Culture Unbound
Journal of Current Cultural Research

Studying Ad Targeting with Digital Methods: The Case of Spotify[i]

By Roger Mähler & Patrick Vonderau

Introduction

Online advertising is a matter of public interest. Ten years ago, many of us would not have cared much about how Facebook, Google or Spotify place ads on their sites, and how they target particular constituencies of buyers or voters. This has changed since 2008, when programmatic advertising was introduced, an automated procedure of ad buying. Programmatic advertising largely lacks human oversight, making advertising an algorithm-driven business. The procedure enabled $100,000 worth of ads being placed during the 2016 U.S. presidential election by inauthentic accounts that appeared to be affiliated with Russia. It allows Facebook ad buyers to define target groups such as "Jew Hater," "Second Amendment," "Hitler did nothing wrong," or "Nazi party," which in turn makes it possible to feed such groups with divisive messages. Platforms have taken an active role in spreading misinformation through advertising. They also monitor user behavior on a large scale. Facebook, for instance, obtains detailed dossiers from commercial data brokers about users' offline lives, and users have limited means to opt out of their data being used (Angwin et al 2016; Madrigal 2017; Meyer 2017).

Spotify, the Swedish music streaming service, has a less controversial reputation. Introducing programmatic ad buying in 2015, however, the company has made no secret of its abilities to collect data on user behavior. In November 2016, Spotify launched a global outdoor ad campaign with ads jokingly showcasing massive aggregate data sets: "Dear 3,749 people who streamed 'It's the End of the World as We Know It' the day of the Brexit vote, hang in there" (Nudd 2016). Spotify does not just collect "an enormous amount of data on what people are listening to, where, and in what context," as one of its executives stated in public (Terdiman 2015). The company also acts as a private data broker, providing this collection of contextual data to marketers for ad targeting purposes. Spotify offers

Mähler, Roger & Patrick Vonderau "Studying Ad Targeting with Digital Methods: the Case of Spotify", *Culture Unbound*, Volume 9, issue 2, 2017: 212–221. Published by Linköping University Electronic Press: http://www.cultureunbound.ep.liu.se

"premium brands" its "extraordinarily engaged, first-party-verified audience at scale" (Spotify for Brands 2015). This offer includes "demographic targeting" as well as "content targeting" to reach users with particular habits, mindsets, and tastes that align with a pre-defined target persona. Playlists, "tailored" to specific urban activities (such as "Morning Commute") and moods (such as "Life Sucks") are combined with data on genre preferences, age and gender, geography, language, and streaming habits alongside broader interests, lifestyle, and shopping behaviors, fueled by third-party data providers.

Spotify's desktop interaction design looks very different today from what it did in 2008 or 2012. In the past, user interaction was organized around tracks and search and community-activating features, such as self made playlists. Today, Spotify's interaction design re-organizes music consumption around behaviors, feelings and moods, channelled through curated playlists and motivational messages that change six times a day. This present situation—where music has become data, and data in turn become contextual material for user profiling at scale—invites us to pause and reflect about the way songs, books, or films are now typically made accessible. How does Spotify's 'service' relate to the European Union's new General Data Protection Regulation and its provisions on profiling? What is the long-term strategy behind this massive data collection? Facebook is rumored to have an interest in acquiring Spotify, and both companies micro-target users based on their emotional states. This may have repercussions for music and beyond. As psychologist Michal Kosinski and others suggest in a paper entitled, "The Song Is You," platforms such as Spotify may strategically exploit the link between music choices and personality traits in the near future (Greenberg et al 2016). Kosinski' model for behavioral prediction is already used by Cambridge Analytica, a firm notorious for "psy-ops" electoral manipulation in support of Brexit and the Trump campaign (Grasseger and Krogerus 2017).

Although ad targeting is in the public interest, there is little reliable information on how it works. Its effects are often under- or overstated. Some see ad targeting as a means to monitor individuals, a view that overlooks that advertising's "targets" are not individuated human beings but inferred ones. Rather than being "you," targets are like you: sets of demographic, psychographic, and other data points (audience segments) aggregated via various online sites and bundled together. Others see the ad targeting in Spotify's free version as largely ineffective. Listeners complain about the lack of proper targeting, noting that some ads did not match basic data sets including age and gender, location of user IP and user language, genre preferences, and listening context. But this wrongly blames on Spotify what in effect are marketer decisions. Most brands do not micro-target their ads but instead opt for broad media reach, depending on overall ad campaign goals and disposable budget.

To study ad targeting, researchers have an inventory of tested methods at their disposal. Media industry researchers often use semi-structured qualitative interviews as direct observation is difficult in media and tech contexts. Our study of Spotify's advertising technology began with such conventional means. We spoke to Spotify, but also interviewed other key stakeholders at business organizations, such as the Internet Advertising Bureau, and companies offering programmatic services, such as Amnet and Starcom. In addition, we retrieved all available information in the public domain related to Spotify's use of such services. Yet as often is the case in media industries research, these kinds of sources often merely provide a work-around because a problem of access to verifiable data persists. Spotify does not reveal with whom the company collaborates in placing ads, for instance. In order to understand who the main actors are in this process, we opted for digital tools to complement data collection. Doing so, we followed the well-established idea to approach "the digital from the inside out, taking up methods that are already embedded in digital infrastructures and practices, and then adapting these to the purposes of social research" (Marres 2017, 84; cf. Hargittai and Sandvig 2015; Rogers 2013).

