DE-DUPLICATION TECHNOLOGIES FOR MARKETING RESEARCH
A COMPREHENSIVE REVIEW
AND
AN INTRODUCTION TO A NEXT GENERATION SOLUTION

By Jason Freeman

COO SHC Universal

You would be surprised if you counted how many internet connected devices you have in your home. I certainly was.

A few months ago, I was commuting to work and heard some recent survey findings on household Internet connected devices. Having worked in Market Research for over 20 years, my curiosity was immediately piqued. I was surprised to hear that, on average, there are 7.8 connected Internet devices per household in the U.S. I quickly started to count our household devices. We had an iPad Mini, a desktop computer, two laptops, two iPhones, and two iPods. Even if I didn't count the iPad Mini, that was currently "out of order" because my four-year-old used it to watch a movie in the bathtub, my family and I were well on our way to being average. Being average now means we have the choice and ability to access Internet content through of number of devices — 7.8 of them, to be exact!!

This ubiquitous access to Internet connected devices poses a real challenge to market research when it comes to survey security and deduplication. This is a challenge, I think it is fair to say, that most companies do not fully understand. Assuring and certifying respondents' answers are not duplicated is a fundamental principle of market research data collection. SHC Universal is working diligently to mitigate the risk and inaccuracy of duplicated responses in data collection, and to understand the whole scope of the problem at hand. To provide an overview of the challenges and opportunities, I have provided insight to question 3 of the European Society for Opinion and Marketing Research's (ESOMAR) "28 Questions to Help Buyers of Online Samples." ESOMAR is at the forefront of international social, market and opinion research, and has designed a standard set of questions a buyer can ask to determine whether a sample provider's practices and samples fit with their research objectives.

How do you deal with the possibility of duplication of respondent answers across sources?

High "survey N" sizes, low qualification incidences, short fielding timelines, and niche markets are driving the need for multi-faceted fielding tactics. In many cases, the delivery approach may require multiple online panel agencies to be in the field concurrently. Employing a robust deduplication solution will ensure data integrity and validity, reduce costs, and eliminate uncomfortable client discussions (e.g. "These two cases have really similar responses. Are they the same person?"). Inevitably, if a comprehensive solution for deduplication is not put in place, the proverbial "can of worms" will be opened, an oversight committee could be appointed, the quality of your data questioned, and future ongoing scrutiny of your deduplication standards will ensue. This could have longstanding implications for your reputation, and therefore, the health of your business.

Let's now look at some of the more prevalent deduplication technologies in use today and see where they fall short.

IP Deduplication

"IP" or Internet Protocol, also commonly referred to as an Internet address, provides a set of rules that govern Internet activity and facilitate online communication. In the same way that someone needs your mailing address to send you a letter, or your phone number to give you a call, a remote computer needs your IP address

to communicate with your computer.

IP Deduplication technologies function by cataloguing the IP addresses of the computers asking to enter a survey. Similar to Caller ID, the calling party must be identified and present their IP address. Once obtained, the receiving end (the survey webserver) will send a response, and a communication channel (traffic) occurs. As new "callers" request entry into the survey, the IP Deduplication technology checks all prior call history to ensure new participants are coming from a unique IP address. If a participant tries to enter with the same IP, they are considered a duplicate, and refused entry.

On the surface, this appears to be a great solution. However, there are several disadvantages to IP deduplication technologies.

It was nice while it lasted.

IPs addresses are not always static - an IP address can change. Typically, home Internet connections are dynamically assigned by an Internet Service Provider (ISP). There is no guarantee the home user will have the same IP address all the time. In fact, may ISP providers will re-issue a new IP address with something as simple as turning off your modem or router.

Knock, Knock... who's there?

Multiple devices and users share the same IP address. This is more commonly referred to as a "Public IP." Potential survey participants located behind the same firewall or router will communicate over the same Public IP. Say we have a hospital campus with the address "21 Jump Street, Beverly Hills, 90210" – there are potentially several willing participants ready to contribute to your research, but IP Deduplication technologies will bar them from entry. In many cases, this degree of exclusion may be undesirable.

