MEDIAPANEL

Metodologia badania Mediapanel

dla Klientów oraz Działu Sprzedaży Gemius

2024

1. Wprowadzenie

Badanie Mediapanel, jest najstarszym produktem Gemius, który przez wiele lat stanowił podstawię oferty firmy. Można powiedzieć, że wiedza i doświadczenie zgromadzone przez lata realizowania Badania stanowią podwalinę pod kolejne generacje produktów Gemius, takich jak gemiusAdReal, gemiusDirectEffect czy gemiusPostBuy. Dzięki temu wiele z nich dzieli z nim definicje, źródła danych, skrypty pomiarowe czy metodologie. Ninijeszy dokument pozwala lepiej zrozumieć nie tylko samo badanie Mediapanel, lecz także metodologię stojącą za pozostałymi produktami z portfolio Gemius.

2. Krótkie definicje

Dane site-centric (uniwersy) – dane pochodzące z bezpośredniego pomiaru po stronie wydawcy, gromadzone za pośrednictwem skryptów pomiarowych Gemius zaimplementowanych na witrynie, w playerze (materiały streamowe) czy w aplikacji mobilnej. Bazują na nich precyzyjnie mierzone metryki, takie jak odsłony czy czas, ale także liczba realnych użytkowników, która jest wskaźnikiem estymowanym i przeliczanym w późniejszym procesie. Pomiar site-centric traktujemy jako punkt odniesienia dla wyników pochodzących z panelu, które są do nich dopasowywane w procesie ważenia lub wdrażania algorytmów korygujących błędy pomiaru. Produktami, które wykorzystują dane site-centric jako wyniki referencyjne są: Mediapanel, gemiusAdReal, gemiusDirectEffect oraz gemiusPostBuy.

Dane user-centric – wspólna nazwa dla wszystkich danych pochodzących z paneli (tj. aktywności i demografii panelistów). Uściślając, są to dane pochodzące bezpośrednio od użytkowników (panelistów cookie/software/hardware). Demografia panelistów (zaczerpnięta z ankiet rekrutacyjnych Gemius) jest łączona z ich aktywnością na stronach internetowych, w playerach i aplikacjach.

Aktywność cookie panelistów pochodzi ze skryptów site-centric, aktywność software panelistów – z naszego rozszerzenia dlaprzeglądarek (dla PC) lub naszej mobilnej aplikacji pomiarowej (dla telefonów i tabletów), natomiast aktywność hardware panelistów - ze smartfonów z zaimplementowanym na stałe naszym oprogramowaniem pomiarowym. Więcej informacji na temat charakterystyki poszczególnych paneli znajduje się poniżej.

Sonary - zaimplementowana w skryptach Gemius metoda pomiaru czasu spędzanego przez internautów na aktywnych zakładkach przeglądarki, której głównym celem jest realizacja dokładnego pomiaru przy możliwie najmniejszej liczbie hitów pomiarowych. Skrypty „decydują" co sekundę, czy powinny wysłać sygnał pomiarowy dla danej audytowanej strony internetowej otwartej w zakładce, czy nie, a decyzja podejmowana jest z niskim prawdopodobieństwem (w Polsce jest to to 1/40). W tym podejściu liczba potrzebnych hitów jest dosyć niska, i choć pomiar dla pojedynczego pliku cookie nie jest wystarczająco dokładny, to dla wszystkich plików cookie odwiedzających poszczególne strony internetowe metoda jest dokładna.

Filtrowanie autoodświeżeń - proces usuwania takich odsłon niezgodnych z naszą definicją odsłony (intencjonalne działanie użytkownika), które charakteryzują się tym, że powstają w wyniku okresowego odświeżania zawartości serwisu (np. aktualizacji wyników sportowych w trakcie meczu). Filtrowanie ich opiera się na analizie interwałów czasowych pomiędzy kolejnymi wejściami tego samego użytkownika na ten sam adres URL. Gdy dany dokładny przedział czasowy występuje nienaturalnie

często (co najmniej 2x częściej niż mediana z 60 przedziałów czasowych wokół), jest on oznaczany jako automatyczne odświeżanie, a odpowiadające mu odsłony są odfiltrowywane.

Cookie Panel – grupa plików cookie, którym wyświetlony został nasz kwestionariusz badawczy (w momencie, gdy dane cookie odwiedziło audytowaną witrynę) i których użytkownicy odpowiedzieli na wszystkie zadane pytania. W przypadku Cookie Panelu znamy zadeklarowaną demografię każdego pliku cookie, rejestrujemy również wszystkie jego wizyty na witrynach i playerach audytowanych. Z Cookie Panelu nie pozyskujemy natomiast żadnych informacji na temat witryn nieaudytowanych oraz aplikacji (nawet audytowanych).

Software Panel – grupa panelistów, którzy po wypełnieniu naszego kwestionariusza badawczego zainstalowali również na swoich komputerach, tabletach lub telefonach specjalne oprogramowanie pomiarowe. Ta metoda badawcza niesie ze sobą więcej możliwości niż Cookie Panel, ponieważ mierzy również ruch na nieaudytowanych witrynach i aplikacjach. Software Panel jest co do zasady mniejszy i trudniejszy do zgromadzenia, a co za tym idzie – wyniki pochodzące z tej metody obarczone są większym błędem statystycznym.

