
Web Site Analysis

**Data Collection Techniques
on Selected Web Sites**

**Daniel Meredith
Sascha Meinrath
Open Technology Initiative
New America Foundation**

Memo prepared for
**The Second NPLAN/BMSG Meeting
on Digital Media and Marketing to Children** for the NPLAN Marketing to Children Learning Community

Berkeley, CA June 29 & 30, 2009
Sponsored by The Robert Wood Johnson Foundation

This memo describes the information collected about visitors to 10 popular food and beverage Web sites including the type of information; where the information is being stored; who has access to it; and how the information could be used. We also describe the techniques Web site maintainers use to gather this information and discuss the ways to prevent those techniques from working.

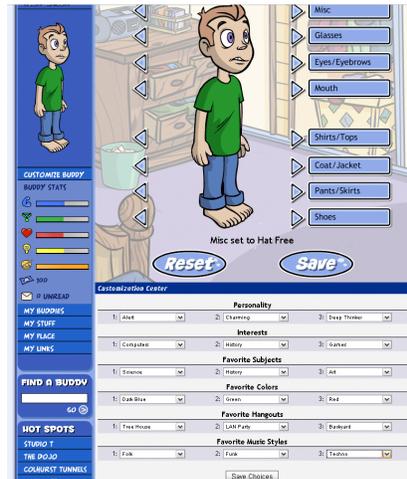
1. <http://www.millsberry.com>

Millsberry is a modern day fictional online town maintained by Mills Online, Inc a subsidiary of General Mills, Inc.¹ This mostly Flash-based site allows the visitor to navigate through Millsberry and requires the visitor to create an account to play games and visit most of the Web site's attractions. To create an account a user name and password along with visitor location information are requested. Visitors then proceed to create a personal avatar (a representation of himself/herself or alter ego)² and asked to provide personality traits, interests, favorite subjects, colors, hangouts, and music styles. Further, Millsberry has unique city districts where the visitor can choose to 'live'. These districts could be used to represent more of the visitor's personality traits. Millsberry has a currency, Mills bucks, that is used to purchase various items. The visitor can earn money by taking part in various activities and selling items to other visitors.

Millsberry.com uses Google Analytics^{3,4} to track visitor interaction with the Web site with the intention of strengthening advertisement potential and marketing initiatives. Google Analytics* makes use of JavaScript executed inside the visitor's web browser to save several HTTP Cookies⁵ on the visitor's computer. Per Wikipedia, a Cookie "is a small string of text stored on a user's computer by a web browser. A cookie consists of one or more name-value pairs containing bits of information such as user preferences, shopping cart contents, an identifier for a server-based session, or other data used by Web sites." Google Analytics uses the following cookies to store the following information:

- __utma - Often referred to as the "persistent" cookie in that will not expire or has an expiration date usually set years into the future. The information stored by this cookie typically tells the server how many times the user has visited the site along with the first and last time the user visited the site.

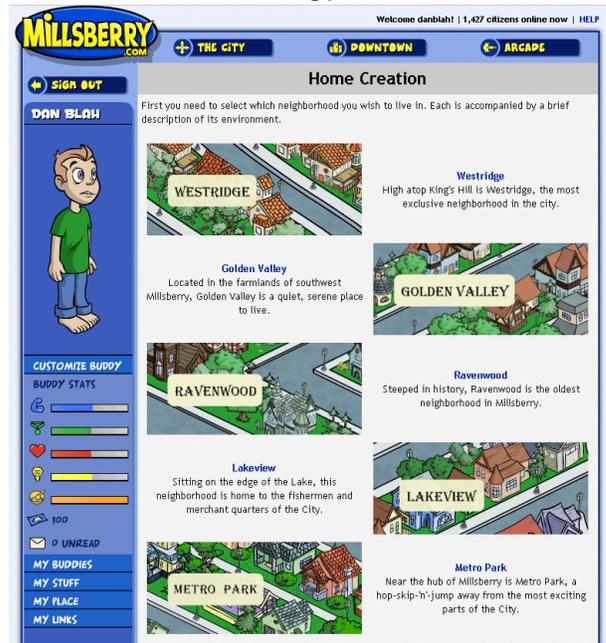
* Since Google Analytics is used by other sites in this paper we describe it only once in detail. Please refer to the above description of Google Analytics for all other sites that are using this service.



