Privacy Protections in Microsoft's Ad Serving System and the Process of "De-identification"

As part of its strong commitment to protecting individual privacy,
by design Microsoft bases its ad selection solely on data that
does not personally and directly identify individual users.

Microsoft Corporation

October 2007

The information contained in this document represents the current view of Microsoft Corp. on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This white paper is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.

Microsoft Corp. • One Microsoft Way • Redmond, WA 98052-6399 • USA

Introduction

As the Internet has matured, online advertising has become the means by which many Web sites offer users rich content and services for free. Online advertising has also become an increasingly sophisticated vehicle for targeting users' interests—based on context (such as car ads on a car Web site) or based on user behavior on a site. The serving of relevant ads benefits advertisers, who are more likely to find customers for their products. It also benefits consumers, who are less likely to see ads that do not interest them. The key is making sure that users' privacy is protected in the process.

As part of its strong commitment to protecting individual privacy, by design Microsoft bases its ad selection solely on data that does not personally and directly identify individual users. The vast majority of ads that Microsoft serves online are not targeted to specific known users—they are based on context or are untargeted. For individually targeted ads, Microsoft's ad serving platform stores the data used for ad personalization separate from contact information or any other data that directly identifies the user. The system also has strong built-in safeguards against unauthorized correlation of these sets of data. The key to these important privacy protections is the use of an "Anonymous" ID (ANID) to enable recording of relevant online user activity without correlating it with data that can be used to personally and directly identify a user. This paper describes how Microsoft uses the ANID as a part of the de-identification process it uses to achieve robust individual privacy protections while still serving relevant targeted ads to users of its Web sites and online services, including MSN® and Windows Live™ sites.

Overview of Ad Targeting

Generally, online ad targeting providers try to correlate the interests of users, as implied by their past behavior or demographics, with the ads those users are served. Users' perceived interests are inferred over time based on information they provide when they register with a Web site or service or actions they take and information they provide when interacting with the site or service. In some cases, their interests are also inferred using publicly available information supplied by third parties. Based on this collection of data, users are assigned to different targeting segments and are accordingly served segment-specific ads. Users can be targeted in this way without the advertiser having any information that might personally and directly identify an individual person. (Similar kinds of behavioral targeting have existed in the offline direct mail and telemarketing industries for years, although they generally require indentifying information such as names, mailing addresses and telephone numbers.)

A generic per-computer ad targeting scenario typically works in the following way: A user visits a Web site, and the site places a cookie on the user's computer. A cookie is a tiny text file into which a Web site stores information called a cookieID that it can later use to recognize the user. The cookieID is also recorded in a database at the Web site. Let's assume that the user is visiting the site for the first time and that he has not and will not register at the site or provide the site with any information that could personally and directly identify him. The user is therefore unknown to the site. Each time the user visits that site, the site reads the cookieID and logs his actions on the site. These actions are stored in the database by the Web server and associated with the cookieID. Over time, the cookieID entries in the database might build up a significant record of actions taken by the unknown user on the site.

When sufficient data has been collected, the Web site's business rules might place the cookieID into one or more segments based on the user actions logged in the database. For example, if a user visits the hotel portion of a travel Web site often enough, the cookieID associated with his computer might be placed into a "Hotel Seekers" buying segment. From that point until the business rules dictate differently, the user might be shown hotel ads when he visits that site. Such behavioral targeting has been shown to significantly increase click-through and conversion rates for advertisers.

Clearing the cookie on the user's computer disassociates that computer from the cookieID and the logs of the user's behaviors and segments on the Web site's database. If the user never clears the cookie, the cookie will persist on his computer and the site can continue to accrue information until the cookie's expiration date or until the computer is recycled or the operating system is reinstalled or replaced.

This scenario becomes somewhat more complicated if a computer or computer user account is shared by two or more people. In general, a separate set of cookies is created for each user account (an account with a separate username and password) on a computer. In the case of Microsoft® Windows Vista® or Windows® XP, if all users of a PC share a single user account, the cookies stored on that PC may represent the totality of all their actions on that computer. So, for example, the records that a Web site attributes to a single unknown user might actually represent the actions of an entire family that shares the same computer account or the actions of all users of a public computer. If each user of the PC uses a separate account, each user will have a separate set of cookies.

Third-party ad networks—service providers that provide ads to a number of Web sites—serve targeted ads to computers in a manner similar to that just described. However, they differ in two significant ways. First, ad networks generally have broader reach because they serve ads across a variety of (often unrelated) Web sites rather than on a single site. Second, they may aggregate information about a user's behavior across multiple sites on which they serve ads, so they might capture a broader range of user activity.