Single Source Panel – Panel jednoźródłowy (nazywany również Hardware Panelem lub Panelem Urządzeń) jest panelem badawczym, w ramach którego Gemius zawiera długoterminowe umowy z panelistami i przekazuje im specjalnie przygotowane smartfony z oprogramowaniem badawczym, umożliwiającym pomiar ich oglądania telewizji i słuchania radia dzięki metodzie „audio-matchingu" (definicja w dalszej części dokumentu).

Paneliści są również zachęcani do instalacji rozszerzenia pomiarowego w swoich przeglądarkach na komputerze (PC) (stają się wtedy Software Panelistami), dzięki czemu możemy też mierzyć ich aktywność na komputerach stacjonarnych lub laptopach. Dzięki temu rozwiązaniu analizujemy zasięg w internecie (smarfony i PC), radiu i telewizji za pomocą wspólnego panelu.

Strategię firmy Gemius można podsumować jednym zdaniem: jedno źródło danych dla wszystkich mediów. Panel jednoźródłowy stanowi bezpośrednie wprowadzenie tej strategii w życie.

Fingerprinty audio - kod numeryczny generowany w stratny sposób z nagrania audio, który jest jednak wystarczająco precyzyjny, aby umożliwić porównanie dwóch różnych nagrań i ustalenie, czy mają tę samą zawartość, ale równocześnie uniemożliwia odtworzenie lub interpretację oryginalnego dźwięku.

Sygnał referencyjny - zestaw strumieni audio pochodzących z kanałów telewizyjnych i radiowych na żywo, stanowiący punkt odniesienia, do którego porównujemy i dopasowujemy fingerprinty audio ze smartfonów z naszym oprogramowaniem pomiarowym, tworzących Panel jednoźródłowy.

Audio-matching - proces znajdowania podobieństw pomiędzy fingerprintami audio, pochodzącymi ze smartfonów z naszym oprogramowaniem pomiarowym, tworzących Panel jednoźródłowy a sygnałem referencyjnym. Celem porównania jest identyfikacja, które kanały telewizyjne lub radiowe słyszał dany panelista. Audio-matching realizowany jest w bardzo bezpieczny sposób, z pełnym poszanowaniem prywatności panelisty, dlatego firma Gemius nie jest w stanie pozyskać z fingerprintów audio żadnych innych informacji niż tych dotyczących samego faktu słuchania określonych programów radiowych lub telewizyjnych oraz reklam.

Gender Automatic Prediction Algorithm (GAPA) - algorytm, którego celem jest przewidywanie płci panelisty na podstawie udziału w jego aktywności stron internetowych i aplikacji, z których korzystają głównie kobiety lub mężczyźni. Dane referencyjne dotyczące udziału płci pochodzą z mobilnego Cookie Panelu oraz mobilnego Software Panelu. GAPA służy do kontroli jakości Panelu Single Source: w przypadku niezgodności pomiędzy deklarowaną a przewidywaną płcią panelisty, uruchamiana jest specjalna procedura mająca na celu sprawdzenie, czy przestrzega on warunków umowy z Gemius, a przede wszystkim czy nie przekazał smartfona z oprogramowaniem pomiarowym osobie trzeciej. Korzystając z tych samych zasad GAPA może zostać wykorzystany do znalezienia aktywności typowej dla dzieci i stwierdzenia niezgodności w przypadku, gdy zadeklarowanym głównym użytkownikiem danego telefonu jest osoba dorosła. Algorytm wykorzystuje niewielki zestaw danych o panelistach (tylko odwiedzane domeny i aplikacje) i jest bardzo prosty w swojej konstrukcji (np. nie wykorzystuje uczenia maszynowego). Dlatego też został zaprojektowany do znajdowania tylko wyraźnych przypadków niedopasowania (tj. wartości odstających) i nie może być używany do uniwersalnego przewidywania płci.

Wymagana wielkość panelu – jedną z podstawowych zasad obowiązujących w statystyce jest zdanie „im większa próba, tym lepsze wyniki". Klasyczna statystyka pozwala nam wyliczać wielkość próby (t.j. liczebność panelistów) potrzebną do osiągnięcia określonego progu błędu. Aby to zrobić, trzeba zdefiniować najpierw margines błędu, poziom ufności (czyli prawdopodobieństwo, że rzeczywisty błąd zmieści się w założonym marginesie błędu) oraz wielkość populacji.

Przyjrzyjmy się przykładowi: populacja 2 000 000 osób, margines błędu ±3pp (punkty procentowe lub procent populacji), poziom ufności 95%. Dla tak określonych parametrów wejściowych wymagana wielkość próby to 1067 panelistów (istnieje wiele intuicyjnych kalkulatorów online do tego typu wyliczeń, nie podajemy więc w tym miejscu konkretnego wzoru). Co ważne, liczba ta nie rośnie wraz ze wzrostem populacji, dlatego dla większej populacji nadal wynosić będzie 1067.

