Privacy-Preserving Audience Measurement in Practice – Opportunities and Challenges

Steffen Passmann
INFOnline GmbH
passmann@infonline.de

Anne Lauber-Roensberg
TU Dresden
anne.lauber@tu-dresden.de

Thorsten Strufe
TU Dresden
thorsten.strufe@tu-dresden.de

***Abstract*—The current practices of Web analytics and independent audience measurement are under legal and societal scrutiny, and the implemented and currently suggested approaches are either impractical, or most likely illegal under the upcoming General Data Protection Regulations of the European Union. While local solutions may achieve compliance for analytics, audience measurement inherently requires an independent third party for the verification of claimed audiences – a special challenge under the GDPR. We hence suggest to move the data processing from the measurement provider to the browser and to submit only unidentified aggregates, possibly over anonymization services, for mere counting. Our solution, though work in progress, hence achieves reliable verification but prevents identifiability, and thus ensures the users' privacy.**

I. Introduction

Webanalytic and Audience Measurement concern every Internet user. Almost every Web site and mobile app apply some sort of tracking, especially when their providers participate in the advertising market.

Meanwhile, privacy on the Internet is becoming increasingly important for users and society as a whole. In the field of legislation, the legal framework will change in the European Union in 2018, as the EU General Data Protection Regulation (GDPR)[1] will be applicable as of 25 May 2018. At the same time, the so-called ePrivacy Regulation shall enter into force, which will provide special provisions for data processing in electronic communications.

This does not only apply to the EU: The FTC investigated Online Behavioral Advertising and retargeting in the US and published quite explicit reports[2],[3], suggesting it may be legal, given explicit protection of the users' privacy.

This challenges especially Web analytics and Internet audience measurement. In an eco system where publishers generate revenues from advertisers for showing paid, alongside their editorial content to their users, both publishers and advertisers need accurate and reliable data, to maximize the alignment of content with the interest of expected audiences. Advertisers need abilities to assess their reach. Publishers naturally want to document the size and characteristics of their users as reliably as possible, also to highlight their attractiveness for audiences and advertisement. Web analytics at the publishers may be legally possible with entirely local solutions, like for instance piwik[5]. Verifiable audience measurement, however, requires an additional party for *"tracking"*, with no original interest in the data, just to guarantee independent verification of the claimed numbers of visits at the sites of the different publishers. This is in a natural conflict with the privacy of the users: it not only discloses the mere existence to third parties, which already is explicitly prohibited by the GDPR unless the third party is crucial for the provisioning of the service requested by the user, but worse, it actually provides this third party with quite extensive behavioral, identifiable data.

Several ideas for analytics, ad targeting, and audience measurement have been published. None, however, are practical and will be legal, to the best of our understanding. Any data processing (even the collection) of identifiable (even pseudonymous) data without explicit *informed* consent from users is prohibited by the GDP and ePrivacy regulations. Technically, the recent developments indicate that tracking by third parties, commonly implemented by third-party cookies, will be prevented by browsers per default in the future. Sophisticated cryptographic solutions [1], [2], [3], [4], however, are not realistic in practice, as audience measurement requires the overhead for publishers and users to be minimal: Experience shows that the latter are not willing to install complex software, like browser add-ons, at scale, to provide functionality to a third party, without explicit benefit.

We hence suggest to split audience understanding and verification into two parts: Profiling and attribution can be achieved by extensive surveys (*panels*), with informed consent of the users, as it's already done, today. Audience measurement, however, shall be implemented in a distributed fashion, processing all identifiable information at the browsers and under the control of the users entirely, to submit only anonymized aggregates to the independent verifier for counting.

II. How Audience Measurement Works Today

We distinguish between *Web pages*, content that is provided to a browser upon a single request, and *Web sites*, the collection of pages that are offered within the context of a product, brand, or company. The first page of a site is commonly the home page or landing page, and all pages within a site commonly share their DNS parent domain.

[1] Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

[2] http://news.cnet.com/2102-1023-871654.html

[3] www.ftc.gov/opa/2009/02/behavad.shtm.

[4] https://piwik.org

978-1-5386-0683-4/17/$31.00 ©2017 IEEE

Web analytics naturally takes a site-centric view and analyses the audiences at the pages of a single site. Its aim is to help understand access and navigation to a Web site, and optimize the Web usage. Given its implementation at and by the content provider, and a sufficient anonymization of the trace a user is creating, local Web analytics may be legal also under the GDP and ePrivacy regulations[5].

