The use of a persistent cookie allows Coremetrics to track a user as he moves from Coremetrics-enabled site to Coremetrics-enabled site. Were this cookie different for each site, the risk posed to the Web user would be less significant, as multi-site profiling would be rendered much more difficult. However, it would require that an opt-out take place on each of the sites that the user visits. Whether to make the cookie global or local is a key decision, one that we would have made differently.

A trivial solution to this problem is the placing of Coremetrics data collectors in the domain name of the site using the service. So, if `www.example.com` is using Coremetrics, instead of having all data reported to `data.coremetrics.com`, it could be reported to `coremetrics.example.com`. Thus, only `example.com` data

would be collected with that cookie. Another site that uses Coremetrics would have a different server, and a Web user who visits both sites will have different cookies for each site that he visits.

4 Defenses

There are three main defenses against this specific threat from Coremetrics, though neither is sufficiently generalized to protect against future threats of the same type but other sources.

Blocking Access to the host `data.coremetrics.com` could be restricted, such that all attempts to report the data to Coremetrics will fail. This, of course, assumes that data do not start being sent to other sites for Coremetrics to start picking up.

Disabling JavaScript Because that is the way that the database is fed, disabling JavaScript will prevent the connections to Coremetrics from being attempted.

Disabling Cookies It is the presence of a cookie that allows one session to be associated with another and for the user to be tracked from site to site. Disabling the use of cookies, though it will not completely eliminate the ability to track, will make the problem more difficult from Coremetrics' perspective, as information leaked in one session will be much more difficult to associate information leaked in another.

5 Conclusions

As the Internet—and the Web in particular—becomes an increasingly important piece of civilized society, marketers and others are engaged in efforts to use this technology surreptitiously. There is money to be made in the collection of information about people. All of the benefit goes to the data collectors and all of the risk is borne by the Internet-using public.

Coremetrics' advertised service—providing Web site operators with information about people use their site—is a potentially useful and legitimate service. However, the system's design and even more importantly, how it is used in some cases, needlessly places users of these sites at risk. The fact that the tracking device is obfuscated *and* hidden from view suggests that despite Coremetrics' rhetoric, it does not place as high a priority on individual privacy as it should. The fact that the data collected are completely centralized, allowing activity across sites to be aggregated into a single profile has no legitimate explanation.

The Internet is becoming an increasingly dangerous place, not because of technology, but because of people. There is nothing new about the technology in question. There is nothing particularly novel about this or any of these user-tracking systems that we have found; they're merely a rather obvious collection of features that were never intended to be used together, resulting in systems that spy on users in ways never envisioned by the technologies' creators. Other systems are merely the result of poorly-considered designs and implementations. Each of us may draw his own conclusions about which systems fall into which categories.

Members of the public who do not understand this technology are often afraid of it, apparently for good reason. Vendors must stop trying to make a buck off of every impression on their Web sites and start giving serious consideration to the long-term harm that is likely to come from making the Internet out to be the most dangerous place on earth to do business.

A Variables Identified

During the course of our investigation, we determined 71 separate variables used for tracking users' activity. Variables of "special interest", that is, personally identifiable information, are listed in Table **??**. The complete list of variables in use can be found in Tables **??** and **??**. We list the purpose of the variable next to its name where we were able to make such determinations.

Variable Name	*Description*
ba1	Postal Mailing Address
ba2	Unknown
be	Email address
bp	Price (of an item being viewed)
bs	State of residence
bt	City of residence
by	Country of residence
bz	ZIP code
ct	City of residence
fn	First name
gd	Unknown
hf	Unknown
ln	Last name
pm	Product Manufacturer or Description
pn	Product Information
rf	Referring URL
ul	URL of page with Coremetrics web bug

Table 3: Variables of Special Interest

Variable Name	*Description*
a1	Unknown (ActionName1)
a2	Unknown (ActionName2)
a3	Unknown (ActionName3)
ag	Unknown
at	Unknown Integer
ba1	Postal Mailing Address
ba2	Unknown
be	Email address
bp	Price (of an item being viewed)
bs	State of residence
bt	City of residence
by	Country of residence
bz	ZIP code
ccf	Unkown Integer
cd	Unknown Integer
cg	Description of items being viewed, e.g., "Men's Watches"
ci	Identifier assigned to the site by Coremetrics (ClientID)
cn	Name of item being viewed (ContentName)
ct	City of residence
fn	First name
gd	Unknown
hf	Unknown
ln	Last name
mf	Manufacturer of item being viewed
n1	Unknown
n2	Unknown
n3	Unknown
n4	Unknown
nl	Unknown
nw	Unknown
on	Unknown Integer