Zgodnie z opisanym klasycznym podejściem zakładamy, że dobrana przez nas próba jest w pełni losowa, a dzięki temu – w wysokim stopniu reprezentatywna. W przypadku, gdy panel nie jest idealnie reprezentatywny i wykorzystujemy procedurę ważenia, musimy podzielić wymaganą wielkość panelu zgodnie z efektywnością ważenia (definicja poniżej). Im niższa efektywność ważenia, tym większa liczebność panelu, której reprezentatywność należy skorygować we wszystkich grupach celowych. Powiedzmy, że doważyliśmy panel i osiągnęliśmy efektywność ważenia na poziomie 40%. Oznacza to, że należy dokonać obliczenia: 1067/0.4 = 2667.5. W rezultacie uzyskujemy wymaganą wielkość panelu liczącą 2668 panelistów, abyśmy mieli 95% pewności, że dopuszczalny błąd mieści się w zakresie ±3pp.

Dane populacyjne (dane strukturalne) – Badanie założycielskie, zestaw danych opisujących strukturę demograficzną danej populacji (a więc udział procentowy kobiet i mężczyzn, rozkład wieku, miejsca zamieszkania itd.).

Good cookie (dobry identyfikator) - plik cookie aktywny co najmniej raz: przed, w trakcie i po analizowanym okresie. Jeśli obserwujemy good cookies przed analizowanym okresem i po nim, mamy pewność, że były one aktywne przez cały analizowany okres. Zakładamy, że pliki cookies, które nie zostały usunięte, są statystycznie reprezentatywne dla całego rynku.

BEAST (Browser Estimation Algorithm Standard) - jest to algorytm predykcyjny, pozwalający ocenić czy dany plik cookie okaże się w przyszłości plikiem good cookie. Przewidywanie to dokonywane

jest na podstawie poziomu wcześniejszej aktywności danego cookie oraz czasu, jaki upłynął od dnia, gdy widzieliśmy go po raz ostatni.

OverNight - część bieżącego procesu produkcyjnego badania Mediapanel, która wykorzystuje:

- metodologię BEAST w celu obliczenia już następnego dnia (a więc po upływie nocy) wyników dla poprzedniego dnia/tygodnia/miesiąca,

oraz

- proces BPS, mający na celu znalezienia duplikacji między platformami, i wyliczenie zdeduplikowanych wartości liczby użytkowników (RU) dla połączonych platform, takich jak PC (PC-Home i PC-Work) lub Total (PC i Mobile).

AppVisitorID - aplikacje na telefony i tablety mogą dostarczać różnych identyfikatorów, z różnymi poziomami ograniczeń, a my chcemy zawsze używać najlepszego dostępnego identyfikatora. Uszeregowaliśmy je według przydatności i jakości, jaką zapewniają naszemu badaniu:

1. **Advertising ID (na androidzie funkcjonuje pod nazwą AAID a na iOS IDFA)** - najlepszy, wspólny dla wszystkich aplikacji na danym urządzeniu identyfikator, który prawdopodobnie zostanie niestety wkrótce wyeliminowany przez twórców systemów operacyjnych,

2. **Vendor ID (na androidzie funkcjonuje pod nazwą ASID a na iOS IDFV**) - trudny do zmiany przez użytkownika, ale nie zapewnia duplikacji pomiędzy aplikacjami różnych dostawców,

3. **Nasz standardowy identyfikator BID/cookie aplikacji** - również nie zapewnia duplikacji i często jest najłatwiejszy do usunięcia (co powoduje w takim przypadku znaczną rotację).

Z technicznego punktu widzenia, aby ułatwić pracę, zdecydowaliśmy się stworzyć nowy identyfikator, który jest oparty na tym z tych 3 identyfikatorów, który w przypadku danej aplikacji jest w największym stopniu możliwy do wykorzystania i nazwaliśmy go **"AppVisitorID"**.

EC (Estimated Cookies) - szacunkowa liczba plików cookie, które zostałyby zarejestrowane na danym węźle, gdyby nie istniało zjawisko kasowalności cookies, a wszystkie przeglądarki akceptowałyby nasze cookie. W takiej hipotetycznej sytuacji każda przeglądarka byłaby reprezentowana przez jeden plik cookie. Innymi słowy, EC to szacunkowa liczba przeglądarek, które odwiedziły dany węzeł w danym okresie.

EC Standard - wzór na szacunkową liczbę przeglądarek, w którym współczynnik skalowania jest oparty na wartościach składowych dotyczących danego węzła (domeny, serwisu lub grupy) n. Jest to metoda przez nas rekomendowana do wprowadzenia do badania, ale jeszcze nie została zastosowana.

$$EC_{Standard} = GC_n * \frac{AH_n}{GH_n}$$

EC Global – wzór na szacunkową liczbę przeglądarek, w którym współczynnik skalowania opiera się nie na poszczególnych węzłach, ale na całym audytowanym internecie. Jest powszechnie stosowany do tej pory, jednak planujemy jego zastąpienie przez EC Standard, który rekomendujemy obecnie.

$$EC_{Global} = GC_n \cdot \frac{AH_{Internet}}{GH_{Internet}}$$

J BRUS – szacowana liczba użytkowników przypadających na jedną przeglądarkę, wyliczona dla danego węzła, dla danej platformy. Wartość ta zależy od dwóch czynników: specyficznego dla danego rynku poziomu nasycenia danym rodzajem urządzeńia (tzw. J strukturalne) oraz względnej wielkości analizowanego węzła. W przypadku większych węzłów wartość J jest zbliżona do wartości dla całego internetu, w przypadku mniejszych – zbliża się do 1.

