
Data sources

SITE-CENTRIC DATA – acquired from scripted websites, stream players and applications. Site-centric research is a direct analysis of the server logs. The measurement is performed directly on the scripted server

USER-CENTRIC DATA – contains socio-demographic characteristics collected from online questionnaires.

COOKIE PANEL – a group of people who have filled in recruitment survey on their devices.

EXTERNAL STRUCTURAL STUDY – quantitative offline surveys on a representative sample of people in the country whose results serve to determine the size and structure of the internet users population.

Below document focuses on explaining gemiusAudience methodology. The detailed definitions of site-centric metrics are available in gemiusPrism interface.

1 Real Users Methodology

In gemiusAudience methodology the basic formula for calculating the number of Real Users on a given node is as follows:

Real Users(n) = Estimated_Cookies(n) * J_Coefficient

In the below chapters each of above components is described.

EC Global

EC Global is the newest statistics for the **Estimated Cookies**, which approximates what would be the number of cookies if all browsers accepted Gemius cookies and no cookie deletion existed.

The EC Global is calculated as follows

$$EC_Global_n = GC_n \cdot \frac{AH_{Internet}}{GH_{Internet}}$$

where:

EC_Global_n - number of estimated cookies on node 'n',

GC_n - number of good cookies on node 'n' (idea explained in the next chapter),

$AH_{Internet}$ - number of all page views on the Internet node,

$GH_{Internet}$ - number of all page views made by good cookies on the Internet node.

Calculating **Estimated Cookies Global** brings higher quality to the Gemius study. EC Global evolved in a process of multiple tests where different prototype EC algorithm versions were considered.

BEAST Algorithm

BEAST Algorithm is used to estimate the number of so called *good cookies* that visited a given node. *Good cookie* is a cookie which existed throughout the whole analyzed period – so was created before analyzed period and was not deleted till the end of this period.

BEAST algorithm allows to estimate the probability value that cookie will not be deleted by the end of research period. Thus, defining a group of cookies that would be representative for all cookies and using them for website reach calculation is possible.

The calculation of probability of being a good cookie in BEAST method takes into account e.g.:

- When was the last activity of a cookie
- How often the cookie was active during 28 days before the last activity
- How long ago cookie was created
- How active was in given period
- How many cookies visited the node and how many page views they made

The method usage guaranties official results on the next day after the research period is finished.

BRUS

In order to calculate the number of Real Users (RU) – i.e. the number of human internet users who visited a given website - the two values are needed: number of browsers from which the website was visited (EC described in previous sections) and the average number of users per browser (so called J_Coefficient).

The average number of users per browser depends on a website and a platform. For websites with high EC the number of users per browser is close to the general number of users per browser on the whole measured Internet on a given market, and this is an effect of a strong statistic representation of the internet users on such websites. On the contrary, in the case of smaller websites, the number of users per browser tends toward 1. This is because, if a website is rarely visited and has low EC, the probability that it will be visited by more than one user by using the same browser is low, and it moves the number of users per browser closer to 1. This observation is based on studying the activity of software panelists on websites many markets.

The method called BRUS (from **BR**owser **US**ers) has been developed to make possible to calculate the number of users per browser as a function of size of a website, expressed in EC.

2 PRES algorithm for Population and Reach

PRES (Population and Reach Estimated Smoothly) – is a method in which Gemius uses external structural data as baseline, but starts to use additional data sources, i.e. Gemius site-centric measurement,

questionnaires and software panel, to introduce adequate trends. Thanks to that we are able to estimate population for upcoming months and make smoother changes between external study waves.

By adding internal data to population estimation it is possible to:

- Improve quality of the study
- Update data every month
- Catch the seasonality
- Get better trends tracing
- Make smoother changes between external study waves
- Correct estimation of Tablets and PC-Work population
- Estimate population for future months.

The aim of PRES is to have an automatic method of population and reach calculation and this goal was fully achieved by introducing this algorithm.

3 Application-Browser Merging (A-B Merging)

Mobile applications usage has been growing in recent years, and currently, it constitutes a significant share of all internet traffic. It became obvious that delivering a valuable image of the audience requires a solution enabling analysis of websites, applications and audio/video players in total. In response to this need, Gemius developed Application-Browser Merging (A-B merging) method.