Meanwhile *audience measurement* takes a client-centric view and reports the Web usage across different sites within (or even across) markets, and facilitates comparison and ranking using accepted and standardized metrics. It is segmented into domestic markets: companies within a national market form an *ABC* (Audit Bureau of Circulation/Certification), or, for publishing exclusively on the Web an *IAB* (Interactive Advertising Bureau) that define measures and currencies for advertising[6], and publish ranks and statistical facts[7]. The ABC's and IAB's are organized in international federations[8] for standardization.

Some markets require all publishers to use the same audience measurement system, to allow for independent and comparable counting and tracking of clients across all participating Web sites (potentially also across the various devices of the user). To ensure comparability, the ABC's not only verify the measured data for plausibility, but also check the technical implementation of the measurement systems that are integrated by the publishers into their Web pages. Audience measurement hence also naturally involves a third party that has to process data of the users, which is not anonymous per se.

A. Measurement Types and Metrics

There are two fundamental approaches to evaluate the online-market: Panels and full evaluations. *Panels* comprise of remunerated user groups who install additional tracking software, like a browser add-on, which regularly submits their browser history, and who complete extensive questionnaires. For a *full evaluation* (also: census), all publishers within a market integrate scripts into their Web sites, pages and Web applications, that submit measurement information about their viewers to an independent third party [5].

A major disadvantage of the full evaluation is its obvious complexity, necessity for cooperation of all market participants, and, correspondingly, its cost. It is also not clear how full evaluations by third parties can be compliant with the GDPR and ePrivacy Regulation, today. Panels, however, suffer from sampling difficulties and hence naturally tend to be biased, misrepresenting little-used or little-known content[5]. Many markets thus choose the full evaluation, enriching the data by the results from panels. This also allows for a detailed profiling of measurements during full evaluation, informed by the comprehensive data from the panels.

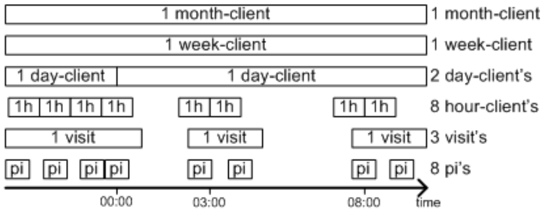


Fig. 1. Relationships between the metrics on a site of a publisher

The ABC's define different *measurement metrics* to be relevant, most importantly the page impressions, visits, and clients (cf. Fig. 1 for examples).

- *Page impressions* are the count of each single access to a page of a Web site. Repeated accesses by the same user are counted separately.
- *Visits* denote sequences of consecutive accesses to a Web site, confined to an activity period: A visit starts when a user creates a first impression on a page within the site. Each additional page impression generated by the user within the site is then assigned to the same visit, even if the user leaves to another site for an intermediate period and returns subsequently. Inactivity at the measured site beyond a certain time threshold, usually 30 minutes, completes the visit.
- *Clients* (also: visitors or unique browsers) count the number of different visits during specified time periods (with common reference periods being the hour of day, day of the week, week, or month). The metric hence counts each set of accesses of a unique client within the interval of reference (for instance Tuesday clients, April clients, or 9 o'clock clients). Unique clients are commonly identified by their browser and its specific characteristics, and a single visit can count towards several clients (for the first and second day in Fig. 1, for instance).

While other metrics, like ”usetime”, ”awareness” and statistically modeled users are defined, they are usually based on the described metrics in the context of audience measurement.

B. Profiling and User Sociodemography

The task of profiling is to describe the characteristics of the stored usage of each client as precisely as possible. The accuracy of a profile increases with the measurement scope. For this reason, it is useful to consider the use of more than a single Web site (cross-site-usage). An example of the characteristics taken into account for a profile is a list of Web pages used by the client with an indication of the intensity (measured, e.g., in page impressions).

External information may allow linking of the profiles to specific sociodemographic characteristics. This external information may be collected in panels, additional surveys, or bought from third party data brokers.

Sociodemographic characteristics of primary interest[9] are the *gender, age* as well as *education, profession, income*,

[5] Outsourcing it to external services without given informed consent by the users may actually be illegal, as it is not strictly necessary for service provision and entails processing of personally identifiable information by third parties.