Ad Targeting at MSN and Windows Live[1]

As a matter of policy, Microsoft takes steps to separate any information that can be used to personally and directly identify a user—such as name, e-mail address or phone number—from the information in its ad selection system. This de-identification adds an important layer of privacy protection while still allowing Microsoft to serve targeted ads based on user behavior. In other words, the MSN and Windows Live sites do not need to correlate personally and directly identifying data with user behavior online in order to take full advantage of behavioral targeting. For example, MSN can target ads to a person who likes coffee, lives in Seattle and is male without knowing the name, e-mail address or any other personally identifying information that the user might have provided when registering for particular services on MSN or Windows Live.

Microsoft uses three different cookies—the Machine Unique ID (MUID), the Windows Live User ID (LiveID) and the "Anonymous" ID (ANID)—in its ad targeting infrastructure.[2] The latter two are part of the process that segregates data used for ad personalization from information that could personally and directly identify a user. We'll look at each of these in turn.

The Machine Unique ID (MUID)

When a user first visits an MSN or Windows Live site, a standard cookie with a randomly generated unique identifier called the Machine Unique ID (MUID) is placed on the user's computer (the "machine"). For the purpose of ad targeting, that cookieID may behave in the same manner as the cookieID described earlier in the generic example. This means that the MUID may be used to target ads based on the behaviors of an unknown user. This behavior is illustrated in Figure 1. Information that could personally and directly identify a user is not associated with the MUID.

[1] This paper does not cover Microsoft's newly acquired Atlas ad serving technology.

[2] Microsoft sites might set other cookies for other purposes, but they are not relevant to the online advertising topics described here and are therefore not discussed in this paper.

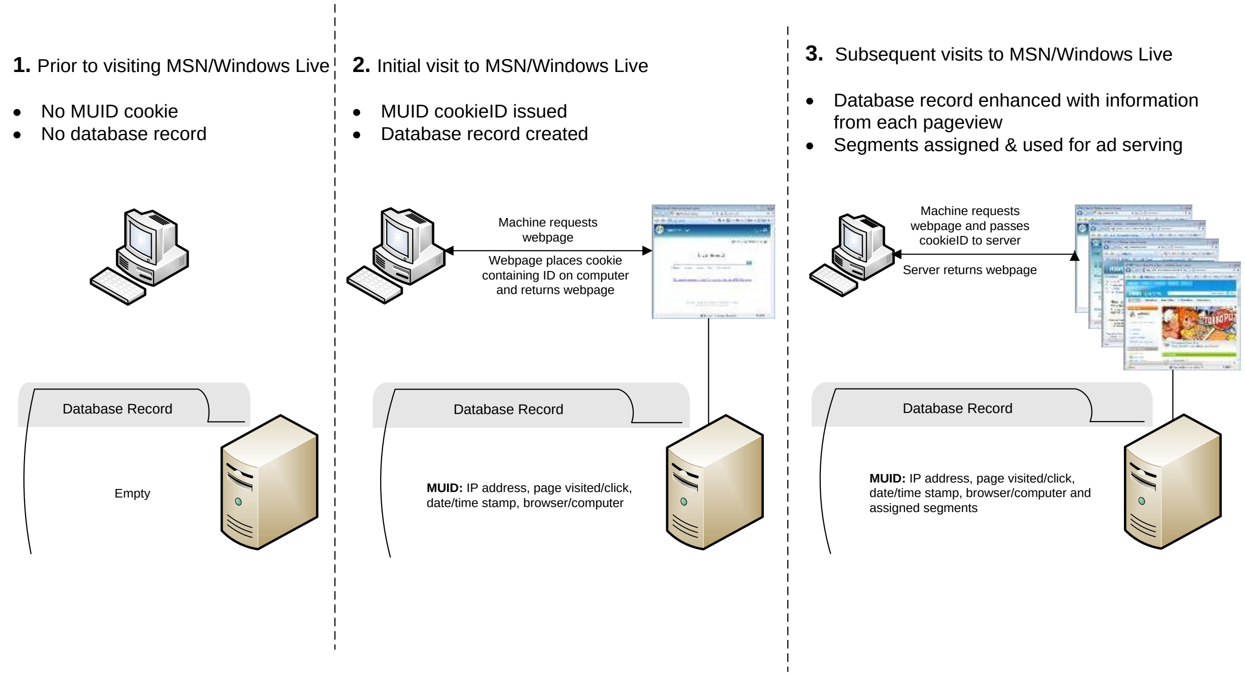
1. Prior to visiting MSN/Windows Live
No MUID cookie
No database record
Database Record
Empty
2. Initial visit to MSN/Windows Live
MUID cookieID issued
Database record created
Machine requests webpage
Webpage places cookie containing ID on computer and returns webpage
Database Record
MUID: IP address, page visited/click, date/time stamp, browser/computer
3. Subsequent visits to MSN/Windows Live
Database record enhanced with information from each pageview
Segments assigned & used for ad serving
Machine requests webpage and passes cookieID to server
Server returns webpage
Database Record
MUID: IP address, page visited/click, date/time stamp, browser/computer and assigned segments

Figure 1. The Machine Unique ID (MUID).