- __utmb and __utmc - These two cookies form a “team” that calculates how long you visited the site.

__utmb will typically expire as soon as you leave the site while __utmc will expire a short time after you leave the site. __utmc does this because it has no way of knowing when a user closes his or her browser or leaves a Web site, so it waits a short amount of time for another page view to happen, and if it doesn't, it expires.

- __utmz - This cookie tells the server how the visitor got to the site (for example, a user could type the domain directly into their browser, connect via a search engine result, or be referred by a link from another site), what link you clicked on or keywords you used to get there, and where you are located (usually by storing your IP address).



- __utmv - This cookie provides custom segmentation allowing reporting for that specific visitor. This is always a “persistent” cookie that will not expire or has an expiration date usually set years into the future.

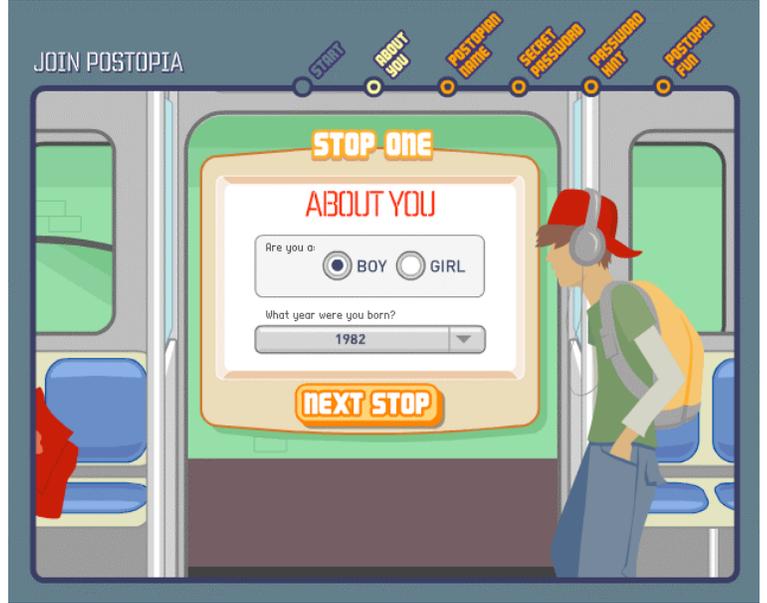
Further, Millsberry.com stores two unique cookies prefixed with gmi which can be assumed to stand for General Mills, Inc.:

- gmiData - This cookie is storing our created username to log into the site as well as our buddy list and favorite places information.
- gmiUnique - In this cookie the data stored is only a time stamp of when the visitor first arrived at the site.

Millsberry's Privacy Policy⁶ states that the company's policy is to create a Web site experience for children under the age of 13 while collecting minimal or no personally identifiable information. Despite this, Millsberry is amassing a large amount of user information. Although, it may not be considered personally identifiable information it is very descriptive of the user's personality and thus very valuable in creating both general and targeted marketing campaigns. They are likely using collected information to develop very specific marketing strategies for children under 13 both within Millsberry and for all General Mills products targeted at children. Further, the Privacy Policy allows General Mills to sell gathered information to 3rd party Web sites and sponsors for their own marketing campaigns or for the development of additional tools.