The A-B merging aims to connect activity of the panelist recruited and measured within a web browser with their activity in the measured application. Technically, it requires pairing proper Browser ID (representing web browser used by the panelist) with Advertising ID (identifying an application installed on the user's device).

AB-merging process merges information from WWW browser and scripted applications, giving as a result a new panel containing activity of both. It is a kind of approximation, which is needed on the markets where otherwise presenting application audience would not be possible because of a lack of mobile software panel.

This new panel with application activity is the main advantage of the use of AB-merging. Other advantages are high resolution of data, as in Browser ID panel and no problems characteristic for software panel-based measurement.

4 FB/nonFB correction

The convenient way to use the Facebook on a phone or a tablet is to install the Facebook mobile application. This is a WebView-type application that uses a build-in "in-app browser" in order to let the user open web pages without exiting the Facebook app. It results in executing the Gemius tracking scripts, whenever a user

opens an audited web page inside this application. This way, even though the Facebook app is not scripted itself, it generates traffic registered by Gemius on audited web pages, more like a browser than application.

The problem is that the Facebook in-app browser does not share the 3rd party cookies/browserIDs with any regular web browser (Safari Mobile, Chrome Mobile, etc.). This means that the same user is registered under two different identifiers when entering the same audited web page via Facebook in-app browser and regular browser, in a given time period, e.g. a week.

This duplication of identifiers, causes overestimation of EC value for Internet node and it then affects the calculated for websites - both on Phone and Tablet platforms.

In order to overcome this problem and receive more accurate results Gemius introduced a new split into two platforms: "Facebook in-app browser" and "not Facebook in-app browser", in short called: FB/non-FB platforms. This split concerns both mobile platforms separately - there are platforms: Phone-FB/Phone-non-FB, and Tablet-FB/Tablet-non-FB.

The FB/non-FB platforms are only technical platforms, created during the process of data calculations - they are not shown in the data outputs. Moreover, the split concerns only site-centric data (not the panel). This is because the Mathematical Function is applied to achieve the deduplicated values of RU for Phone or Tablet platform.

Mathematical Function is applied to the data of the FB and non-FB platforms (RU, visits and populations) in order to receive the final deduplicated RU of each node on the platform. The corrected RU received this way for each node is then applied as the behavioral conditions in the RIM weighting process of the panel.

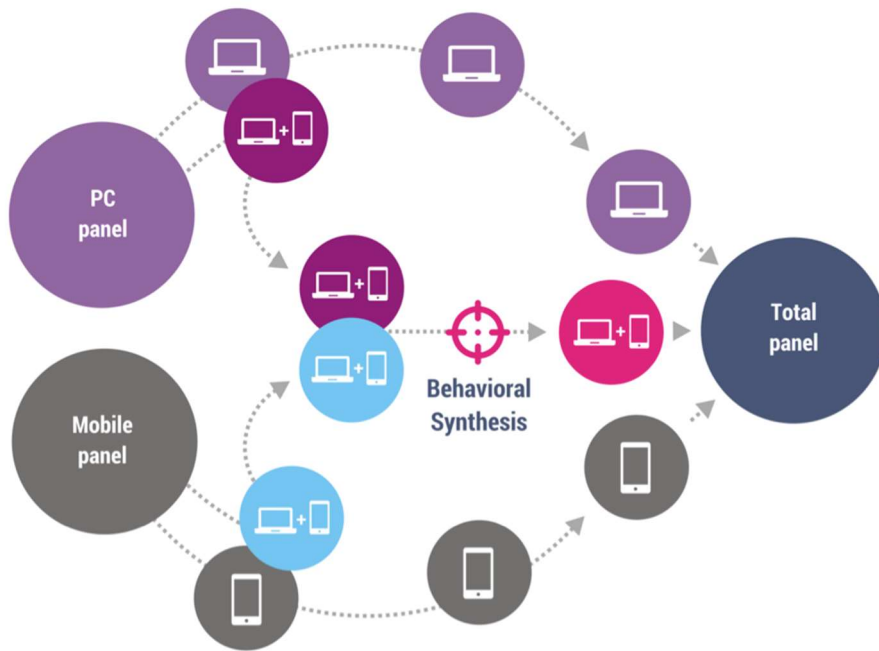
5 Behavioral Panel Synthesis

Behavioral Panel Synthesis (BPS) algorithm allows to calculate the total number of *Real Users* for any kind of platform combination, websites and target groups. All available panels are combined into one *cross-platform panel*. To perform BPS algorithm the following data are needed:

- panels on each analyzed platform;
- *calibration panel* (consists of panelists for whom all of their online activities are measured on least two platforms);

BPS algorithm consists the following steps:

1. calibration panel construction:
 - *calibration panel* – set of IDs from different platforms that represent the same user;
2. prediction model construction:
 - main goal: probability estimation that a pair of panelists from two different devices is the same person;
3. metrical clustering – division of research panels into separable clusters due to the socio-demographic convergence.
4. panelist merging:
 - probability that pair of the panelist is the the same person;
 - merging procedure;
 - demographics calculations – demographics is based on the PC panelists. If there is no PC panelist – on the most current panelist.



The last step of the BPS procedure to calculate Total number of Real Users is the weighting process. There are three main groups of boundary condition to which Total panel is weighted:

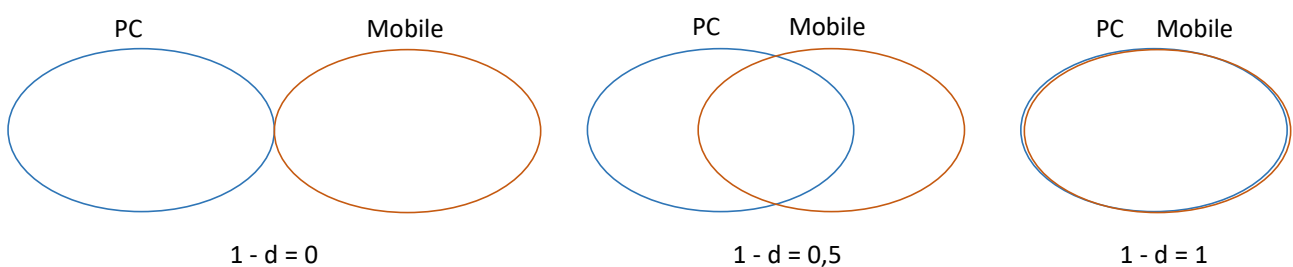
1. Real Users for each website on each platform
2. Socio-demographic structure of each platform (based on structural study),
3. Population sizes of users who use a particular set of devices, e.g. Desktop only, Desktop and Mobile, etc. (based on structural study)

Finally the Total number of Real Users is equal to the sum of weights of panelists that visited particular website.

BPS Lite

The BPS Lite algorithm is used instead of standard BPS algorithm in case of too small calibration panel has been recruited.

In this case the Mathematical Function is used to estimate for each website the duplication d of RU on different platforms.



This duplication corresponds to the coefficients that define the importance of behavioral characteristics (i.e. visited a website or not) during panelists merging, as described in previous section.

6 Constant Panel

Constant Panel is a modelled panel, which contains very high number of virtual panelists with equal weights. Instead of changing weight of one panelist, we simulate larger number of panelists from large target groups and smaller number of panelists from small target groups. Panel is called “constant”, because number of panelists and their weights do not change in time.

By using non-rotating panelists with equal weights, we directly simulate situation from real-world population, where all people are counted equally and natural rotation is quite low. Thanks to this stability, we might easily calculate consistent results for any given custom time period directly from the panel.

Constant Panel might be formed from any real input panel, and also from many real panels if each one of them provides different part of panelist activity. On the markets, where it is available, Constant Panel consists Single Source Panel as one of the inputs, providing unprecedented quality of cross-media measurement, combining internet, TV and radio data.

Constant Panel – our revolutionary approach of modelling directly real-world population, will transform market research industry forever, by bringing new levels of quality, usefulness and convenience.



Gemius S.A.

D48 Building, Domaniewska 48
02-672 Warsaw, Poland

+ 48 22 390 90 90

+ 48 22 378 30 50

contact@gemius.com

gemius.com