[6] http://ausweisung.ivw-online.de/

[7] www.agof.de/studien/digital-facts/studiensteckbrief/

[8] IFABC www.ifabc.org, IAB www.iab.com

[9] several others are collected in panels and available to enrich the profiles, cf.: https://support.google.com/adwords/answer/2580383.

and the household size [6]. Considering advertising today, however, we observe a strong emphasis on gender and age of the users, as potential customers[10], while more precise characteristics are way more uncommon than anecdotes suggest.

III. The Issues

The current measuring systems collect information through sensors that are implemented in JavaScript and integrated into each measured Web page, which, upon each page impression submit information to a centralized server (cf. Fig. 2).

While the sensors are comparatively simple scripts, they retrieve as much characteristics as possible from the user's browser, OS, and device, to ensure that unique clients can be identified and the visits- and clients-metrics, as described above, can actually be measured[11,12].

In addition to the extensive data that is transmitted by the sensors, the receiving server extracts additional, potentially identifying information from the underlying protocol (IP address, linked information), to enrich the dataset and increase the accuracy of the metrics. The calculation of the metrics then is performed at the server, which combines all accesses of the different unique users, and attributes them to the visits and clients for the different Web pages and Web sites.

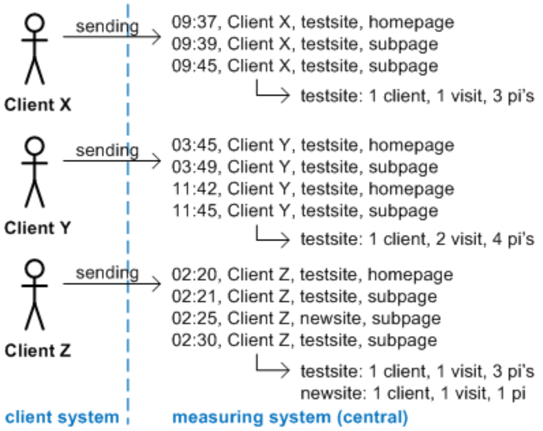

sending
09:37, Client X, testsite, homepage
09:39, Client X, testsite, subpage
09:45, Client X, testsite, subpage
testsite: 1 client, 1 visit, 3 pi's
Client X
sending
03:45, Client Y, testsite, homepage
03:49, Client Y, testsite, subpage
11:42, Client Y, testsite, homepage
11:45, Client Y, testsite, subpage
testsite: 1 client, 2 visit, 4 pi's
Client Y
sending
02:20, Client Z, testsite, homepage
02:21, Client Z, testsite, subpage
02:25, Client Z, newsite, subpage
02:30, Client Z, testsite, subpage
testsite: 1 client, 1 visit, 3 pi's
newsite: 1 client, 1 visit, 1 pi
Client Z
client system
measuring system (central)

Fig. 2. Audience measurement as implemented today

The collecting servers generate live ad-hoc data online, to give the publishers near to real-time information about the usage on their sites. Page impressions, for example, can simply be summed up for this purpose.

Calculating visits and clients requires intermediate storage of measurements, in so-called session-tables. Especially in the case of heavily frequented Web sites, this intermediate data can become very large[13].

The collected data set is stored in addition, to have a data set for later evaluations with potential additional requirements.

A. Identifying the Users

Audience measurement and Web analytics requires the systems to identify unique users to assemble navigation trails and calculate the complex metrics of visits and clients.

The traditionally easiest solution is to identify users by storing id's within their browsers, as cookies. We distinguish between first-party cookies, stored in the name of the content provider, and third party cookies, stored in the name of another entity, for instance through scripts or external links within the requested Web page. While first-party cookies suffice for Web analytics, audience measurement naturally requires third-party cookies for cross-site measurements and to safeguard the trustworthiness of results, which will no longer be available.

The cookie value stores a unique entry, which is used as part of a client-id. It often consists of a random number and additional values, like timestamps, to increase the reliability of identification. Mobile devices share unique advertising identifiers by default, which in similarity to third-party cookies provide identifiability across sites. It is hence used when available, as it conveniently allows exact tracking of clients over all participating applications of all publishers.

