DOTMETRICS

DAILY FRESH AUDIENCE DATA

© 2016 Ipsos. All rights reserved. Contains Ipsos' Confidential and Proprietary information and may not be disclosed or reproduced without the prior written consent of Ipsos.

DotMetrics Audience: Daily Unique Users Using Predictive Modelling

Legal notice

All information contained in this document is the legal property of Ipsos and DotMetrics. This document is intended for internal use by clients and other stakeholders using the DotMetrics system, however, it cannot be shared, disclosed or reproduced, neither in written or oral form, with third-parties, before obtaining a written consent from Ipsos.

Introduction

DotMetrics Audience is a joint big-data predictive analytics solution developed by Ipsos, and used to deliver daily ratings and demographic data for web publishers. The system is based on the DotMetrics infrastructure provided by the web-analytics company DotMetrics, and predictive modelling and research data provided by Ipsos.

This document is conceived as a light-weight whitepaper targeted at industry professionals, marketers, media planners and web analysts, who don't necessarily have research or IT experience, to help them better understand the foundation of the DotMetrics system, and the background behind the numbers presented in it.

DotMetrics is a multi-layered web-analytics and audience measurement solution that can be delivered in three variants:

1. A site-centric predictive model using pop-up, or other available device-level demographic data to predict probable user metrics
2. A user-centric panel based solution, utilising software meters to accurately capture browsing data of a representative subset (panel) of the population.
3. A hybrid solution, utilizing site-centric data as a base, with user-centric data used as a corrective input to the predictive model

This document is first in a series of documents, focusing on the site-centric implementation of the predictive model, using pop-up surveys provided by the DotMetrics system to gather device-level demographic data.

The site-centric system is built on three-layers:

1. The analytics layer gathering census-level device data on all browsing performed on measured websites
2. The pop-up layer gathering devices into a weighted device-level panel with ascribed demographics for calculation of demographic data
3. The predictive model utilising establishment survey data and the other two layers to provide unique user calculation

DotMetrics Web Analytics

The DotMetrics web analytics layer is based on data collected by site-centric Javascript[1] tags, which the web publishers participating in the measurement install on their own webpages, embedding the tags in their website's HTML code.

The tags are implemented in order to gather all traffic-level data that can be detected (number of page loads, session's opened, session durations and active time spent at the page), as well as to try to identify the device accessing the website, as the goal is to measure the number of different devices that have accessed a website in a certain time period.

In order to be able to identify devices, DotMetrics uses three layers of device identification. The devices are first identified using HTTP cookies[2], and in case that method fails, they are identified using other established types of locally stored identification data (Flash local storage, e-Tags, perma-cookies etc.). In case the device can't be identified using previously delivered and stored data, DotMetrics implements a digital fingerprinting algorithm[3] to try to match the device to a previously recorded fingerprint produced by a device.

The tags are then executed each time a web page is loaded on a user computer. The tag immediately collects user-agent[4] data, and classifies the device into one of four device categories (PC, Mobile, Tablet, Smart TV) based on user-agent data. A record of the device's public IP is temporarily made, until the IP can be geolocated using geographical IP databases, at which point, a country of origin is attributed to the traffic data. Afterwards, the IP is no longer stored in connection to other data, but is stored separately for click-fraud tracking – so that the IP can be banned in case of determined fraud.

After the user-agent is recorded, the tag stores information about the page load, recording the page view and timestamp and starts to record active time spent[5] on the website. Finally, the device enters the identification funnel, and tries to be identified using the methods described above (shown in figure 1). In case no identification can be made, the device is treated as a newly arrived device, its fingerprint is recorded and a device record is created in the database. The newly created device record receives a new set of cookies to be easily identified on subsequent visits.

All data is attributed to a specific website or set of websites based on website ID's which are present in each tag script. All ID's have domains attributed to them, and in case a script tag was executed under a wrong domain, the traffic is discarded.