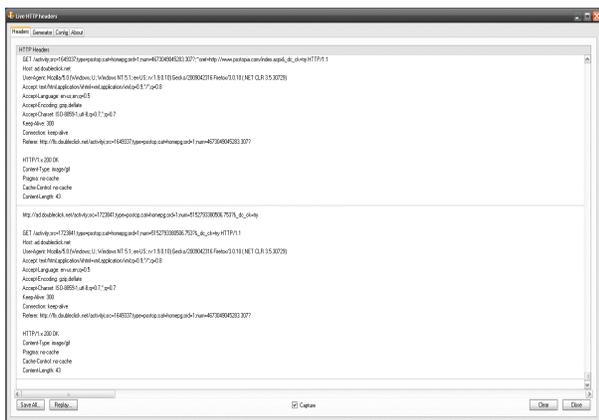
2. <http://www.postopia.com>

Postopia is another fictional city inhabited by the Flintstones, among other characters, and maintained by Post Foods LLC. Postopia would like the visitor to sign up for an account but does not require it for most areas in the site. The advantage of creating an account is the ability to save the fictional currency, PosTokens, earned when playing games and completing other activities. Creating a Postopia account requires basic personal information, the visitor's sex and birthday. Next, the visitor is asked to provide a name and create a password.



Postopia is less straightforward in its Web site analytics gathering than most. At first glance it seems as though the site is not using a 3rd party service to actively track what visitors are doing on the Web site. There are only three cookies being saved:

- **SNIFFER** - Despite the ominous name, this cookie is only identifying what version of Flash the visitor is using.
- **ASP.NET_SessionId** - This cookie doing exactly what it says, it is assigning a Session ID to identify the visitor and separate two clients that try doing identical things.
- **PERSIST** - The visitor will receive this cookie after creating an account. It holds an encrypted hash (a non-human readable string of alpha numeric characters) representing the user name to login.



Initial examination using Firefox as a web browser⁷ and the Firebug debugging extension⁸ to view Web site source and cookies saved, we were not able to identify any tracking service. To dig a bit deeper we can use the Firefox LiveHTTPHeaders extension⁹ which allows the visitor to view any information sent and received by the web browser. Using this tool, we can see in real time that the Postopia Flash application is sending nearly identical information stored in the Millsberry Google Analytic cookies to DoubleClick.net.

DoubleClick^{24,25} is another 3rd party Web site analytics firm recently purchased by Google and heavily geared towards the enterprise marketing needs of large online Web sites who maintain high visitor numbers.

Although Postopia is not storing nearly as much voluntary personal information, the site is storing an equivalent amount of browsing statistics including the user's IP address, its browser capabilities, what pages were viewed, and how long the user spent on the site. With the knowledge of the page content and browsing statistics, Postopia maintainers would also be able to create sound marketing strategies and develop advertising campaigns. For example, knowing where a user is located and which pages he or she spent the most time on, Postopia could easily make general and targeted assumptions based on the pages content to refine existing marketing campaigns and develop new campaigns.

3. <http://vw.happymeal.com>

This Web site is home to McWorld, another Flash based preteen-centric virtual world managed by the McDonald's Corporation. Unlike the first two virtual environments, McWorld requires the visitor login by creating a guest account or registering a new account before using any of the site features. After submitting a username, password, and date of birth, the visitor has the option to create an avatar. Within the world users are able to add other visitors as buddies, chat, also input special codes received in Happy Meals to 'Unlock Cool Stuff'.

McWorld is the first Web site we tested reporting to both Google Analytics and WebTrends.^{11,12} We will focus on WebTrends[†] since McWorld's Google Analytic settings are identical to Millsberry. The cookies WebTrends is saving on the visitor's computer are:

- WT_FPC cookie - This is WebTrends equivalent to the Google Analytics __utma cookie. Here again we see a "persistent" cookie that will not expire or has an expiration date usually set years into the future. The information stored by this cookie will typically tell the server how many times the user has visited the site along with the first and last visits.
- WEBTRENDS_ID cookie - This cookie stores the unique ID for the Web site assigned by WebTrends used internally by WebTrends.
- ACOOKIE - This is an encoded (not easily readable by humans) cookie that contains the visitor's IP address and the cookie creation time.



[†] Since WebTrends is used by other sites in this paper it will only be described once in detail. Please refer to the above description of WebTrends for all other sites that are using this service.