Real Users (RU) – szacowana liczba rzeczywistych osób (nie komputerów, plików cookie czy adresów IP), którzy odwiedzili dany węzeł. Wartość ta jest wyliczana jako liczba przeglądarek (EC) razy liczba osób przypadających na daną przeglądarkę (J brus).

$$RU = EC \cdot J\ brus$$

Real Users dla aplikacji mobilnych - analogicznie do powyższej definicji, jest to szacunkowa liczba rzeczywistych osób korzystających z aplikacji mobilnych. Metoda wykorzystuje również to samo równanie, co powyżej. Jednakże w tym przypadku potrzebny jest krótki komentarz:

1. Zalecanym identyfikatorem do użycia podczas obliczania EC aplikacji jest AppVisitorID (na razie używamy jeszcze Advertising ID),

2. Zakładamy, że wszystkie identyfikatory AppVisitorID (i Advertising ID) są "dobrymi identyfikatorami", więc nie musimy używać w tym przypadku metody BEAST, aby wyliczyć EC aplikacji. W procesie wyliczania tej wartości mnożymy jedynie liczbę identyfikatorów przez stosunek liczby wszystkich odsłon do liczby odsłon zawierających informację o Advertising ID, aby uwzględnić w badaniu też użytkowników, którzy nie przesyłają nam tego identyfikatora.

3. Używamy metody J_BRUS, gdzie J i EC dla całego internetu pochodzą z danych strona internetowych z platformy Phones, stanowiących najlepsze dostępne przybliżenie.

Zasięg audytowany badania – udział internautów na danym rynku, którzy odwiedzili przynajmniej jedną audytowaną przez Gemius witrynę w miesiącu.

PRES (Population and Reach Estimated Smoothly) – jest to metoda szacowania populacji na wszystkich platformach oraz estymowania audytowanego zasięgu badania na danym rynku. Podstawową zasadą metodologii estymowania populacji PRES jest wykorzystanie jako punktu wyjścia zewnętrznego badania, a następnie nałożenie na nie trendów, analogicznych do miesięcznych zmian EC. Audytowany zasięg badania szacowany jest bezpośrednio na podstawie Software Panelu

Klasyfikator home/work– wiele osób korzysta z osobnych komputerów stacjonarnych lub laptopów w domu i w pracy, również tego samego dnia. Z tego powodu, aby uzyskać poprawną (zdeduplikowną) liczbę RU dla całej platformy PC, musimy znać poziom duplikacji pomiędzy tymi dwoma grupami urządzeń.

Klasyfikator H/W jest klasyfikatorem wykorzystującym machine learning, który rozpoznaje, czy dany adres IP pochodzi z domu czy z pracy użytkownika i zgodnie z tym rozpoznaniem przypisuje ruch na platformie PC.

W kwestionariuszach PC pytamy panelistów, czy w danym momencie znajdują się w pracy czy w domu. Dzięki temu zyskujemy zestaw uczącyzłożony z adresów IP przypisanych do obu grup. Następnie, wykorzystując go, uczymy klasyfikator za pomocą machine learningu odpowiedniego przypisywania wszystkich pozostałych adresów IP na danym rynku do grupy home lub work.

Podział "Facebook / non-Facebook" – aplikacje Facebook na telefonie lub tablecie mogą pełnić role przeglądarki (tzw. web-view lub in-app browser), jednak nie dzielą one plików cookie z żadną tradycyjną przeglądarką. Zjawisko to powoduje, że w przypadku niektórych węzłów mamy do czynienia z dwoma plikami cookie pochodzącymi od jednej osoby: ze zwykłej przeglądarki oraz z aplikacji Facebook. Może to prowadzić do przeszacowania liczby RU dla węzłów popularnych na obu typach przeglądarek.

Podział "Facebook / non-Facebook" to metoda rozdzielania plików cookies pomiędzy dwie platformy technologiczne (zarówno na telefonach jak i tabletach), nazwane Facebook oraz non-Facebook, a następnie szacowania duplikacji pomiędzy nimi. Dzięki temu możemy przewidzieć, jaki procent plików cookies z obu platform (czyli Facebook i non-Facebook) był w rzeczywistości wygenerowany przez tę samą osobę. Informację tę wykorzystujemy, aby odpowiednio skorygować końcowy wynik.

Jest to metoda opracowana specjalnie na potrzeby tej konkretnej aplikacji. Analogiczne rozwiązanie stosujemy obecnie też w przypadku ruchu generowanego w przeglądarce wbudowanej w aplikację Wiadomości Google (Google News). Nie wykluczamy, że w przyszłości wprowadzimy analogiczne rozwiązanie dla innych aplikacji ze znaczącym ruchem "in-app browser" oraz własną przestrzenią cookies.