[1] Javascript is a technology utilised by web browsers, which allows for execution of short scripts on the client-side (i. e. on your computer).
[2] HTTP cookies are small textual files that websites deliver to a client computer, in order to be able to recognize the computer at a later date.
[3] Digital fingerprinting is a technology that tries to passively create a unique identifier of a device without delivering any data to the computer (such as in the case of cookies). The data is gathered from a list of publically available data about the device, that browser need in any case – to be able to render the page correctly, and monitor for server communication. For a review of articles on digital fingerprinting, refer to the references.
[4] User-agent is a piece of information that each browser reports to the server as soon as connection has been established. It holds information on OS version, device type, and other data about the device.
[5] Active time spent is a metric implemented by DotMetrics, which will be explained at a later section.

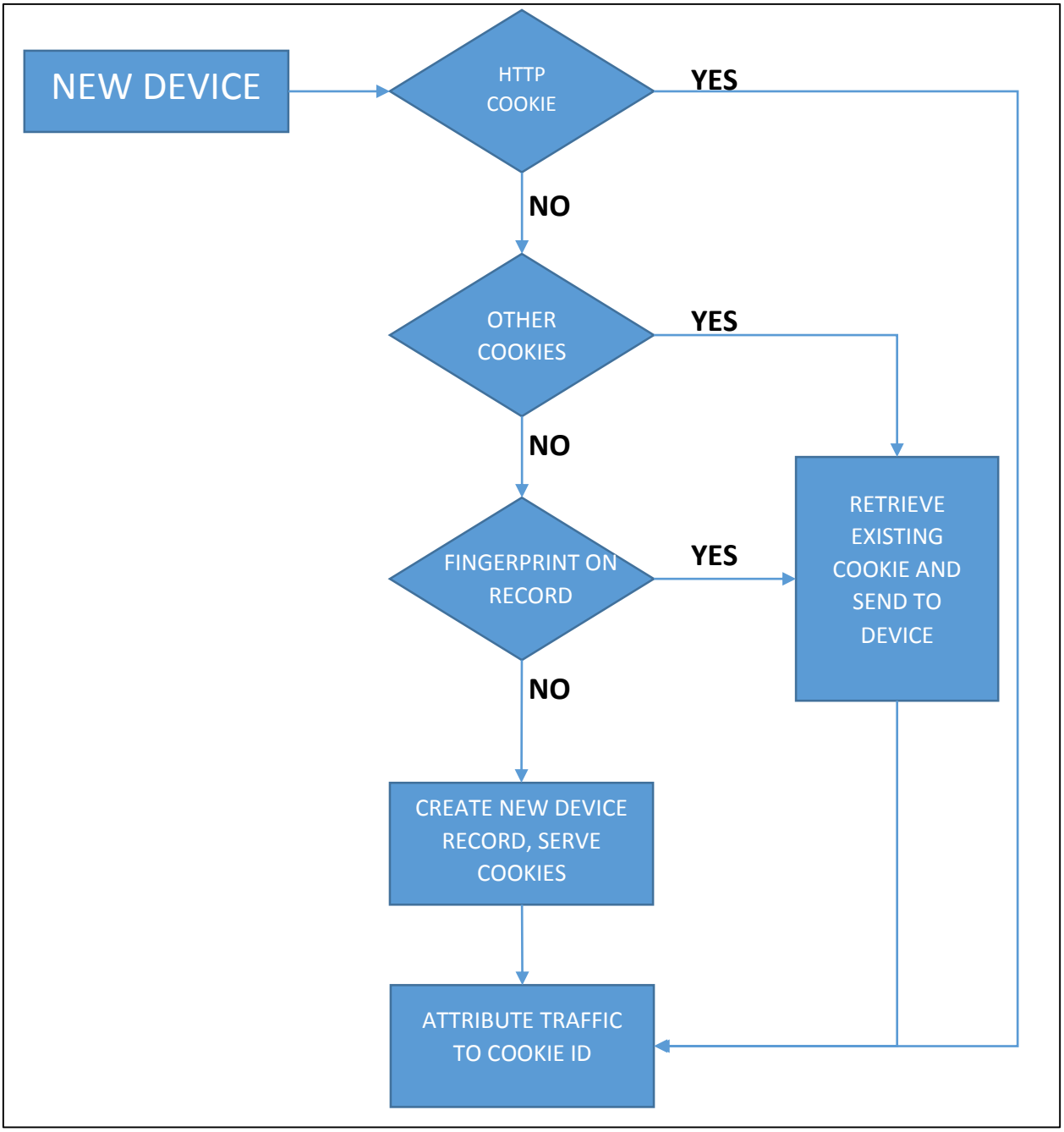

Figure 1: Device identification process

In cases where no identification can be made[6], all traffic is recorded, but the device doesn't go through the device identification funnel. Instead, all traffic is attributed to a set of unverified devices, for which the number of device remains unknown. The number of devices in the unverified section of measurement is calculated by projecting the number of devices in verified traffic (i.e. if the verified traffic accounts for 85% of page views, then we assume that the remaining devices form 15% of the total device count).

[6] For instance, because of private browsing options, or using secure browser such as the TOR browser.

The data is stored in an intermediate database in a heavily unstructured format, akin to a log, until 02:00, when the database is structured and stored in the final big-data warehouse, which is used as the data production website in DotMetrics Audience.

Additional layers of data quality assurance are implemented on top of this process, so each device is constantly checked against its previous browsing. In cases of conflict[7], where the same device ID is seen performing mutually exclusive traffic, the devices are split and tracked separately.

Web Analytics Metrics

1. Page view

A page view is recorded upon each page load. A page view represents the number of HTML page loads that were performed on the website[8].

A page view is the simplest web metric, which is tied to a per-request way of serving ads, so each page view means a new add could've been loaded on the website.

2. Visits

Visits in DotMetrics are a way of tracking the number of different user sessions started on the website. A visit is started each time a device performs a page load, unless it already has a running visit. A visit is closed when a device closes a browser it was browsing in, or when 30 minutes pass that a device hasn't made a new page load in the system.