You can't take it with you.

If you travel, your home or work IP doesn't travel with you. You'd be using another network to connect to the Internet. As you move from the airport to your hotel to the local coffee house, your IP address will change each and every time. You could even be at the office right now, on your cell phone – over a cellular network, and on your desktop computer – connected to your employer's network. They would have different IP addresses. Survey participants likely have an abundant number of potential network connections leading to active or passive security bypass breaches.

We are only as blind as we want to be.

IP Deduplication in and of itself is not a bad technology. However, having a survey with a unique IP address does not guarantee a duplicate-free project. Conversely, a survey with duplicated IP addresses does not necessarily signify duplication issues.

"Cookie" Drop Deduplication*

Cookies, as we all know, are tasty, little morsels. I prefer the *hot* chocolate chip kind with a tall glass of cold milk. There are also cookies in the computer world, but we don't get to eat them. Cookies are very small text files stored in a computer's browser directory or program data subfolders. Cookies are used by websites for lots

of practical reasons, such as keeping track of movement within a site, helping resume a session, remembering logins and preferences, etc. In short, they are used to make online experiences go as smoothly as possible.

Cookie Deduplication Technologies function by requesting information that can be retrieved at a later time. The survey website server can then recognize a user when he/she returns to the website. Only the server that sent the cookie can read, and therefore use, that cookie. As new participants enter a survey, their browsers can be evaluated, and the server will determine if they have previously entered and allow or reject entry to the survey instrument.

This sounds like a great solution. However, there are several disadvantages to Cookie Drop Deduplication Technologies.

Who doesn't like cookies?

Many Internet users do not like cookies. Some inadvertently associate them with pop-up ads or viruses. Others simply do not want to be identified or tracked by websites. Today, many popular browsers conveniently allow participants to go "Incognito" or allow "private" browser settings. These browser modes will block websites from placing cookies, and subsequently disable Cookie Deduplication Technologies.

Who emptied the cookie jar?

Most popular browsers will let you decide whether you keep or delete cookies after each online session. At any time, a participant may choose to clear his/her browser history and remove all cookies. In addition, other software programs will perform regular "clean-ups" or "health checks," purging cookies

Baking cookies. How many do you need?

Growing up in a house of eight brothers and sisters has taught me it takes lots of cookies to feed an army. Luckily, my mom had two ovens so she always had a choice about which one she used. Similarly, many participants today have a choice of multiple browsers (Chrome, Firefox, IE, Opera, etc.) on their device(s). Each browser (or oven) has its own cookie repository. A cookie set in one browser is not physically found in another. Similar to my siblings, browsers do not like to share their cookies. In other words, cookies are not cross-browser compatible. To add to the complication, participants can switch and use one of their other 7.8 Internet connected devices. Survey participants have an abundant number of potential browsers and devices which can lead to active or passive security bypass breaches.

* Local Storage, Flash cookies and ETags are other similar techniques, but are more obscure and harder for a user to detect. However, they all have the same fundamental weakness as cookies – they rely on things the user can delete.

Browser Fingerprint / Digital Fingerprint Deduplication

Browsers do not have a unique identification. If there were such a thing, it would be the Holy Grail for tracking, as well as an online advertising company's dream! But, browser fingerprint technology comes close.

A browser fingerprint, or device fingerprint, is information collected about a computing device for the purpose of identification. Comparable to a human fingerprint, a computing device has several forensic values that can be studied to determine uniqueness. These points of identification include items like:

- The user agent string from each browser.