Ważenie wieńcowe (RIM weighting) – proces przypisywania wagi każdemu paneliście w celu określenia, jak wielu ludzi w rzeczywistej populacji powinien on reprezentować a w konsekwencji niwelowania ewentualnych odchyleń panelu i zwiększania jego reprezentatywności. Jeśli na przykład wiemy z badania strukturalnego, że rozkład płci powinien wynosić 50% dla kobiet i 50% dla mężczyzn, a w naszym panelu obserwujemy 40% kobiet i 60% mężczyzn, przypisujemy panelistom płci żeńskiej wyższą wagę, tak by w efekcie sumowała się ona do 50% wszystkich wag panelistów. Proces, w którym usuwamy odchylenia dotyczące zmiennych strukturalnych, takich jak płeć, wiek, wykształcenie itp., nazywamy "ważeniem strukturalnym". Możemy również niewelować odchylenia dotyczące aktywności panelistów, np. jeśli mamy zbyt małą liczbę panelistów aktywnych w danym węźle, musimy zwiększyć wagi tych panelistów, aby spełnić warunki referencyjne (RU z uniwersów). Takie działanie nazywamy "ważeniem behawioralnym". Ważenie jest bardzo powszechnym procesem w naszej pracy z danymi. W standardowej produkcji badania Mediapanel jest ono używane dziesiątki razy, aby zyskać pewność, że na wielu różnych etapach produkcji zbiór danych panelu zawsze spełnia określone warunki.

Efektywność ważenia (Weighting Efficiency, WE) – to wyrażona procentowo miara, opisująca, jaki jest poziom ujednolicenia wag panelistów. Maksymalne ujednolicenie wag jest pożądaną właściwością procesu ważenia, ponieważ zapewnia to prostsze wyliczenia oraz dokładniejsze wyniki dla małych węzłów. Efektywność ważenia równa 100% oznacza, że wszystkie wagi są takie same, natomiast zbliżona do „0" świadczy o maksymalnym zróżnicowaniu wag.

Warto zauważyć, że w przypadku Constant Panelu (szczegółowy opis poniżej) wszyscy paneliści mają równe wagi, możemy więc powiedzieć, że technicznie rzecz biorąc jego efektywność ważenia WE=100%. Constant Panel jest jednak wirtualnym, modelowanym panelem, który nie podlega bezpośredniemu ważeniu, dlatego określanie dla niego efektywności ważenia nie ma uzasadnienia. W związku z tym, w przypadku Constant Panelu po prostu jej nie definiujemy.

Census Safeguarding Age Correction (także Central Statistical Office correction) - zestaw warunków ważenia dotyczących wieku (lub przecięć wiek-płeć), które wykraczają poza standardowe dane strukturalne i są oparte na oficjalnym spisie ludności danego kraju. Są one skonstruowane w taki sposób, że suma wag panelistów w określonym wieku i określonej płci nie może przekraczać (np. O więcej niż 3000) liczby osób odpowiadających określonemu wiekowi i płci w całej populacji danego kraju.

Są to warunki mające wyłącznie charakter zabezpieczający, więc jeśli proces ważenia przebiega normalnie, nie wywołują żadnych zmian. Wchodzą one w życie tylko w skrajnych i rzadkich przypadkach, gdy standardowe ważenie, ze względu na silne odchylenia w panelu lub wyjątkowo niską skuteczność ważenia, zwróciłoby nielogiczne wyniki dla wąskiej grupy wiekowej, przekraczające liczbę osób w określonym wieku w populacji. W takim przypadku dodatkowe warunki zmniejszają wagi dla tej wąskiej grupy i ponownie rozdzielają je na inne grupy wiekowe w tym samym strukturalnym przedziale wiekowym, w wyniku czego zachowują zarówno standardowe, strukturalne warunki, jak i nie przekraczają wartości wynikających ze spisu ludności.

AB-merging (Aplication Browser merging) – metoda przypisywania aktywności z aplikacji mobilnych do mobilnego Cookie Panelu, który nie wykazuje takiej aktywności (nie rekrutujemy panelistów w aplikacjach mobilnych). Główne założenie tej metody polega na łączeniu identyfikatorów aplikacji (tzw. Advertising ID) z plikami cookie (BID) panelistów ze standardowych przeglądarek poprzez sprawdzanie, czy te dwa identyfikatory pojawiły się w tym samym czasie, w ramach tego samego adresu IP.

Balance Amendment – metoda komplementarna względem strukturalnego i behawioralnego ważenia panelu. Zaprojektowana w celu usunięcia potencjalnego systematycznego błędu estymacji węzłów nieaudytowanych, spowodowanego odchyleniami panelu (tj. zbyt wysoką lub zbyt niską średnią aktywnością panelistów). Metoda ta automatycznie znajduje optymalną funkcję korekty RU, która daje nam najniższy średni błąd RU, w pierwszej kolejności na witrynach audytowanych (gdzie mamy do porównania zarówno dane site-centric, jak i ważone panelowe). Następnie funkcja korekty stosowana jest do węzłów nieaudytowanych w celu zminimalizowania na nich błędów. Metoda ta ma również wbudowany mechanizm zabezpieczający, który dodatkowo zmniejsza prawdopodobieństwo przeszacowania w węzłach nieaudytowanych.

Fuzja Cookie-Software – to proces łączenia informacji pochodzących z dwóch paneli. Cookie Panel dostarcza informacji o witrynach audytowanych, natomiast Software Panel – o nieaudytowanych. Cookie Panel jest znacznie większy niż Software Panel, a co za tym idzie – lepszej jakości. Z tego powodu traktowany jest priorytetowo i to do niego dołączane są informacje na temat witryn nieaudytowanych pochodzące z Software Panelu. Optymalne połączenia mają miejsce wówczas, gdy zidentyfikujemy przypadki podobne, czyli połączymy oba panele bazując na możliwie najbardziej podobnej aktywności panelistów. Warto zauważyć, że w przypadku znalezienia optymalnych

przypadków podobnych, aktywność z Software Panelu jest pomijana, ponieważ jej lepsza wersja dostępna jest w Cookie Panelu.