3. Time spent

Time spent represent the active time spent browsing the website. Since this is a cumulative metric, it shows the cumulative time of all users who have accessed the website in a unit of time. Time spent is recorded when a page load starts. In case a user switches out of the tab (the tab stops being the active window), time spent measurement pauses. When the user switches back to the tab, time measurement continues.

In cases where the device was idle for more than 30 minutes, the idle time is subtracted from the total time measured on that website. Idleness is defined as a period without any mouse, keyboard, video-watching or scrolling activity.

DotMetrics Pop-up Surveying

Along with traffic measurement, the tagging scripts also serve as delivery tools for the DotMetrics Audience pop-up survey. The pop-up survey is designed to pop-up on a set of devices at a daily level. Users who receive the pop-up have the option to fill out the pop-up survey, in which case, it won't be shown to that user again while we are able to track the user, or until the survey's been valid for a certain period (determined for each market differently). After the set period passes, the device can be surveyed again to account for changes in demographics. If the user decides not to fill out the pop-up survey, it won't target him again for a shorter period (two weeks standard).

[7] Although rare, occasionally two devices can produce a similar fingerprint (such is the case in iOS devices), and be falsely identified as the same device. By applying additional layers of data quality, we are able to correctly identify those devices.

[8] Depending on the market requirements, page views can also include the number of AJAX request for asynchronous websites. To be sure what is represented, check with your local DotMetrics supplier or joint industry committee body.

The answers from the pop-up surveys are then attributed to the device on which it popped out. To account for device sharing, we ask a question "Does anybody else use this device?" as part of the survey on PC's and tablet devices, and use the information in the weighting of the device afterwards.

After the demography from the survey has been attributed to a device, that device becomes part of the sample for all traffic performed on websites which it has visited[9]. Since the method of delivery and surveying is not random in nature (we choose devices randomly, however, we cannot speak of a truly probabilistic sample with response rates as low as they are in pop-up surveying), we apply weights from population characteristics determined through establishment surveys. The pop-up survey pools are thus weighted to represent the Internet population of a certain country.