Table 4: Coremetrics Variables We Identified, Part 1

Variable Name	*Description*
pa	ProductName
pc	Boolean (PageCount)
pi	Identifier of the Current Page (ClientPageID)
pm	Product Manufacturer or Description
pn	Product Information (PageName)
pn1	Unknown (PageExtraNumeric1)
pn2	Unknown (PageExtraNumeric2)
pr	Unknown Number
ps1	Part of the Referring URL (PageExtraString1)
ps2	Unknown (PageExtraString2)
pt	Unknown One-Letter Code (PluginType)
qt	Unknown Integer
rf	Referring URL
rnd	A pseudorandom number generated by the browser in JavaScript
s1	Unknown
s2	Unknown
s3	Unknown
s4	Unknown
sa	State
sa1	Postal Address
sa2	Unknown
sc	Category of product being viewed—see cn (Parent ContentName)
sd	Unknown
se	Search terms (e.g., “movie”, “video”) (Search)
sg	Unknown number (formatted 9.99, perhaps some kind of product price)
sl	Unknown boolean
sp	Phone number
sr	Unknown boolean
ss	State
st	City
su	Unknown product category (e.g., “product view”)
sy	Country
sz	ZIP Code
tp	Boolean value of TestPerm cookie (TestPerm)
tr	Unknown number (formatted 99.9)
ts	Boolean value of TestSess cookie (TestSess)
ul	URL of page with Coremetrics web bug
vn1	Part of Coremetrics JavaScript code version (Version1)
vn2	Part of Coremetrics JavaScript code version (Version2)
zp	ZIP Code

Table 5: Coremetrics Variables We Identified, Part 2

B Coremetrics JavaScript Used to Report Users' Activity

This appendix is composed of the JavaScript program that is embedded in pages of a site that uses Coremetrics' tracking service. We have performed basic deobfuscation in an attempt to make the code more readable. This code was retrieved from the `www.toysrus.com` site; other sites might have different versions of this software in place.

```
<!--
/* Pageview Data-Transport-Tag v.2.2.8, 04/13/2000;
   COPYRIGHT 1999-2000 COREMETRICS, INC. ALL RIGHTS RESERVED. U.S.PATENT PENDING. */

/* This line is used to set your Coremetrics client ID.
   It should match the client ID that you were given.
   Please replace 99999999 with your client ID.*/
ClientID = "90000002";

/* Please do not modify anything below this line. */
PluginType="C";
Version1="e2.2.8";

var Version2;
if(Version2==null){
    Version2="e2.2";
}
var date=new Date();
var Rdm=date.getTime()%10000000;
var ReferralURL;

if (ReferralURL == null || ReferralURL == "" || ReferralURL == "(none)") {
    if (navigator.appName == "Microsoft Internet Explorer" &&
        parseFloat (navigator.appVersion) < 4) {
        ReferralURL="unavailable";
    } else if(document.referrer=="undefined"){
        ReferralURL="";
    } else ReferralURL=document.referrer;
}

URL=window.location.href;
TestSess=x06530905("TestSess");
TestPerm=x06530905("TestPerm");

if (TestSess != "Yes") {
    document.cookie="TestSess=Yes";
}

if (TestPerm!="Yes") {
    expiredate=new Date();
    expiredate.setHours(expiredate.getHours()+5);
    document.cookie="TestPerm=Yes;expires="+expiredate.toGMTString()+";"
}

TestSess=x06530905("TestSess");
TestPerm=x06530905("TestPerm");

arg="pt="+PluginType+"&vn1="+Version1+"&vn2="+Version2+
```

```
    "&ci="+ClientID+"&rf="+x08226(ReferralURL)+"&ul="+x08226(URL)+
    "&se="+x08226(Search)+"&pn="+x08226(PageName)+
    "&pi="+x08226(ClientPageID)+"&cn="+x08226(ContentName)+
    "&sc="+x08226(ParentContentName)+"&ps1="+x08226(PageExtraString1)+
    "&ps2="+x08226(PageExtraString2)+"&pn1="+x08226(PageExtraNumeric1)+
    "&pn2="+x08226(PageExtraNumeric2)+"&a1="+x08226(ActionName1)+"&a2="+
    x08226(ActionName2)+"&a3="+x08226(ActionName3)+"&pa="+
    x08226(PromotionName)+"&pc="+x08226(PageCount)+"&ts="+TestSess+
    "&tp="+TestPerm+"&rnd="+Rdm;

pl=document.location.protocol;

if (pl!="http:"&pl!="https:") {
    pl="http:";
}

prearg="<img width=\"1\" height=\"1\" src=\"";
prearg+=pl+"//";
prearg+="data.coremetrics.com/cgi-bin/eluminate.cgi?";
postarg="\">";
dummyImageURL="http://data.coremetrics.com/cgi-bin/eluminate.cgi?"+
    x08226(arg);
Display=(prearg+x08226(arg)+postarg);
function x36273(ff) {
    var i=0,j=0;
    while (ff.charAt(i)==" ") i++;
    j=ff.length-1;
    while (ff.charAt(j)==" ") j--;
    return ff.substring(i,j+1);
}

function x08226(s){
    var tmp;
    s=""+s;
    s=x36273(s);
    s=escape(s);
    while(s.indexOf("+")>=0){
        tmp=s.indexOf("+");
        s=s.substring(0,tmp)+"%2B"+s.substring(tmp+1,s.length);
    }
    return s;
}

function x06530905(name){
    var arg=name+"=";
    var alen=arg.length;
    var clen=document.cookie.length;
    var i=0;
    while(i<clen){
        var j=i+alen;
        if(document.cookie.substring(i,j)==arg)
            return x05844687532(j);
        i=document.cookie.indexOf(" ",i)+1;
        if(i==0)
```

```
            break;
    }
    return null;
}

function x05844687532(offset){
    var endstr=document.cookie.indexOf(";",offset);
    if (endstr==-1)
        endstr=document.cookie.length;
    return unescape(document.cookie.substring(offset,endstr));
}
//-->
```