BPS (Behavioral Panel Synthesis) - metoda łączenia panelistów z różnych platform (PC, Phone, Tablet) w jednego panelistę, który może wykazywać się aktywnością w różnych typach mediów. Algorytm łączenia panelistów oparty jest na podobieństwach behawioralnych . Oznacza to, że dążymy do skojarzenia panelisty, który odwiedził dany węzeł na urządzeniu PC z panelistą, który odwiedził ten sam węzeł za pośrednictwem telefonu. Podejście to skutkuje określeniem współoglądalności pomiędzy platformami. Zadaniem BPS jest też odtworzenie duplikacji w ramach tzw. panelu kalibracyjnego (grupa panelistów, których aktywność zaobserwowano na więcej niż jednej platformie). Głównym celem metodologii BPS jest obliczenie łącznego zasięgu na danym węźle, a więc określenie wartości Total Real Users, niezależnie od tego, z której platformy korzystali użytkownicy.

Constant Panel – jest to modelowany panel, zawierający bardzo dużą liczbę wirtualnych panelistów o równych wagach. Zamiast zmieniać wagę pojedynczego panelisty, symulujemy istnienie większej liczby panelistów z dużych grup celowych oraz mniejszej liczby panelistów z małych grup celowych. Panel nazywany jest „stałym", ponieważ liczba panelistów i ich wagi nie zmieniają się w czasie.

Wykorzystanie stałej puli panelistów o równych wagach pozwala nam odwzorować sytuację populacji z realnego świata, gdzie wszyscy ludzie są liczeni tak samo, a naturalna rotacja utrzymuje się na niskim poziomie. Dzięki tej stabilności możemy bezpośrednio z panelu w łatwy sposób wyliczać spójne wyniki dla dowolnego okresu.

Constant Panel może być tworzony na bazie dowolnego realnego panelu wyjściowego lub wielu realnych paneli, jeśli każdy z nich dostarcza informacji o innej części aktywności panelistów. Constant Panel wykorzystuje jako jeden z paneli wyjściowych również panel Singel Source. Dzięki temu uzyskujemy bezprecendensową jakość pomiaru cross-mediowego, obejmującego dane internetowe, telewizyjne oraz radiowe

Daily Fusion - proces mapowania panelistów z wczorajszych wyników na poprzedni 27-dniowy Constant Panel, osiągając docelową 28-dniową wartość RU pochodzącą z procesu BPS, dla każdego węzła i platformy jednocześnie.

Fill Panelist Activity (FPA) - jest algorytmem edytowania aktywności panelistów w taki sposób, aby w najwyższym możliwym stopniu odpowiadała wynikom pozyskanym z pomiaru site-centric (odsłony, wizyty, czas). Cel ten realizowany jest poprzez dodawanie lub odejmowanie aktywności na poszczególnych węzłach. Pozornie jest to proces zbliżony do ważenia, ponieważ mają one zbieżny cel, jakim jest dopasowanie wyników do danych referencyjnych site-centric. Główna różnica pomiędzy tymi procesami polega na tym, że w przypadku ważenia panelistom przypisywane są różne wagi, co wpływa na zmianę wartości wszystkich metryk (w tym RU) i powoduje, że są jedynie w przybliżeniu równe uniwersom. FPA jest algorytmem dostrajającym, w ramach którego precyzyjnie modyfikujemy niewielkie części aktywności panelistów (już po ważeniu), w celu otrzymania dokładniejszych wyników takich metryk, jak odsłony, czas i wizyty. W procesie tym staramy się nie wpływać na wartość Real Users.

Heavy Panelists Filtration (HPF) - klasyczna metoda usuwania wartości odstających, której celem jest znalezienie i zniwelowanie nienaturalnie wysokich wartości odsłon, wizyt i czasu, pochodzących z aktywności pojedynczego realnego panelisty. HPF – podobnie jak FPA - jest częścią końcowego post-

procesu przetwarzania danych, przy czym HPF dotyczy tylko węzłów nieaudytowanych, a FPA precyzyjnie koryguje węzły audytowane, wykorzystując jako punkt odniesienia dane site-centric.

Gdy HPF wykryje aktywność heavy panelisty (tj. panelisty, który sam wygenerował znaczną część odsłon lub czasu węzła), algorytm radykalnie zmniejsza wszystkie statystyki dla tego panelisty, ustawiając je jako równe liczbie wizyt panelistów na tym węźle (w rezultacie każda Wizyta liczy jedną Odsłonę i trwa 1 sekundę). Nie usuwamy całej aktywności danego panelisty, a jedynie tę jej część dotyczącą pojedynczego węźła, która okazała się nadmierna.

JAR (Joint Apocalypse Response) – wspólny termin dla określenia kompleksowego planu firmy Gemius, mającego na celu rozwiązanie problemu likwidacji Third-Party Cookie (TPC).

Przez wiele lat społeczności internetowe prosiły o poprawę poziomu prywatności, w wyniku czego producenci przeglądarek zrezygnowali z obsługi TPC. Obecnie nie są one już domyślnie obsługiwane w przeglądarkach Safari, Microsoft Edge i Mozilla Firefox. Google Chrome zapowiada zakończenie ich obsługi do końca trzeciego kwartału 2024 roku. Głównym celem JAR jest umożliwienie ciągłości badań Mediapanel w świecie bez TPC.