- The HTTP ACCEPT headers sent by the browser.
- Screen resolution and color depth.
- The system's time zone.
- The browser extensions and plugins, like QuickTime, Flash, Java, or Acrobat, which are installed in the browser, and the different versions of those plugins.
- The fonts installed on the computer, as reported by Flash or Java.
- Whether your browser executes JavaScript scripts.
- "Yes/No" Information identifying whether the browser accepts various kinds of cookies.
- A hash of the image generated by Canvas fingerprinting.
- A hash of the image generated by WebGL fingerprinting.
- Whether your browser is sending the "Do Not Track" header.
- Your system platform (e.g. Win32, Linux x86).
- Your system language (e.g. en-US).
- Your browser's touchscreen support.

The standouts for uniqueness are clearly user agent (1 in 86,939*) and browser plugins (1 in 1.8 million). The cumulative uniqueness just for user agent and Plugins: 86,939 x 1.8 million = 156.490 billion. Add in the others, and you have an estimated 18.1 bits of entropy, which means if you take two browsers at random, you have one chance in $2^{18.1}$ (≈280,000) that they will have the same "fingerprints." That's pretty unique, right?

Am I really unique?

Unfortunately, browser-fingerprinting isn't fool-proof. It mistakenly assumes points of identification distributed among users are random, when in fact, mainstream devices fall into a more homogenous ecosystem. Usage on smartphones, tablets and iDevices tends to fall under a much smaller subset of forensic values. For example, the number of iPhone and iPad versions and models are not diverse and will more often incidentally fingerprint identically between two unique survey participants. Moreover, some browsers are more popular than others. Findings in a recent internal review discovered nearly 50% of survey participants were using the same user agent (browser) within the top 10,000 user agents (browsers) available.

Another environment to consider is the corporate network. Here, you are likely to find many dozens or hundreds of potential participants with the exact same browser, plugins, fonts etc., leading to false-positive results.

Talk to the hand.

Similar to cookies, participants may prefer to remain anonymous and choose to limit their browsing footprints. Comparable to "privacy" or "incognito" modes for cookies, many popular browsers now have add-ons (Privacy Badger, Disconnect, Tor, I2P, Tails, Freenet, Freepto, etc.) that purposely modify and/or block Browser Fingerprinting Technologies. Changes to the browser extensions or plugins, browser version updates, fonts installed on a computer, etc., can meaningfully impact detection results. In addition, JavaScript must be enabled for fingerprinting to work. Although it is not realistic to expect users to have JavaScript disabled on the modern web, it is still a user preference that can be turned off by those who may be security conscious.

(To be fair, browser fingerprinting isn't automatically a bad thing. In some cases, it offers potential benefits to end-users. For example, my bank can detect if the device logging into my online account is different from a device or browser that I have previously used. Based on that observation, my bank will present me with additional security validation and send me a passcode to my primary phone that I must enter before access is granted.)

Cross Browser Detection

Until 2017, researchers had been unable to develop reliable techniques to track users when they use different

browsers on the same device. This new technology is still being reviewed and has not yet been released for commercial usage. In the meantime, traditional browser fingerprinting remains a single browser check. This limitation makes it impossible to link the fingerprint left behind by a Firefox browser to the fingerprint from a Chrome or Edge browser running on the same device. Browser fingerprinting does not resolve the reasonable assumption participants can move to one of their other 7.8 Internet connected devices to gain survey entry.

* The total number of user agents is always evolving. This is a current estimate as of January 2017.

Exclusions Lists

An EL (Exclusion List), also commonly known as DNR (Do Not Recruit), is the exchange of panelist information between panel recruitment providers. Basic member information of the first three characters of the first and last name, along with the zip or state code, are exchanged between providers to avoid the recruitment of the same individual.

What is your name?

If you were to ask my Dad his name, he would tell you, "Bob." But, his real name is Richard. He prefers to go by his nickname. This isn't unusual. There are many examples where someone may choose to use different names across time or in different situations. Consider self-identifying by your middle name, a change in marital status, moving to a new state/zip, or a variation between legal or common name on file. These seemingly static points of data are actually highly variable. This inconsistency can cause EL and DNR lists shared with panel providers to be inaccurate and result in an inability to catch all duplicates. Additionally, simultaneous fielding of more than one panel provider will inherently have overlap, which cannot be fully controlled, even under the most frequent and diligent sharing between panel providers.