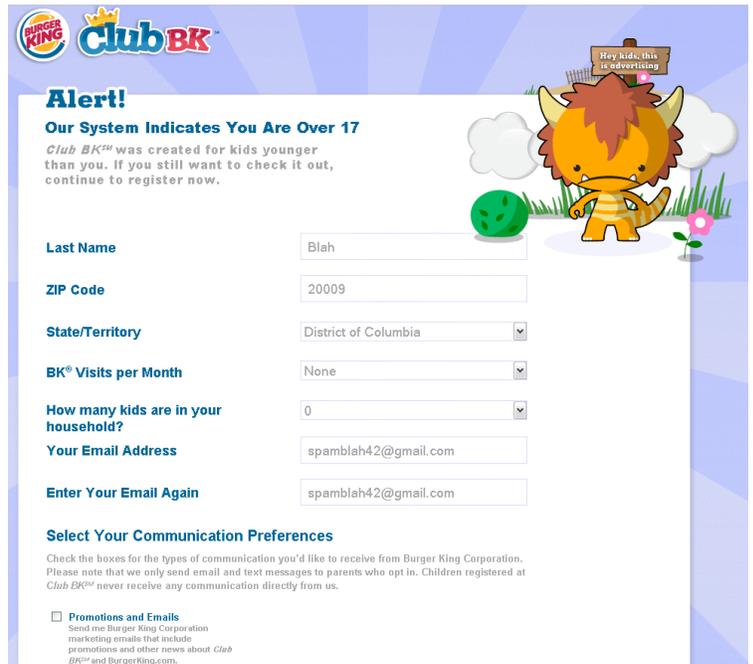
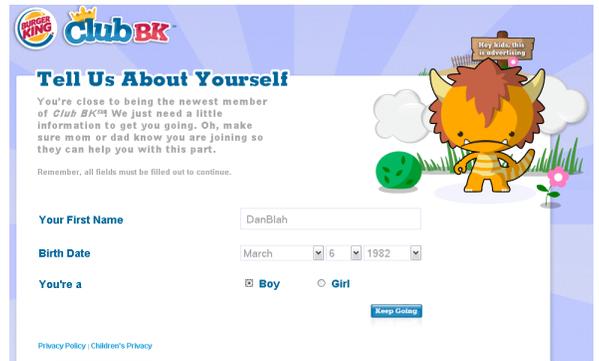
McWorld is collecting basic information about the visitor. Using both Google and WebTrends analytic software, the Web site will be able to identify the city and state of the user, what browser they are using, how long they stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed. Depending on how McWorld is using special codes entered by a visitor, the Web site could be used to identify more demographic information. For example, if the special codes are unique to a specific McDonalds location, once entered into the McWorld Web site the visitor becomes unique to a specific McDonalds location along with the area’s publically known demographic information.

4. http://www.clubbk.com

This Web site sponsored by Burger King is yet another virtual universe with three planets to visit. As with McWorld, access to the ClubBK Universe requires registration of new account. The registration process requires submission of a Zip Code along with an email address. Further, visitors have the option to sign up for newsletters and submit their cell phone numbers for SMS notifications. As with previous sites, there is an in-game currency used to buy various items. The visitor earns currency by completing quests within the universe.

Using both Google and WebTrends analytic software, the Web site will be able to identify the rough location of the user, what the browser capabilities are how long they stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed.

Along with using Google and WebTrends to provide basic analytics, the ClubBK site allows visitors to enter codes found in the BK Kids Meals for special items. With the registration process asking for specific location information and the analytics software, ClubBK maintainers would also be able to create targeted marketing strategies and develop advertising campaigns. For example, by asking for the visitor’s location, marketers would be able to relate the visitor’s response to marketing campaigns within ClubBK and the visitor’s response to marketing campaigns at their local Burger King location.



5. <http://www.whoppervirgins.com>

This is another Burger King Corporation project showcasing a documentary video describing the process and results of a demonstration where people in isolated regions around the world are presented with Burger King Whoppers and McDonalds Big Macs for comparison. The site is a Flash application with the primary purpose of playing the eight minute video. At the end of the film the visitor is given the option of sharing the video with a third party Web site like Facebook. The Web site itself is not actually asking for any information.

WebTrends is used on the site to collect visitor information including the location of the user, what browser they are using, how long they stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed.