In some cases, no cookies or ad identifiers are available, for instance when the user's browser blocks or regularly deletes cookies. The identification then commonly takes into account other suitable characteristics of browser, OS, and device, like the browser's UserAgent, or additional JavaScript fingerprints, as well as the IP address. The latter is commonly truncated, as it is considered directly personally identifiable information already today. The challenge of identifying unique clients hence is already addressed by correlating other characteristics, to reduce uncertainty and increase accuracy. This, nevertheless, is not always possible, and the remaining uncertainty hence is reflected in all calculated metrics[14].

So calculating the complex metrics and assembling navigation trails requires identification of unique clients. The utilized client identifiers, however, lose their anonymity through the uniqueness of the clients' Internet usage, and have to be regarded as pseudonyms [7], which complicates data processing under the GDP and ePrivacy regulations.

This situation is aggravated by the way that profiles are enriched with sociodemographic information, today: Even in the case that no sociodemographic data is available from panels, the user behavior of the clients is clustered to similar profiles, and representative users are chosen and asked to complete surveys (potentially in return for remuneration or participation in lotteries). The aim of this approach is to provide sociodemographic information for a sufficiently large number of clients. The acquired sociodemographic data then can be associated with the other clients. Once behavioral templates are extracted, they can quickly be identified in traces online, and the users hence be enriched with likely sociodemographic information. While this isn't immediately necessary for audience measurement, this approach is commonly used for retargeting.

[10] https://support.google.com/analytics/answer/2799357

[11] https://developers.google.com/analytics/devguides/platform/

[12] www.infonline.de/en/unternehmen/leistungsspektrum/

[13] https://developers.google.com/analytics/devguides/collection/analyticsjs/limits-quotas

[14] https://developers.google.com/analytics/devguides/collection/analyticsjs/cookies-user-id

B. Legal Aspects of Current Measurements

As far as audience measurement and analytics involve the processing of personal data, they have to comply with data protection laws. The term *personal data* means any data relating to an identified or identifiable person.

Even when information does not directly allow for the identification of a person, it is still considered as personal data, when there is a certain probability that the data processor may identify the data subject with the assistance of other parties, e.g. the Internet service provider, the publisher, or any other entity. So according to the jurisprudence of the European Court of Justice, even a dynamic IP address registered by a website provider falls within the ambit of this definition, when the website provider can legally obtain additional data on the user from the Internet service provider, e.g. when taking legal action against cyber attacks[15]. It hence also includes the profile [8] and user-id, and of course all records linked to any of the information above. In contrast, data protection laws do not cover the processing of securely *anonymous* data.

The present European legal framework for data processing is set by the Data Protection Directives, especially the ePrivacy Directive[16] . According to Art. 5 (3) ePrivacy-Directive, the storing and processing of information in the user's terminal equipment, e.g. cookies and other tracking techniques, is only allowed, if the user has given his or her consent after having been provided with clear and comprehensive information about the immutable purpose of the processing. Only in those cases, when the data processing is strictly necessary for providing the service requested by the user, there is no need for an informed consent. So under the current legal regime, any collection of personal data for the purpose of audience measurement on principle requires the data subject's informed consent. When the data are processed by an external organization, there may be additional requirements for lawful data processing, such as data processing agreements.

The future legal framework within the European Union is not entirely clear. The proposal for an ePrivacy-Regulation tabled by the EU Commission permits data processing necessary for Web audience measurement, but only if it is carried out by the provider of the service requested by the end-user (Art. 8 (1) (d)). This provision has attracted a lot of criticism for its imprecise wording. The advisory body representing European data protection authorities emphasized that the provision will have to be framed more precisely in order to clarify that the provision only applies to usage analytics necessary for the analysis of the performance of the service, but does not permit any profiling[1]9. So the final version of this provision will most likely not cover data processing for any other purposes, such as advertising. So judging from today, under the GDPR and the EU-Commission's proposal for an ePrivacy-Regulation, data processing for Web analytics and audience measurement will only be permitted upon informed consent, unless this *(a)* is either necessary for transmitting the electronic communication or for providing the service requested by the user (which remains hard to argue), *(b)* if the end-user has given his or her consent accordingly (Art. 8 of the proposed ePrivacy-Regulation), or *(c)* the data is anonymized and not reasonably relatable to any natural person.