Where is the sourcing?

In general, panel providers work well by sharing basic panel information for deduplication purposes. It is viewed as a necessity, but if not carefully managed, could lead to exposure of panel members. And while this works well in the US, there can be more strict privacy guidelines overseas. Some recruiting agencies will simply refuse to share information or deem the release in violation of their privacy policies. In some situations, prior wave exclusion information may not be available or limited to just an IP address.

Didn't you previously attempt?

Due to these privacy concerns, most panel providers will only share suppression lists for completed cases. Therefore, not all of those previously disqualified will be excluded. This scenario has the potential to allow "re-attempted" entries into the same survey a second time, which may be undesirable. Healthcare or niche panels tend to have strong overlaps, whereas universe sizes are small or have a limited scope.

Other Considerations

IP address, cookies, digital fingerprints and exclusion lists are the primary means employed in deduplication processes. However, a modern deduplication solution should provide protection for other scenarios not considered by current methods.

Testing Bypass

A common technique in the industry is to ensure those invited to a survey are correctly re-routed and credited. In this case, security is often temporarily removed or bypassed with a testing link or added URL pass parameter. In other words, there is a security "on/off switch" that the panel provider or survey host system can "flip." An ideal security solution will have a way to eliminate an accidental error in security bypass.

Resetting IDs – Allowing disqualifies to attempt again

In small universe groups, or low qualification studies, it may become necessary after live fielding begins to relax certain survey qualification points. In these conditions, there is a very desirable need to have a simple mechanism to allow past participation re-entry without the requirement to modify, script or turn off security. A flexible security solution will have this ability.

Blacklisting/Whitelisting

Customized deduplication solutions will facilitate the concept of blacklisting and whitelisting of IP addresses, cookies, and digital profiles. Examples of these cases would be: suppression of prior wave IP addresses, the blacklisting of a digital or IP profile that you never want allowed in any survey, or whitelisting certain recruiting methodologies (i.e. Face to Face or a Call Center), where unchecked entry is allowed while security is still active for other survey hits.

Flexibility

Depending on the case, it may be desirable to make adjustments for more or less selective detection. For example, a consumer-based survey may always require the IP address to be unique, but a healthcare or B2B survey should be flexible to allow the same IP (i.e. unique respondents at the same location) to participate.

Fraud Prevention

A robust deduplication solution will prevent fraud. Participants who intentionally try to obscure or hide their identity should be disallowed. Inherently, detection of proxy, TOR network, masked user agent, device emulator, and datacenter/VPN should be viewed as anonymizing behavior and be key areas of defense.

Filling the gaps

A deduplication solution should employ all reasonable techniques. No single technique in and of itself provides fail-safe technology. Therefore, a modern, innovative enterprise system should employ all current methods together. It should go beyond the scope of boxed solutions and combine these tools in a way that enhances the digital and physical tracking of entries beyond a single survey entry and across concurrent surveys. In other words, it is best to develop a systematic way of knowing all the devices, IP addresses, and digital fingerprints of any single survey participant.

Coming Soon - SHC AuthenticID, SHC's Answer to Deduplication

At SHC Universal, we are committed to quality and Perfect Data. Our years of experience have taught us what it

takes to have a robust deduplication solution. We believe both that a true solution needs to solve all of the challenges inherent with current deduplication tools, and further that we have developed the best answer to this problem– we call it SHC AuthenticID – and we plan to begin offering it to clients imminently. Soon, the days of worrying about duplication will be over. We plan to launch SHC AuthenticID shortly – stay tuned, we have your solution ready!

Please contact Jason at Jason.Freeman@SHCuniversal.com or our Client Relationships team at sales@SHCuniversal.com for more information on our SHC AuthenticID product.

www.SHCuniversal.com