JAR składa się z 2 metod:

1. **Cookie Matching (CM)** - odtworzenie panelu cookie,
2. **Browsers Number (BN)** - przywracanie informacji o liczbie przeglądarek na podstawie danych site-centric.

Cookie Matching (CM) - metoda oceny prawdopodobieństwa, że pliki First-Party Cookie (FPC) pochodzące z różnych domen, które obserwujemy jako oddzielne, w rzeczywistości pochodzą z jednego urządzenia, a następnie odpowiedniego łączenia ich w odpowiednie grupy.

Prawdopodobieństwo, że dwa FPC pochodzą z jednego urządzenia jest obliczane przez klasyfikator uczenia maszynowego, który posiada poniższą zasadę działania:

- jeśli dwa identyfikatory są często obserwowane w tym samym miejscu (tj. pod tym samym adresem IP) w tym samym czasie, prawdopodobieństwo, że pochodzą z jednego urządzenia wzrasta,
- natomiast jeśli dwa identyfikatory są często obserwowane w różnych miejscach w tym samym czasie, prawdopodobieństwo, że pochodzą z jednego urządzenia maleje.

Na podstawie tego prawdopodobieństwa łączymy FPC w grupy zwane „communities", które są odpowiednikami Third-Party Cookie. Próbujemy odtworzyć wszystkie TPC na danym rynku, jednak tym, co rzeczywiście leży w zakresie naszego zainteresowania jest nasz Cookie Panel, dlatego aby zaprezentować profil społeczno-demograficzny zachowujemy tylko te communities, które mają przydzielony kwestionariusz naszego badania, a więc communities ze znaną demografią.

Browsers Number (BN) - zestaw metod obliczania na podstawie danych site-centric szacunkowej liczby przeglądarek (do tej pory nazywanych EC) dla domen, grup oraz węzła łączącego cały audytowany ruch internetowy, bez użycia plików third-party cookie (TPC). BN wykorzystuje wyłącznie dane dostępne po wycofaniu TPC i ma zastąpić obecnie wykorzystywane metody EC Global i EC Standard, które opierają

się na TPC. Źródłami danych wykorzystywanymi przez BN zamiast TPC są: first party cookie (FPC) i adresy IP. Obecnie wykorzystywana metoda -- wymaga reprezentatywnej próbki plików cookie, tak zwanych "dobrych plików cookie", których charakterystyka jest ekstrapolowana na wszystkie pliki cookie (przy użyciu stosunku liczby dobrych cookie do liczby odsłon wykonanych przez te cookie). BN również potrzebuje reprezentatywnej grupy, której charakterystyka byłaby przekładana na cały ruch, ale tym razem tę rolę pełnią reprezentatywne grupy FPC i adresów IP.

Metody działające w ramach BN są zaprojektowane i zoptymalizowane dla kilku wymiarów:

- dwóch dotyczących czasu - dnia i pseudomiesiąca oraz
- dwóch dotyczących obiektu zainteresowania - pojedynczej domeny i wielu domen (w tym węzła całego audytowanego internetu).

W sumie istnieją więc 4 kombinacje, a więc 4 nieco różne, ale dobrze uzasadnione sposoby podejścia do wyszukiwania reprezentatywnych grup danych

Znalezienie takich reprezentatywnych podgrup jest sercem BN i najcenniejszym know-how Gemius.

3. Informacje ogólne oraz schemat badania Mediapanel

Mediapanel jest badaniem służącym pomiarowi audytorium online. Naszym celem jest dostarczanie najlepszych, dostępnych w jak najkrótszym czasie oraz najbardziej przydatnych wyników na temat ogółu internautów odwiedzających witryny, aplikacje i oglądających materiały streamowe. Teraz, dzięki danym pozyskiwanym z nowego panelu Single Source (jednoźródłowego), dostarczamy rynkowi również wyniki cross-mediowe, uwzględniające informacje o poziomie konsumpcji telewizji i radia.

Obecnie badanie Mediapanel opiera się na czterech filarach:

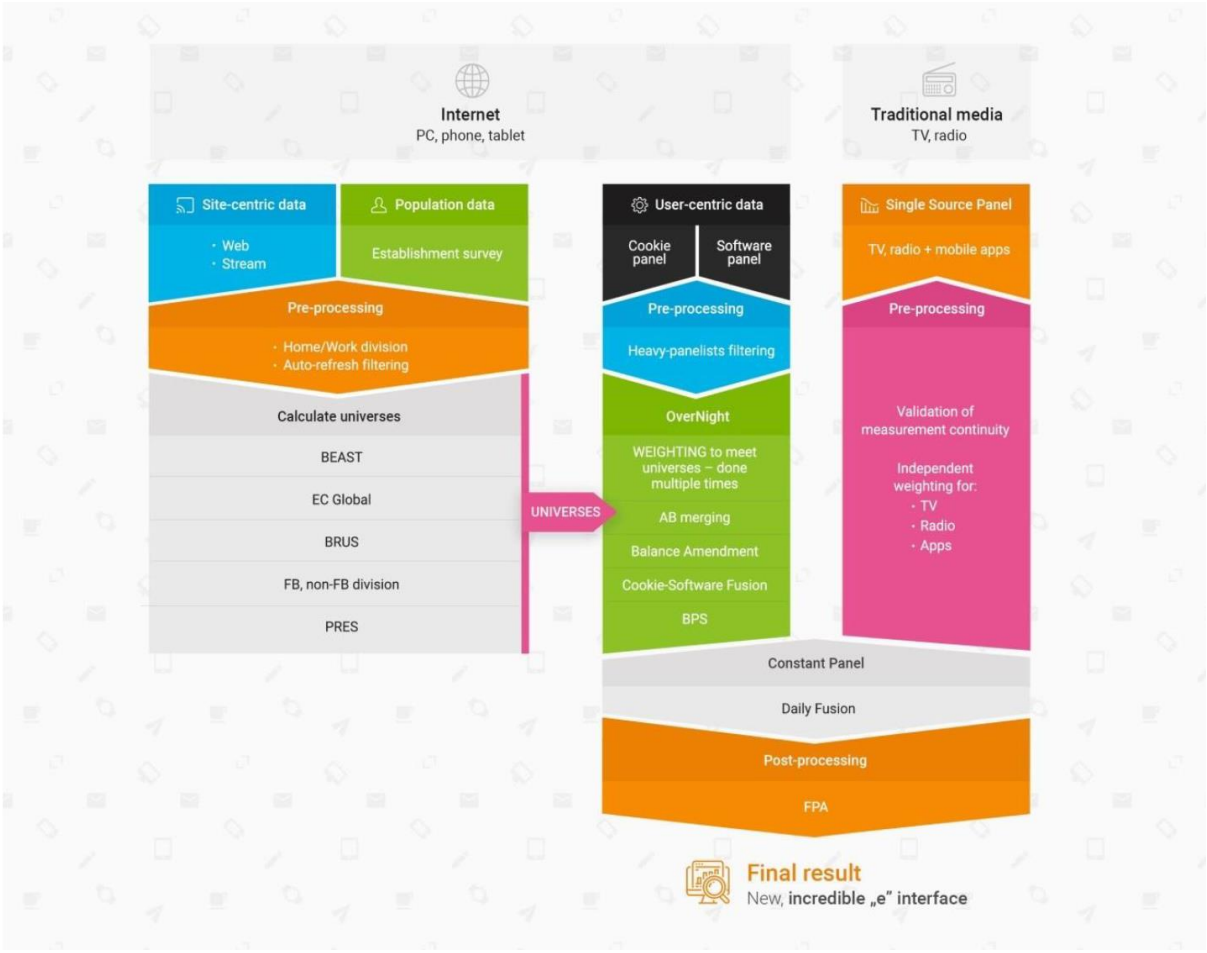
1. Pierwszym filarem są oczywiście źródła danych. Jest ich wiele, przy czym każde z nich posiada inny zakres danych, własną charakterystykę, wady i zalety, co najważniejsze jednak – dostarczają informacje o różnych częściach obrazu całego rynku.

2. Drugim filarem są metodologie OverNight, wypracowane na bazie wielu lat doświadczeń i ciągłego rozwoju procesu łączenia danych z różnych źródeł w jeden spójny obraz.

3. Trzecim filarem jest Constant Panel, który jest metodologią, dzięki której wyniki OverNight są spójne i stabilne w czasie, my zaś zyskujemy dostęp do danych z dowolnych okresów.

4. Ostatnim, choć niemniej istotnym filarem jest panel Single Source. To nasze najnowsze źródło danych, które dostarczyło nam tak wiele informacji i otworzyło tak liczne możliwości, że zasługuje na to, aby wymienić je osobno. To przepustka do analizy świata mediów innych niż internet (a więc telewizji i radia), która całkowicie redefiniuje zakres i znaczenie badania Mediapanel.

Poniżej prezentujemy graficzny opis badania Mediapanel, obrazujący połączenia i zależności pomiędzy poszczególnymi metodologiami oraz schemat przepływu danych.



Na początku procesu gromadzimy wszystkie źródła danych - a więc pomiar site-centric, dane o populacji, dane user-centric (zarówno z Cookie Panelu, jak i Software Panelu) oraz dane z panelu **Single Source**, które – jak wierzymy – **stanowią przyszłość cross-mediowego pomiaru rynku**.

Kolejnym krokiem są procesy związane ze wstępnym i głównym przetwarzaniem danych OverNight. Na tym etapie wykorzystywana jest większość metodologii wypracowanych przez Gemius, opisanych wcześniej w niniejszym dokumencie.

Następnie wyniki OverNight są za pomocą codziennej fuzji przekształcane w Constant Panel. Metodologia CP może wydawać się skomplikowana, jednak jej główny cel jest prosty: utworzenie stabilnej wersji panelu BPS z niskim, kontrolowanym poziomem błędu, w długich okresach.

Ostatnim etapem procesu jest końcowe przetwarzanie danych z zastosowaniem algorytmu Fill Panelist Activity (a także HPF) oraz prezentacja wyników w nowym, intuicyjnym interfejsie online - mediapanel.gemius.com.

Gemius S.A.
ul. Domaniewska 48
02-672 Warszawa, Polska

Kontakt:
mediapanel@gemius.com
+ 48 22 390 90 90
+ 48 22 378 30 50