The European legislator concedes that given the ubiquitous use of cookies and other tracking techniques, end-users are increasingly requested to provide consent. In order to prevent an overload, the EU Commission proposes that users should be able to provide consent by using technical means, such as privacy settings of a browser, where they also request privacy by default. An unofficial draft of the Proposal, which leaked in November 2016, suggested establishing an obligation to provide for a tracking ban by default in the browser settings, in this sense. And even though this proposition is not reflected by the EU Commission's proposal, it will be an important issue for publishers and advertisers to provide users with privacy-preserving options in order to enhance user acceptance of audience measurements and thus to motivate users not to choose too restrictive no-tracking-options.

IV. State Of The Art

Different approaches to implement privacy preserving analytics have been suggested in literature.

Several papers propose systems providing results with differential privacy [9], when querying centralized databases. They support different types of queries, like linear and histogram queries [10], [11] or even implement general programs [12]. These and similar approaches require the data to be collected and processed at a trusted third party, first. While ABC's are trusted third parties between publishers and advertisers, they are commonly not trusted by visitors to Web pages, and hence this assumption is not realistic in the given environment.

Randomized response [13] offers subjects to lie with some probability. It was implemented quite prominently by google [3] and in other approaches [4]. This may hide the count of impressions, it gives away the specific pages that are visited.

Privad [14] and Adnostic [1] are probably the closest to our requirements for audience measurement with privacy. They process behavioral cross-site data within the browser, and amongst adding parties require the users to install browser extensions. Experience shows that users are not willing to change any of their settings, let alone install additional software, for the benefit of the advertisement industry, and hence we allege that none of the approaches that have been suggested so far is applicable to audience measurement in reality.

V. Towards Privacy in Audience Measurement

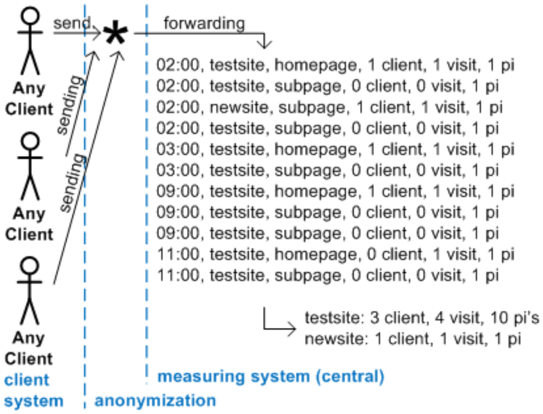
Analyzing issues and proposals, we suggest to follow the path of *(a)* separating audience understanding (profiling) from audience measurement completely, *(b)* decentralizing the processing of personally identifiable information to the realms of the users, within their browsers, and *(c)* keeping the local processing as simple as possible (cf. Fig. 3).

[15] ECJ, Judgement C-582/14 of 10 October 2016 – Breyer/Germany.

[16] Directive 2002/58, as amended by Directive 2009/136/EC.

[17] Art. 29 Data Protection Working Party, Opinion 01/2017 on the Proposed Regulation for ePrivacy Regulation (2002/58/EC), adopted in April 2017.

Acknowledging the fact that the algorithms for processing will have to be provided as scripts within the Web pages, as users do not install browser add-ons, we restrict them to very simple aggregation, and as data is not available across sites, we restrict the aggregates to the separate sites. The respective sensors are transmitted with the requested Web pages, and hence are naturally open source and available for auditing by interested users. Result submission will still be attributable to IP addresses at the collecting servers. We are partnering with an audience measurement institution that will gladly refrain from collecting IP addresses, but this statement of course offers weak protection only, and we hence propose to anonymize the aggregates through layer-4 proxies (set up and provided either at the users affiliation, or trusted institutions like universities or DPA's) in addition.



```
02:00, testsite, homepage, 1 client, 1 visit, 1 pi
02:00, testsite, subpage, 0 client, 0 visit, 1 pi
02:00, newsite, subpage, 1 client, 1 visit, 1 pi
02:00, testsite, subpage, 0 client, 0 visit, 1 pi
03:00, testsite, homepage, 1 client, 1 visit, 1 pi
03:00, testsite, subpage, 0 client, 0 visit, 1 pi
09:00, testsite, homepage, 1 client, 1 visit, 1 pi
09:00, testsite, subpage, 0 client, 0 visit, 1 pi
09:00, testsite, subpage, 0 client, 0 visit, 1 pi
11:00, testsite, homepage, 0 client, 1 visit, 1 pi
11:00, testsite, subpage, 0 client, 0 visit, 1 pi

testsite: 3 client, 4 visit, 10 pi's
newsite: 1 client, 1 visit, 1 pi
```

Fig. 3. Simplified sketch of the proposed architecture

Browsers allow for access and storage of site-specific data, so we extend the sensors to store the access patterns of the user to the currently measured site on the client system. The stored data simply contains the requested page together with a time stamp. The sensors are only invoked upon accessing a page. They then check for the data stored during the last access. Setting the current time stamp relative to the time stamp of the previous contact suffices to update all described metrics.