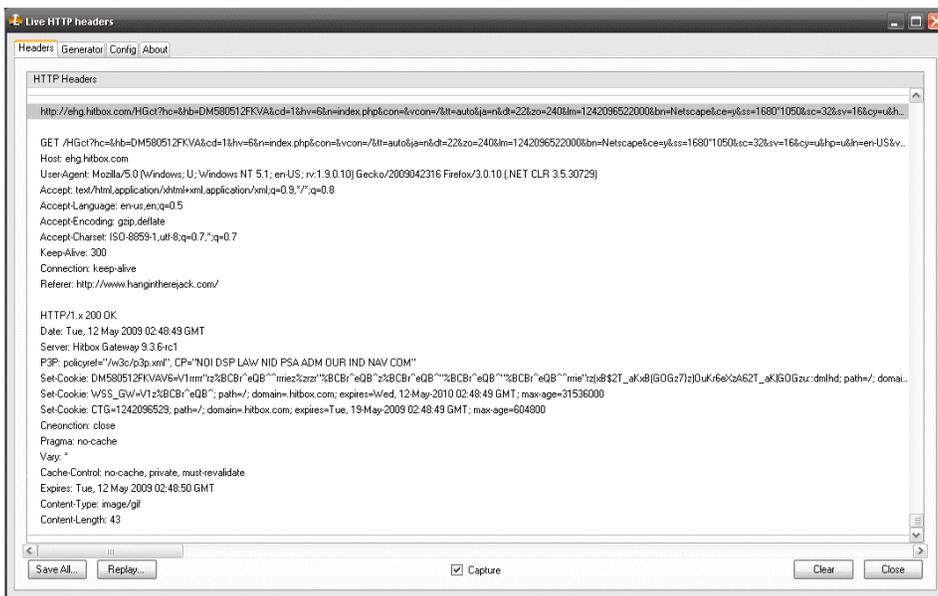
This is a very focused ad campaign designed to promote the superiority of the Burger King Whopper over the McDonalds Big Macs. Using data collected by WebTrends, along with the collected trends and third party social network interaction, Burger King will be able to track how the campaign propagates through the internet.

6. <http://www.hangintherejack.com>

Hang In There Jack was a marketing campaign by Jack in the Box, Inc. to promote its message of “Order anything on the menu any time of day” as well as hype its new Web site at <http://www.jackinthebox.com>. The campaign started when the fictional CEO Jack was hit by a bus in a commercial shown during the Superbowl. The scene and recovery progress were documented blog style and posted on third party social networking services like YouTube¹³, Facebook¹⁴, and Twitter.¹⁵ Visitors were encouraged in the videos and blog posts to comment and reply using these third party social networks.

Data collection for Jack in the Box is conducted on this site by an old player in the third party analytics

business, Hitbox. Hitbox is now a service of Omniture¹⁶ a company that was notorious for its analytic services provided exclusively to adult entertainment Web sites. It has since graduated to providing services for folks like Apple and Adobe. Omniture acquired infamy in the past for being one of the first to try and obscure the fact that it collected visitor information using techniques



that are now common practice. Using the same methods as WebTrends and Google Analytics now, these were once considered a violation of privacy in the early industry of computer security. Now, Omniture's use of inserting JavaScript into the footer of HTML code is the same method we have seen in earlier Flash applications allowing Web sites to identify the city and state of the user, what browser they are using, how long visitors stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed. Attempting to be somewhat elusive, using the Firefox web browser⁷ and the LiveHTTPHeaders extension⁹ was needed to view analytic information being sent to Hitbox servers.

Using Hitbox and third party social networking services, this site's goal is to drive visitors to a new Web site and establish brand recognition in a viral lower-budget method.

7. <http://www.facebook.com/papajohns>

This is a promotional page hosted by Facebook¹⁴ and maintained by Papa Johns Pizza. Facebook pages allow companies and organizations to create interactive social oriented sections where visitors can leave general comments, view and post both videos and photos, while sharing items with other friends using Facebook. To engage the page, a visitor will need to create a Facebook account. Although only a name and an email address is required to create a Facebook profile, everything from personal interests and employment history to phone numbers and a current address is requested and able to be accessed by Papa Johns or Facebook or both.