The Windows Live User ID (LiveID)

The example of the MUID involves a cookieID that is assigned per computer account to users who are not known to the Web site. Now we'll discuss cookieIDs that are assigned on a "per-login" basis to users who have established a relationship with the Web site.

In general, Web sites that require a user to log in to access a user-specific service, such as Web-based e-mail, use a user-based cookieID as a part of their system for granting access. At MSN and Windows Live, the core user-based ID is the Windows Live ID (LiveID). When a user first registers at the site, she typically chooses a username and password and provides Microsoft with a first and last name, plus a few pieces of non-identifying demographic information such as country, Zip code, age, gender and language. In scenarios where a user is creating a billing account, additional pieces of personal information might also be collected at this point. A unique LiveID is then generated and associated with this data. The LiveID is the unique ID number specific to that user account.

The LiveID is stored on the Windows Live ID servers and, once the user has presented a valid username and password, is placed in a cookie on her computer. The presence of the LiveID cookie is the signal to an MSN or Windows Live service that it should continue to grant access—for example, to the user's e-mail in the case of Windows Live Hotmail. When the user logs out of the service or ends the session, the LiveID cookie expires (unless the user has opted to make the cookie permanent by clicking “Save My Password” so she does not have to log in each time she accesses the service). Granting a user access to her Hotmail e-mail is an example of when the LiveID needs to be associated with personal information. Other Windows Live services that require this type of authentication via a LiveID include Windows Live Messenger and Windows Live Spaces.

Because the LiveID database contains data that could be used to personally and directly identify individual users, by design Microsoft’s advertising system *does not* use the LiveID to select and serve ads—even though it would be technically far simpler to have it do so. One of Microsoft’s online advertising principles is that its ad targeting platform can select appropriate ads based *only* on data that does not personally and directly identify individual users.[3]

The “Anonymous” ID (ANID)

One of Microsoft’s goals is to serve targeted ads in a manner that protects user privacy. To avoid using the LiveID cookie to serve per-user ads—because, as described earlier, it is directly associated with information that could personally identify the user—Microsoft has created an “Anonymous” ID, called the ANID, on which its ad serving capabilities are based.

When a user first registers with Windows Live or MSN, a LiveID and an ANID are created simultaneously. The ANID is derived by applying a one-way cryptographic hash function to the LiveID. A one-way cryptographic hash function ensures that there is no practical way of deriving the original value from the resulting hash value—that is, the process cannot be reversed to obtain the original number.

What this means in practical terms is that each time a registered user logs in, Microsoft’s system applies the hash function to the LiveID to generate an ANID, and each ID is put in a separate cookie on the computer. The advantage of using a one-way cryptographic hash function is that although the same number is guaranteed to be generated each time it is applied to a given LiveID, it is virtually impossible to reverse the process. In other words, it is extremely difficult to use a given ANID (with or without knowing the hashing algorithm) to derive the original LiveID value. Because all personally and directly identifying information about a user is stored on servers in association with a LiveID rather than an ANID, there is no practical way to link data stored in association with an ANID back to any data on Microsoft

[3] Microsoft’s online advertising principles can be found at http://download.microsoft.com/download/3/7/f/37f14671-ddee-499b-a794-077b3673f186/Microsoft’s%20Privacy%20Principles%20for%20Live%20Search%20and%20Online%20Ad%20Targeting.doc.

servers that could personally and directly identify an individual user. Figure 2 illustrates this relationship between the two IDs.

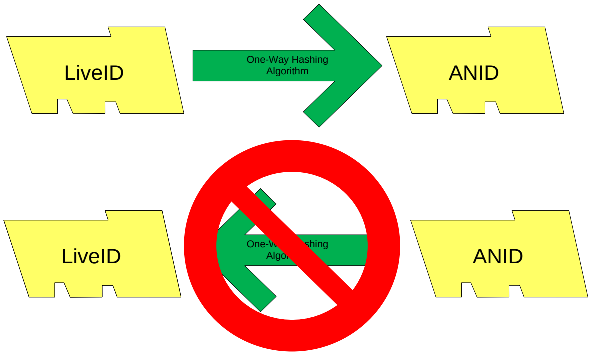

Figure 2. One-way cryptographic hash.