Local aggregation is hence executed continuously and submitted to the servers for counting upon return to the site, when the interval of the last visit has expired. The submitted data simply contains *(a)* a generalized time stamp, *(b)* the name of the site, *(c)* the page the user visited first, and *(d)* the aggregates as stated above.

The sensor finally sends the aggregates to the measuring system, which can simply add the metrics towards the different visited sites directly (cf. example in Fig 3). Identifying the users no longer is necessary in this case, as the metrics that require linking of records are already calculated at the client. This also applies to the IP address, so the submission can be performed via proxies, or even through anonymizing mix cascades or TOR. Aggregation and anonymization also reopen the venue to applying randomized response, and the sensors may increase their protection from potential profiling by submitting only a specified percentage of aggregates, which could be perturbed according to statistical models.

We implemented the overall system as described above using JavaScript for the sensors and nginx at the server, which we are currently testing in restricted setups.

VI. Opportunities and Challenges

Our approach, albeit simple, comes with several opportunities and challenges that are not quite obvious.

A. Opportunities

Let's consider the different stakeholders of the play. Advertisers and publishers do not perceive any change compared to audience measurement today, and can still rely on the independence and quality of the measured data.

The users benefit from extensively increased privacy that is gained entirely transparently without their interference, although they may be left with some threats that we will discuss, below.

The providers of audience measurement services initially seem to loose, as they no longer have access to vast collections of behavioral data. This, however, comes with two distinct advantages: first, they save significant storage and computation resources. These are necessary today to retain session tables and to compute the metrics – they will now be provided by and shared between the users and their browsers. Second, the providers no longer need to worry about storing and protecting highly sensitive data, the loss of which reportedly has caused significant financial and reputation damages[18,19]. Given the reduced privacy concerns the providers could also benefit of the increased acceptance and hence participation, and reductions in active countermeasures by the users.

Finally, and probably most importantly, we allege that approaches of this flavor will be the only solutions that will be both practical and legal as of May 2018 in the EU.

B. Challenges

Some challenges persist that remain to be addressed, which we group into security, data, and legal challenges below.

With respect to *adversaries*, the system provides only limited protection from collusion: *(a)* colluding measurement providers and proxies may aim at re-identifying users. In addition to the fact that the providers have a reputation, and in case of being brought to court also money to loose, they also only have little learn: linking the IP addresses to the aggregates still does not provide any detailed usage information. *(b)* colluding measurement providers and publishers could potentially link very characteristic aggregates to entries in the Web server log files or their local Web analytics systems. Note, however, that European law already requires providers to truncate IP addresses to 24 bits before storage and any processing, and hence the risk of re-identifying natural persons

[18] www.alstonprivacy.com/comscore-reaches-14-million-settlement-in-electronic-privacy-class-action

[19] https://techcrunch.com/2016/10/06/report-verizon-wants-1-billion-discount-after-yahoo-privacy-concerns/

through this attack remains low. Randomized response could make such attack at least difficult, if not impossible, when the access patterns to a site are sufficiently perturbed to prevent re-identification in the logs. *(c)* collusion between the three parties: measurement provider, proxy, and publisher would yield the possibility to identify users and link their usage of the Web sites of all colluding publishers. This, however, would be much easier for the colluding publishers without the other parties by ignoring their duty to pseudonymize, and aggregating their corresponding Web analytics logs.

There are many different types of *fraud traffic* from human-click fraud to automated retrieval. Derived from today's trends towards fraud detection solutions, the problem will need outsourcing into other specialized products.

Expressiveness and quality of the measured data is another challenge: although profiling is not in the primary interest of audience measurement providers, this added data of course represents a significant value to their customers. Linking profiles to audiences explicitly is made immediately impossible by our approach (this, after all, is one of the primary requirements following from the legal situation).