Despite using its own program that installs cookies on the visitor's computer, Facebook's analytics solution is doing the same thing as DoubleClick, Google Analytics, or WebTrends including storage of the city and state of the user, what browser they are using, how long they stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed.

The information potentially garnered from this site is highly specific. With substantial personal information saved in the visitor's profile and Facebook's tracking abilities, Papa Johns could end up with very specific demographic information on visitors to this page. This Web site, representing a partnership between a corporation marketing a product and the largest social networking platform, could provide marketing researchers a vast amount of data to create highly focused ad campaigns. This is mutually beneficial for both Facebook and Papa Johns. Facebook is able to take its comprehensive unmatched collection of user information used for social interaction a step further to identify a user's consumer trends. Papa Johns is able to ascend its proven and time tested commercial marketing knowledge with social trends. This would specifically be beneficial on college campuses where both Facebook and Papa Johns would be able to do a joint targeted marketing campaign in an area where they already have established recognition. Further, by creating an account, the visitor allows Facebook to share any collected personal information with third parties for any reason.

8. http://www.pepsiusa.com

This particular site is the primary portal page to the many Web sites managed by PepsiCo, Inc. As such this particular site is not asking for any information but redirects the visitor to specific marketing sites that proceed to solicit a good amount of information.

WebTrends is used to identify the city and state of the visitor, what browser they are using, how long they stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed.

With the use of WebTrends, PepsiCo is identifying what search strings lead visitors to the site and what specific PepsiCo Web site they are trying to visit.

9. http://www.mycoke.com

MyCoke is a Flash based Web site highlighting Coca-Cola Company's "Secret Formula" ad campaign, "Twist txt get" where special SMS text message codes are found under bottle caps, in a set of simple online games, and on an introductory site to the virtual online environment geared towards adults (13 or older) called CC Metro. CC Metro is the only area requiring registration of a Thirst ID. To go further into the CC Metro environment the visitor is required to install special software that will run on the visitor's computer. This is a similar approach to SecondLife.¹⁷

Here we have a similar situation to Postopia where again at first glance it seems as though Coca-Cola is not using a tracking service. There are only two cookies being saved:

- JSESSIONID - This cookie doing exactly what it says, it is assigning a Session ID to identify the visitor to the server.
- cTest - This is a common cookie assigned to test if the visitor's browser is accepting cookies.

However, using Firefox⁷ and LiveHTTPHeaders⁹ we were able to determine that the Web site is sending analytic information to WebTrends and DoubleClick^k to identify the city and state of the user, what his or her browser capabilities are, how long he or she stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed.

Along with the analytics data, it can be assumed that the companion application for CC Metro would allow the Coca-Cola Company to amass a large amount of visitor personality information. From the description on the introductory pages, CC Metro is likely to collect an equal or greater amount of information as Millsberry including user interests, personality traits, and would be physical traits. In a

Step 1

Ok, let's get started creating your **My Thirst ID**.

Please enter your date of birth, email address and let us know if you are a resident of the United States.

* Birthdate:

* Email Address:

* I am a resident of the United States: Yes No

Submit **Cancel**

Step 2

If you do not accept our [Privacy Policy](#) and [Terms and Conditions](#), you will not be able to register and therefore, not be able to access many of the features on MyCoke.com

* Screen Name:

* Enter a Password:

* Re-Enter Password:

* Gender: Male Female

* Zip Code:

* I understand and agree to The Coca-Cola Company privacy policy on this site

* I accept the terms and conditions of MyCoke.com

* Wanna hear about what's going on at MyCoke.com? Check here to get updates! Yes No

Would you like to receive emails from The Coca-Cola Company about other brands, new products or promotions? Yes No

Your security key is: 