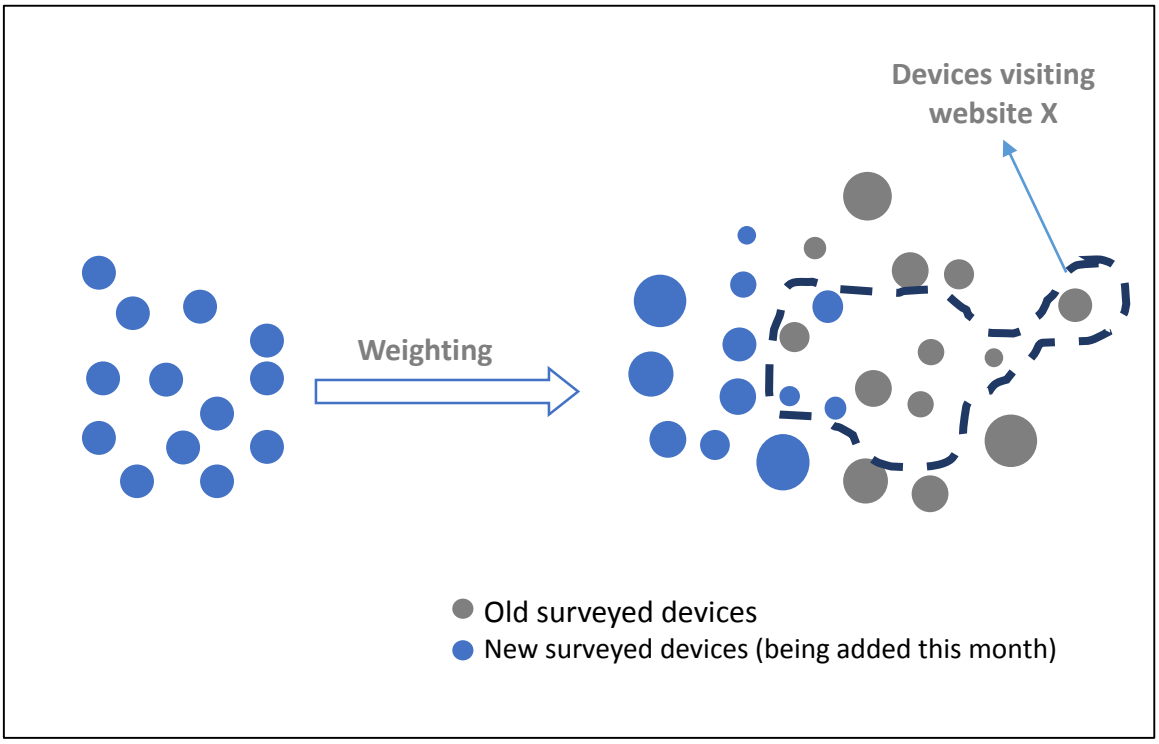

Figure 2: Demographic data sampling

It is worth to note, that the pop-ups are in no way tied to the website on which they were shown. Once a device has been surveyed, it is free to visit other websites – and since we can track it, we will attribute its demographics to that website traffic as well. In that way, the pop-up surveyed devices become parts of a surveyed "pool", akin to a panel[10]. Figure 2 illustrates this point.

Finally, it's worth noting that all demographic calculations have a minimum sample of 80 users.

[9] In other words, when looking at the demographics of website X, if the surveyed device has visited website X, it's answers will be used to calculate demography for that website.
[10] This not being a true panel, in the sense it is usually used in audience research, because it's not maintained.

Establishment survey data

In addition to site-centric measurement provided by the tag scripts, an establishment survey is also performed. Using contemporary surveying methods of collection[11], we establish certain key Internet usage statistics present in the population.

The data we are interested in entails – how many Internet users are there on a certain market? How many of those access domestically available websites in general? What devices are they using to access those websites? Are there differences in device usage between certain demographic groups, and if so, what are the demographic variables which are most significant for this difference? What is the device overlap between device groups? How many mobile users are also PC users, on a daily level?