As mentioned earlier, a user might input particular pieces of demographic information when a LiveID is created. When the LiveID and ANID are created, the demographic information that cannot be used to personally and directly identify the user is copied to a database that is indexed on only the ANID. Microsoft's ad serving infrastructure consumes data associated with the ANID but not the LiveID, so copying the demographic data in this way allows Microsoft to make it available to the ad serving infrastructure. As a user with an ANID cookie on her computer navigates around the Microsoft sites, data associated with her online behaviors, such as searches and pageviews, is associated with the ANID. All of this information can then be used to assign ad targeting segments to the ANID in the same manner as described previously in the generic description of ad targeting. (Figure 3 illustrates this process.) Most importantly, because of the one-way hash used in creating the ANID, none of the specific behaviors associated with the ANID or the ad targeting segments consequently assigned to the ANID are linked back to the personal information associated with the LiveID.

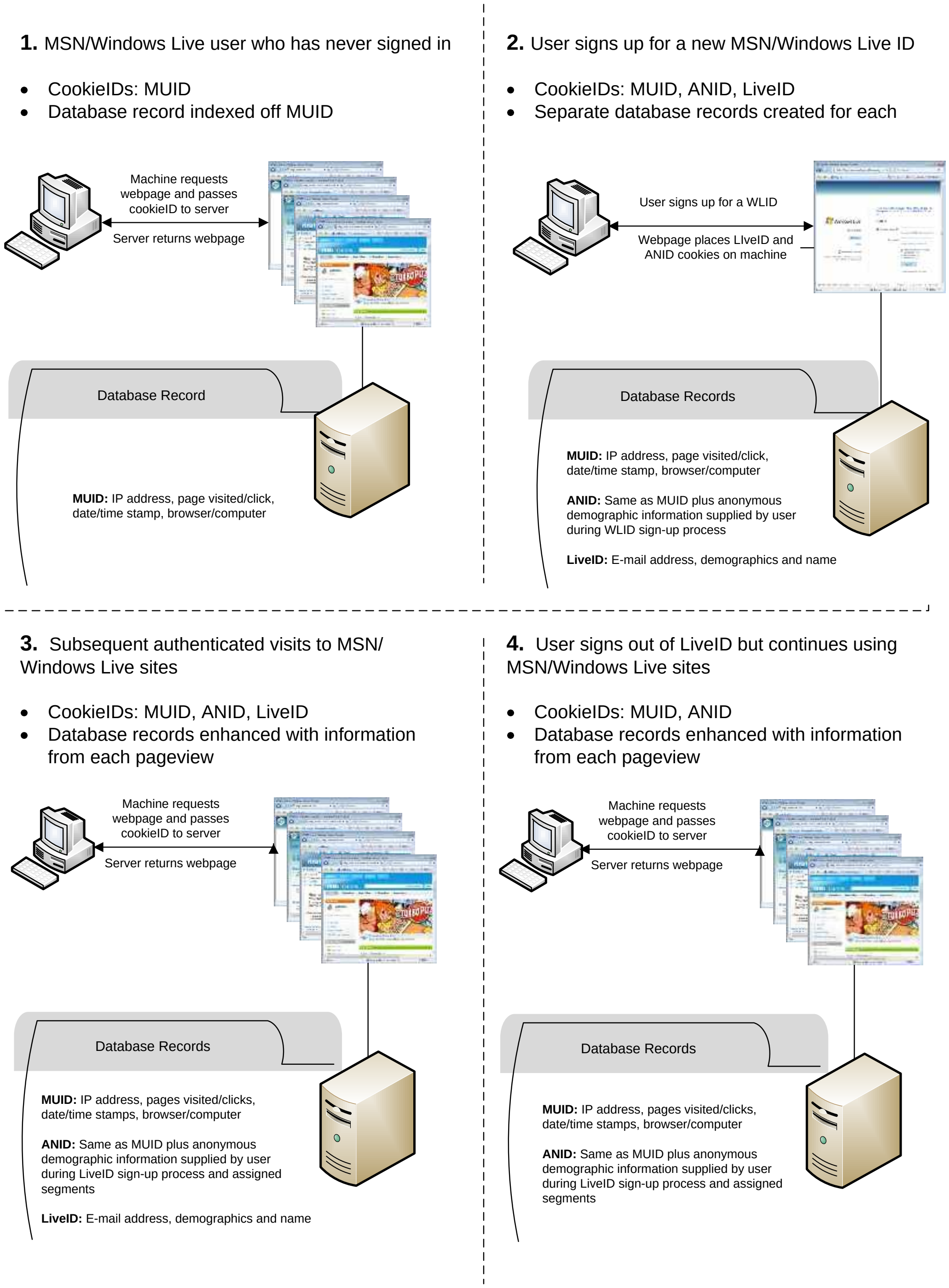
1. MSN/Windows Live user who has never signed in
CookieIDs: MUID
Database record indexed off MUID
Machine requests webpage and passes cookieID to server
Server returns webpage
Database Record
MUID: IP address, page visited/click, date/time stamp, browser/computer
2. User signs up for a new MSN/Windows Live ID
CookieIDs: MUID, ANID, LiveID
Separate database records created for each
User signs up for a WLID
Webpage places LIveID and ANID cookies on machine
Database Records
MUID: IP address, page visited/click, date/time stamp, browser/computer
ANID: Same as MUID plus anonymous demographic information supplied by user during WLID sign-up process
LiveID: E-mail address, demographics and name
3. Subsequent authenticated visits to MSN/ Windows Live sites
CookieIDs: MUID, ANID, LiveID
Database records enhanced with information from each pageview
Machine requests webpage and passes cookieID to server
Server returns webpage
Database Records
MUID: IP address, pages visited/clicks, date/time stamps, browser/computer
ANID: Same as MUID plus anonymous demographic information supplied by user during LiveID sign-up process and assigned segments
LiveID: E-mail address, demographics and name
4. User signs out of LiveID but continues using MSN/Windows Live sites
CookieIDs: MUID, ANID
Database records enhanced with information from each pageview
Machine requests webpage and passes cookieID to server
Server returns webpage
Database Records
MUID: IP address, pages visited/clicks, date/time stamps, browser/computer
ANID: Same as MUID plus anonymous demographic information supplied by user during LiveID sign-up process and assigned segments