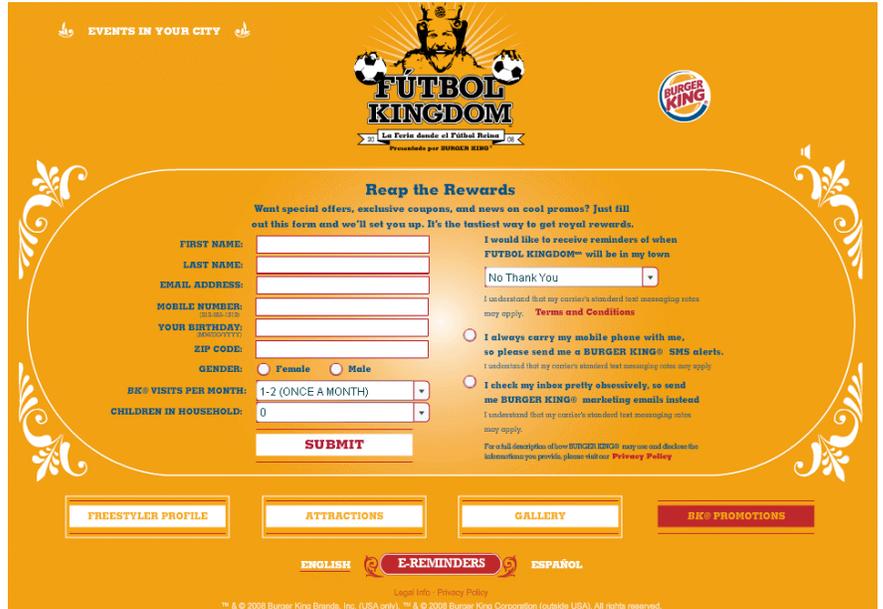
* Enter security key here: (case sensitive):

Submit **Cancel**

virtual environment controlled by the marketers, users will be used to test ad-hoc low cost marketing campaigns likely as precursors to campaigns used outside of the CC Metro virtual world.

10.http://www.futbolkingdom.com

Futbol Kingdom is a simple Flash site maintained by the Burger King Corporation to promote specific traveling events showcasing freestyle soccer players, various attractions, along with a 3vs3 pickup soccer tournament. The site has a single form asking for personal contact information including a cell phone to send out text messages and an email address for special promotions and coupons not related to the Futbol Kingdom event. Despite storing very



personal information including the visitor's full name and cell phone, the privacy policy²⁶ for Futbol Kingdom also allows 3rd party organizations contracted by the Burger King Corporation to have access to any collected information.

WebTrends is being used to identify the location of the user, what browser they are using, how long they stayed on the Web site, what Web site referred them, and what Web sites the visitor viewed.

SUMMARY

Analytic Systems Used

When visiting a Web site, a visitor is likely to be sharing information with more than just the Web site owners. Further, when using a third party data collection service, the Web site maintainers are automatically allowing a third party company access to their marketing data to be used in any way the third party company sees fit. On the Web sites we examined, we found third party data collection services WebTrends used five times, Google Analytics used three times, DoubleClick (owned by Google) used twice and Hitbox used once. Note that two companies (Google having recently purchased DoubleClick) make up the majority of the tracking systems we observed. This is likely to be the case for most Web sites. Maintainers are either using Google Analytics or WebTrends with the majority choosing Google Analytics due to its low to no cost. The significance here is that Google itself generates the majority of its revenue selling advertisements online, in newspapers, and on television. In most cases unknowingly, millions of Web sites choose to use Google Analytics for free in exchange for feeding Google marketing information on the Web sites' millions of visitors. This not only allows

Google to know where marketing trends are going, it is one of the many things that puts Google into the very powerful driver seat of technology.

How to Avoid Becoming a Marketing Statistic

Can a visitor prevent a Web site from storing information? For 99% of the information being stored, the unfortunate answer is going to be no for the average person browsing the Internet. However, with the proper tools, several preventative measures can be taken.