Using the establishment survey, we answer the above stated questions, and construct a clear picture of the size, structure, and habits of the Internet-visiting population on a given market. The establishment survey is usually performed on a quarterly basis, and the data is then used to make the predictive model.

The establishment survey is usually performed on a specific subset of population (very young children are unlikely to be surveyed, as well as very old people). In tables 1-3, we have examples of data gathered through the establishment survey.

	DAILY	WEEKLY	MONTHLY	DAILY	WEEKLY	MONTHLY
Number of people in the market (aged 10-75)				3 550 000		
Percentage of Internet users	64%	70%	73%	2 272 000	2 485 000	2 591 500
Using domestic websites	62%	85%	93%	1 931 200	2 286 200	2 487 840
On PC	85%	92%	96%	1 772 160	2 112 250	2 228 690
On mobile	78%	85%	86%	363 520	497 000	492 385
On tablet	16%	20%	19%	<22 720	<24 850	<25 915
On smart TV	<1%	<1%	<1%	2 272 000	2 485 000	2 591 500

Table 1: Establishment survey data example

Age group	**Average number of used devices**		
	PC	**MOBILE**	**TABLET**
Total	1,3	1,1	1,1
<20	1,2	1,0	1,1
20-30	1,8	1,3	1,3
30-45	1,6	1,1	0,9
45+	1,1	0,8	0,75

Table 2: Demographic difference in device usage example

[11] For details on collection methods and methodology used in your market, contact an Ipsos DotMetrics representative.

DEVICE OVERLAP	PC	Mob	Tab	STV
PC		69,89%	75,53%	88,00%
MOB	68,39%		77,91%	87,19%
TAB	15,47%	16,31%		42,66%
STV	2,31%	2,34%	5,48%	
PC+STV		69,92%	76,51%	
PC+TAB		72,22%		95,60%
PC+MOB			86,66%	89,37%
MOB+STV	68,45%		78,71%	
MOB+TAB	70,18%			93,42%
TAB+STV	16,87%	17,67%		
PC+TAB+MOB				95,60%
PC+TAB+STV		72,22%		
PC+MOB+STV			87,46%	
MOB+TAB+STV	70,24%			

Table 3: Example of daily device overlap

Single device type unique user modelling

The DotMetrics Audience system delivers audience estimates by taking the traffic data for a certain time period, and using a predictive model to calculate the probable number of unique users for the report parameters on-the-fly. This is done by setting the foundations of the model beforehand, using establishment survey data.

Audience estimates are founded in the fact that reach of a certain website in relation to all websites behaves the same both for the devices and people in a certain period[12]. The model is established by calculating the average expected number of users behind the average number of devices available for a certain time period. Thus, using the example shown in the previous chapter, if we know that, on average, 1 800 000 PC devices access all the websites available on a market in a day, and we've calculated that, on average, 1 772 160 people access domestic websites daily – we can calculate that, for any day, the average expected number of devices per person would be 1,02.

$$dU = \frac{\bar{N}_{devices}}{\bar{N}_{users}}$$

Equation 1: Average number of devices per person

This finding can then be used to calculate the expected number of people behind each website. After calculating the number of devices that have visited a website X in a day, we can divide it by 1,02, and get the basic estimate of people who have visited that website.

This flat-rate approach, while simple, ignores the principle that not all websites have the audiences that behave the same. Since we have no way of knowing what is the average number of devices that people use on a website-per-website basis, we have to model different behaviour in some way. We do

[12] I.E. if a certain website holds 34% of reach of all devices accesing the web in that day, it's assumed that 34% of all the people are also accesing that website in that day.

that by applying the differences in average device usage in each demographic group to the demographic structure of the website (or a group of websites) we need the estimate for.

To do that, we construct a weighted average of different device usage patterns by specific demographic structure (calculated from the pop-up surveys available on the website), and then transpose the final coefficient by using the difference from survey findings and calculated averages. To break that down, firstly, we fetch the demographic structure of the analysed website from our pop-up database. The example of a fetched demographic structure for website X is shown in table 4.