We do suggest to separate the task of assessing sociodemographics in panels from the census. We allege that, given convincing results from the panels the data collected at remunerated volunteers can easily by extrapolated to the measured audiences. Our system is also inherently compatible to the panels: The volunteers traditionally install browser add-ons in any case, which can enrich the data of the sensors with cross-site and profiling information. The sensors for this purpose could also be extended to store some historical usage data together with the last access, separately for each visited site locally. After getting informed consent and an opt-in, this data could be linked to the user-ID by the browser extension and submitted as part of the panel.

Another problem arises with *multi-origin* sites, which contain pages relating to multiple domain names, as the sensors will not be able to store and access data relating to domains other than the one at which they have been accessed. This is a general problem for audience measurement with disabled third-party cookies. Several solutions have been suggested, like for instance creating expressive JavaScript fingerprints, and external session-services, or browser extensions. None of these meet our requirements of privacy-preservation and transparency for users and providers. The best current solution we can imagine is to implement URL-rewriting to integrate session-IDs within the pages of the multi-origin site, to achieve linkability within this realm. Even ignoring this fact would not cause severe loss in data quality in the suggested system, as for multi-origin sites the approximation error for visits and clients can easily be determined as the number of domains that the site spans.

There also remain some *legal* challenges. Audience measurement requires processing of Web usage data by independent, third parties. It remains unclear, which processing exactly will be covered by the legislation. Even our suggestion entails processing that isn't strictly necessary for the service provision, and data is transmitted to third parties. We allege, however, that aggregation and network address anonymization minimizes the data collected at the measurement provider to the necessary minimum. The anonymizing layer-4 proxies, which represent a new party in the service, do not learn anything from the https submission but participation of a user behind an IP address in the measurement system. A solution to prevent that third parties learn anything about the fact that an IP address is actively accessing Web pages it is well feasible to place the proxies directly at the publishers, which learn about this access within their Web servers in any case.

VII. Summary and Outlook

In this paper, we described the service of Internet audience measurement and highlighted the current implementations and privacy concerns they raise. We then suggested a work-in-progress idea that allows to measure the essential metrics, while preserving the privacy of the users. The approach collects and aggregates data within the browser of the user, anonymizes it and then transmits it to a measurement provider, where it subsequently is only accumulated. Meeting the legal requirements, it still provides results without any loss in quality as compared to audience measurement today. The idea is illustrated using the three known metrics of page impressions, visits, and clients, but it can reasonably be adapted to additional aggregate metrics.

We are currently analyzing existing data to quantify the risk of re-identication in the log-files of typical Web sites, to potentially parametrize perturbation and randomized response accordingly. We are also in the process of implementing and deploying the suggested system at a large scale in the German market, to assess its feasibility and compare the results.

Acknowledgment

This work in parts was supported by INFOnline GmbH and DFG, through RTG RoSI and the excellence cluster CFAED.

References

[1] V. Toubiana *et al.*, "Adnostic: Privacy preserving targeted advertising," in *NDSS*, 2010.

[2] F. Günther, M. Manulis, and T. Strufe, "Cryptographic Treatment of Private User Profiles," in *FC/RLCPS*, 2011.

[3] E. Úlfar *et al.*, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *ACM CCS*, 2014.

[4] D. L. Quoc *et al.*, "Privapprox: Privacy-preserving stream analytics," in *Usenix ATC*, 2017.

[5] L. Berekhoven *et al.*, *Marktforschung*. Gabler, 2009.

[6] P. Shukla, M. Banerjee, and P. T. Adidam, "The moderating influence of socio-demographic factors," *Jnl of Consumer Behaviour*, 2013.

[7] S. Zahoor *et al.*, "Uniqueness in user behavior," in *ICICT*, 2016.

[8] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Oakland*, 2008.

[9] C. Dwork, "Differential privacy," in *ICALP*, 2006.

[10] C. Li *et al.*, "Optimizing linear counting queries under differential privacy," in *Symposium on Principles of Database Systems*, 2010.

[11] M. Hay *et al.*, "Optimizing linear counting queries under differential privacy," in *VLDB*, 2010.

[12] A. Haeberlen, B. C. Pierce, and A. Narayanan, "Differential privacy under fire," in *Usenix Security*, 2011.

[13] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Jnl of the American Statistical Association*, 1965.

[14] S. Guha *et al.*, "Serving ads from localhost for performance, privacy, and profit," in *HotNets*, 2009.