Figure 3. Sample MUID, ANID and LiveID interactions.

When a user logs out of a Windows Live account, the LiveID cookie is deleted from her computer. However, the ANID cookie remains on the user's system until a different Windows Live account is accessed from that computer account (which would replace the old ANID cookie with a different one), until the user takes steps to delete the ANID cookie or until the cookie expires.

Privacy protections tend to be strongest when implemented as a part of the fundamental architecture of a computer system. Microsoft’s ad serving system was designed expressly to work with the ANID, and the ANID was designed expressly to enhance user privacy. These safeguards help ensure that information associated with a LiveID will not leak into the ad serving environment.

Of course, the ANID infrastructure itself does not guarantee complete and irreversible anonymity. But it does provide strong technical protection, which, combined with stringent internal policies, is designed to keep the data used for ad serving separated from information that identifies an individual.

For example, because the system is ANID-based, Microsoft employees with access to the company’s ad serving system alone cannot identify users who are served ads based on the data in the system. Furthermore, to associate any of the ANID-based data in the Microsoft ad system with an individual user, an internal or external attacker would not only need access to the ad serving system (to access the data), the Windows Live ID system (to access all LiveIDs ever issued) and the hashing algorithm but would also need a massive computing infrastructure to run the algorithm on each and every LiveID ever created to try to find the ANID in question. Each of these components is separately protected with strong internal security measures, rendering this scenario virtually impossible.

Further, the use of the ANID is part of the company’s overall approach to protecting user privacy, which includes strong and meaningful protections from the time that behavioral data is first collected. These protections also include the recently announced policy of anonymizing search query data after 18 months. (This includes the complete and irreversible deletion of full IP addresses and cookieIDs—including ANIDs—from search terms.)

Conclusion

Microsoft’s use of the ANID enables the delivery of relevant ads to users while basing ad selection solely on data that does not personally and directly identify individual users. As a fundamental element of Microsoft’s ad targeting infrastructure, the ANID underscores the company’s strong commitment to privacy. It is complemented by the recent announcement of

Microsoft's Privacy Principles for Live Search and Online Ad Targeting,[4] the public release of the company's Privacy Guidelines for Developing Software Products and Services[5] and its advocacy for comprehensive federal privacy legislation in the United States and strong public policies worldwide to protect consumer privacy. In a dynamic industry where rules and best practices are continually evolving, Microsoft is committed to ensuring that its current and future products and services implement industry-leading technologies and processes that protect individual privacy.

[4] Available at http://download.microsoft.com/download/3/7/f/37f14671-ddee-499b-a794-077b3673f186/Microsoft's Privacy Principles for Live Search and Online Ad Targeting.doc.

[5] Available at http://www.microsoft.com/downloads/details.aspx?FamilyId=C48CF80F-6E87-48F5-83EC-A18D1AD2FC1F&displaylang=en.