Age group	Average number of used devices	
	PC	Website X structure
<20	1,2	26%
20-30	1,8	44%
30-45	1,6	23%
45+	1,1	7%

Table 4: An example of a fetched demographic structure

Using the demographic structure, we calculate the weighted average using the formula shown in equation 2.

$$dU_w = \sum_{d}^{n} P_d \times dU_d$$

$dU -$ Average number of used devices
$P -$ Percentage of demographic group
$d -$ Demographic group
$n -$ Total number of demographic groups
$w -$ weighted

Equation 2: Weighted average of used devices

The weighted average is calculated based on survey data, however, in equation 1 we've shown that the average number of used devices is calculated more accurately from data available in the traffic system and the establishment survey. Thus, we need to rebase the average number of used devices to the value gotten in equation 1. We do that by calculating the difference between the weighted average for website X and survey average for the total sample (see Table 2), and applying the difference to the calculated average number of devices from equation 1. The resulting equation is shown in Equation 3.

$$dU_w = \frac{\sum_{d}^{n} P_d \times dU_d}{dU_t} \times \frac{\bar{N}_{devices}}{\bar{N}_{users}}$$

$dU -$ Average number of used devices
$P -$ Percentage of demographic group
$d -$ Demographic group
$n -$ Total number of demographic groups
$w -$ weighted
$t -$ total
$\bar{N}_{devices} -$ Average daily number of devices
$\bar{N}_{users} -$ Average daily number of users

Equation 3: Rebased weighted average

The resulting average is then applied to the measured number of devices on website X, using the following formula:

$$N_{unique\ users} = \frac{N_{devices}}{dU_w}$$

$N_{devices}$ – Number of devices on website X in a certain day

$N_{unique\ users}$ – Number of unique users on website X in a certain day

Equation 4: Final number of unique users

After obtaining the final number of devices for the requested time frame, we can calculate the number of users in certain demographic groups by applying the audience composition percentages calculated from the pop-up demographic structure. If the website has at least 80 surveyed devices in the requested time frame, a percentage of any demographic groups can be calculated from the database of surveyed devices, and applied to the total number of unique users.

Calculating different device-type overlap

In the previous chapters, we've shown how we calculate the number of unique users available on each website on a single platform. However, websites can be accessed from multiple platforms, and each unique user on a certain platform can also be a user on a different platform – which is why, instead of just adding the number of users together, we must account for user duplication in different device types.

User duplication is, firstly, calculated in the establishment survey for each device type and for three time levels (day, week and month). The resulting percentages are shown in table 3, and they represent the probability that a known device of type A is also a known device of type B (where A represents columns, and B represents rows).

After calculating the resulting probabilities, we can calculate the probable overlap. If we want to calculate the total number of PC and mobile users for a website, our first step is to calculate the expected number of unique users separately, using the method described in the previous chapter. An example is shown in table 5.

Device type	Unique users
PC	150 000
Mobile	142 000
Tablet	36 000

Table 5: Example of calculated unique users per platform

After obtaining the expected number of unique users, we can calculate how many people that are known to be using PC's, are expected to use mobile phones. The expected number will be 68,39% of the 150 000 known PC users – or 102 585. Concurrently, we calculate the expected number of people who are known to use mobile phones, to also use PC's, and determine it to be 69,89% of 142 000 known mobile users – or 99 244. The expected number differs because the structure of device types on this particular website differs from the structure in the total Internet population. To account for this, we always assume the expected number of duplicated users to be the smaller of two estimates.

In this case, we would assume 99 244 overlapping users. The total number of unique users can be calculated by adding the unique users from each device type, and subtracting the overlap – which results in 192 756 unique users accessing website X from mobiles or PC's.

When establishing this model, we've run a simulation of different website structures in terms of device types, and the resulting cumulative function. The simulation was based on overlap data from the Croatian market, for Q2 2016. The results of the simulation are shown in figure 3 and 4.