Most of what's being stored on the visitor's computer in cookies is also being stored on the server in log files. Why do these sites bother storing information on the visitor's computer in a cookie and then send it off to a third party when it is also being stored on the server? Typically this is done because a third party service is better at aggregating this information. Third party systems choose to store information on clients' computers because it is available on demand, quicker, and is not dependent on Web site maintainers to update their systems. Rather than be in the web analytics business, Web site maintainers have only to insert a small amount of JavaScript code into the Web site's HTML or the Flash application allowing the third party service direct access to visitor information.

Users can disable the execution of JavaScript and can not install Flash on their web browsers. Major browsers like Internet Explorer, Firefox, Safari, and Opera allow this. However, a good amount of Web sites today will not work properly without JavaScript execution enabled. Presumably, due to Google being in the analytics business and hosting heavy JavaScript web applications, it is not surprising that Google's new browser, Chrome, has no easy way to disable JavaScript.¹⁸

With JavaScript disabled a common saying in web development communities comes to mind: "If JavaScript can't do it, Flash can." Flash can because it runs as a separate program along with the browser, allowing it to do lots of interesting things (for example, interacting with a webcam, microphone, or speakers; storing files on the visitor's computer; and changing other system settings¹⁹). Although the default settings are fairly restrictive, it is not very difficult for an application to gain additional access. Using a built-in feature and for legitimate reasons, a Flash application can send and receive information from the visitor's computer, the Web site server, and a third party server. This is a feature allowing Web sites to display dynamic often changing content without the visitor having to refresh the browser. With respect to tracking, the application can send information to a third party server without the visitor even knowing about it. Further, due to the legitimate and frequent use there is no way to disable this. A technically savvy visitor can spot the tracking information being sent using Firebug's Console feature⁸ or another Firefox extension named LiveHTTPHeaders.⁹ Both of these tools allow the visitor to watch where and what the Flash application is sending and receiving.

The best way to stop Flash from collecting data is not to use it; disabling JavaScript and uninstalling Flash will ensure that Web site visitor's information is not being tracked by third party services. But since so many sites now use JavaScript and Flash, users may severely limit what sites they can access.

###

Overview of Data Collected by Selected Food and Beverage Web Sites

Web site	Requires Login	Post-login Incentives (games, coupons, etc.)	Asking for Email, Phone, or Zip Code	Asking for Personality Information	Using Web Site Tracking or Analytics
www.millsberry.com		X	X	X	X
www.postopia.com		X			X
vw.happymeal.com	X	X	X		X
www.clubbk.com	X	X	X		X
www.whoppervirgins.com					X
www.hangintherejack.com					X
www.facebook.com/papajohns		X	X	X	X
www.pepsiusa.com					X
www.mycoke.com		X	X		X
www.futbolkingdom.com			X		X

References (Cited URLs)

1. <http://www.generalmills.com>
2. [http://en.wikipedia.org/wiki/Avatar_\(computing\)](http://en.wikipedia.org/wiki/Avatar_(computing))
3. <http://www.google.com/analytics>
4. http://en.wikipedia.org/wiki/Google_Analytics
5. http://en.wikipedia.org/wiki/HTTP_cookie
6. http://www.millsberry.com/town_hall/terms.phtml
7. <http://www.firefox.org>
8. <http://www.getfirebug.com>
9. <http://livehttpheaders.mozdev.org>
10. <http://www.postopia.com/miscContent/privacy.aspx>
11. <http://en.wikipedia.org/wiki/Webtrends>
12. <http://www.webtrends.com>
13. <http://www.youtube.com>
14. <http://www.facebook.com>
15. <http://www.twitter.com>
16. <http://www.omniture.com>
17. <http://secondlife.com>
18. <http://nethead.blogspot.com/2008/09/google-chrome-how-to-disable-javascript.html>
19. http://www.macromedia.com/support/documentation/en/flashplayer/help/settings_manager.html
20. <http://adblockplus.org>
21. <http://www.privoxy.org>
22. <http://www.pps.jussieu.fr/~jch/software/polipo>
23. <https://www.torproject.org/overview.html.en>
24. <http://www.doubleclick.com>
25. <http://en.wikipedia.org/wiki/DoubleClick>
26. http://www.futbolkingdom.com/privacy_eng.html