DotMetrics Audience: Daily Unique Users Using Predictive Modelling

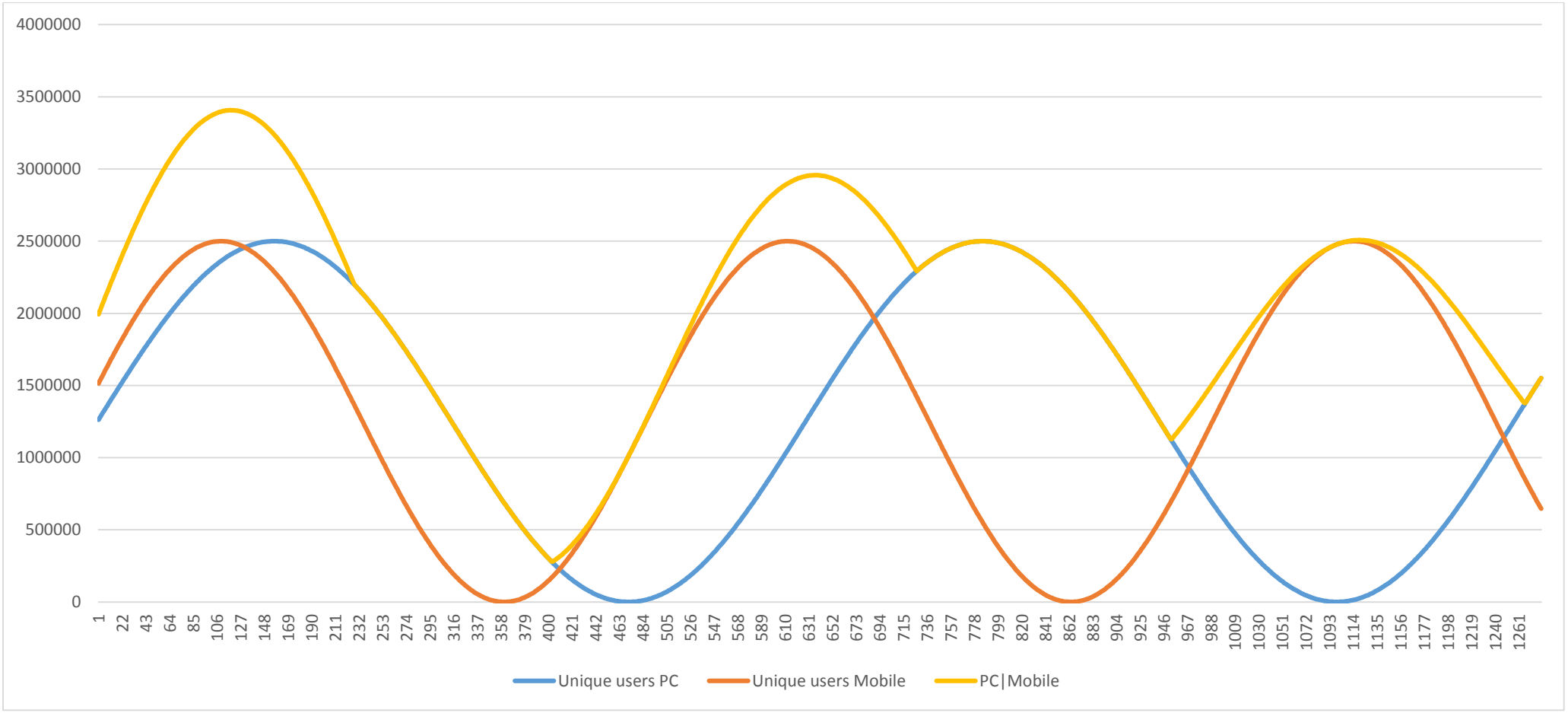

Figure 3: Simulation of PC + Mobile accumulation

DotMetrics Audience: Daily Unique Users Using Predictive Modelling

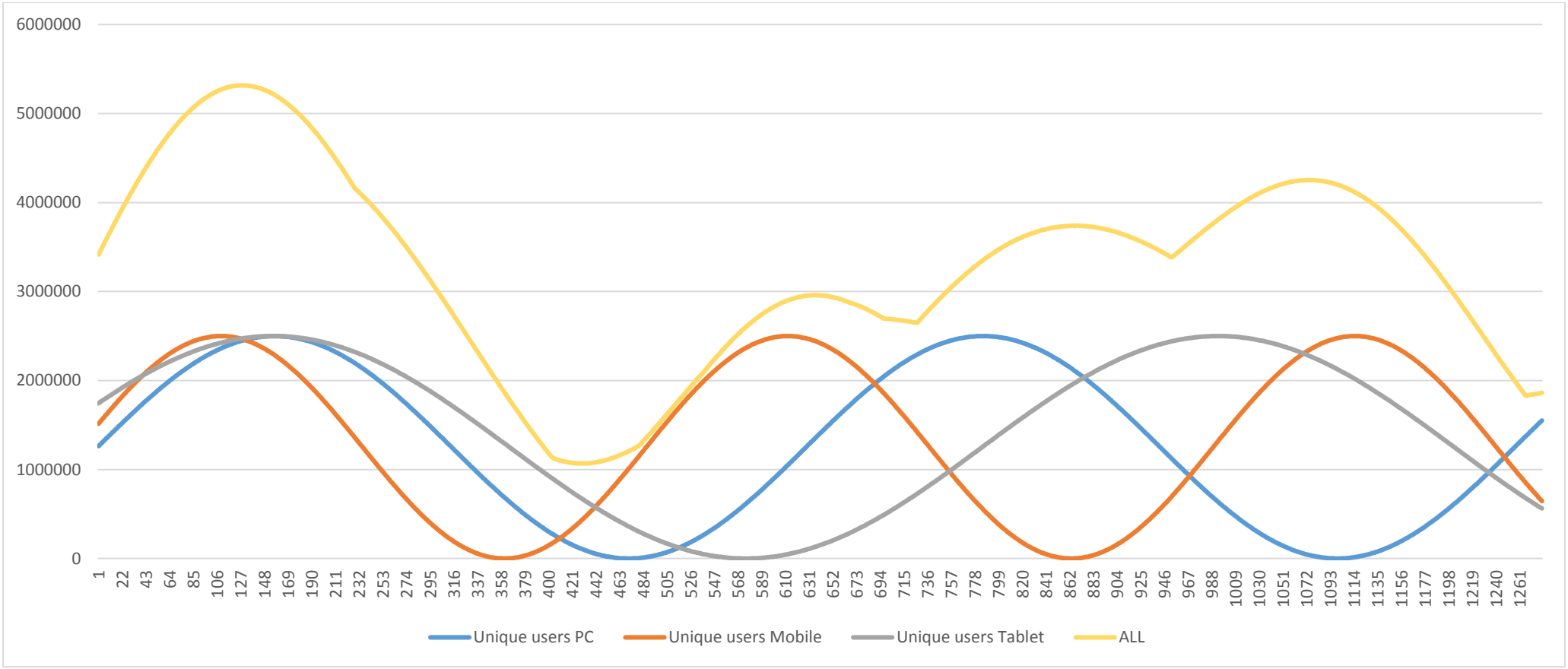

Figure 4: Simulation of PC + Mobile + Tablet accumulation

References

Eckersley Peter, “How Unique is Your Web Browser?”, Electronic Frontier Foundation, Berlin-Germany (2010). Accessed from: https://panopticlick.eff.org/static/browser-uniqueness.pdf on 12/12/2016

Kohno, Tadayoshi, Andre Broido, and Kimberly C. Claffy. "Remote physical device fingerprinting." IEEE Transactions on Dependable and Secure Computing 2.2 (2005): 93-108.

Zdziarski, Jonathan, Weilai Yang, and Paul Judge. "Approaches to phishing identification using match and probabilistic digital fingerprinting techniques." Proceedings of the MIT Spam Conference. 2006.