The remainder of this article gives a brief technical account of how we proceeded. The overall aim of our research was to map the infrastructure of ad serving companies in order to better understand strategies of intermediation in the current platform-based media marketplace. Platforms such as Spotify, Facebook or Google act as market makers in the digital media value chain. Given the degree "multi-layered platformization" (Hölck 2016) is currently developing, we need to have a precise understanding of intermediary strategies and of their key actors (cf. Skeggs and Yuill 2016). Preliminary results of the research are published separately (Vonderau 2017).

Ghostery and Fiddler: Simple Tools for Studying Ad Tech Infrastructure

Our data work began with opening a small program or plug-in called Ghostery in the browser while being logged in to Spotify's free desktop version. This tool allows us to track or monitor the activities of trackers related to ad programs, analytics, and other functions, and to capture those activites on a trackermap. Ghostery has an extensive, crowd-sourced database of companies that are active within the complex advertising landscape. This allows to identify and chart advertising supply chain vendors.

One limitation, however, is that Ghostery can only be used in conjunction with a Web browser (i.e. the Spotify Web player). In our case, we needed a workflow applicable for all kinds of Spotify clients, not just the Web player, but also desktop applications and Spotify's mobile clients. The purpose was thus to

find an alternative (but similar) workflow that collects Web traffic between the local machine and the remote Web resources such as, for instance, Google's subsidiary DoubleClick, and ad serving service—all in order to grab (and graph) a snapshot of the underlying ads serving infrastructure. To do this, we needed technical means to intercept and store the communication between actors and companies within the ads landscape.

There are several tools that can be used to capture network data—from generic tools such as WireShark to more specialized tools such as Charles and Fiddler that only capture http and https (encrypted) Web traffic. We used Fiddler, which is a commonly used tool in software development. On Fiddler's home page

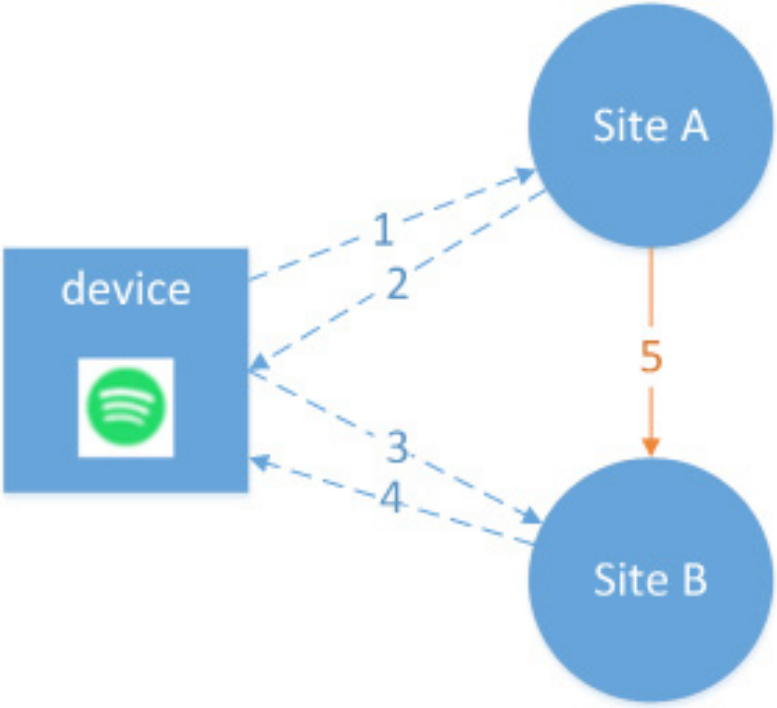

Figure 1. The Spotify client requests a resource from site A and receives a response. This response spawns a new request to a resource from Site B.

(www.telerik.com/fiddler), the tool is described as a "debugging proxy," which in essence means that the software positions itself as an intermediary layer between the client software and the internet, and in such a way that all client requests, and the associated responses, are routed through this layer. Fiddler can then store and even change or "fiddle with" these messages (hence the name). Fiddler allowed us

to capture data from any of Spotify's applications since they, in one way or another, all rely on Web (http/https) traffic. It also allowed us to capture data from other kinds of platforms (Windows, Linux, MacOS, iOS and Android)—at least as long as we were able to route Web traffic via Fiddler. A tool like WireShark can be used

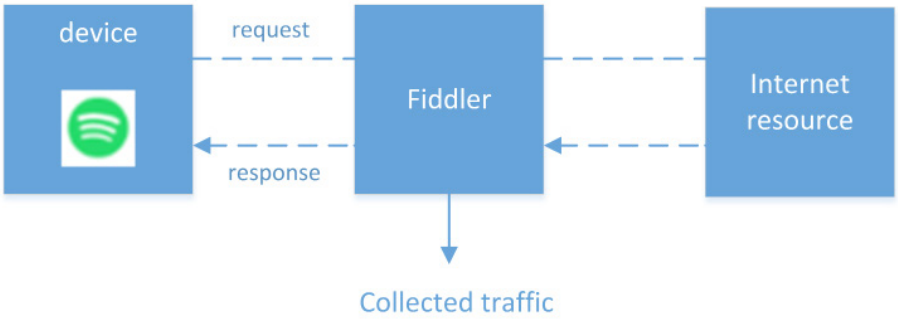

Figure 2. Fiddler acts as an intermediary between a client (e.i. Spotify Desktop application) and a Web resource such as play.spotify.com.

simultaneously to ensure that one does not lose any important non-http(s) data.

The collected data was very detailed and contained Web (HTML) documents, cookies, images, audio streams, source code, and so on. However, it had a low signal-to-noise ratio. We were basically interested in the ads-related traffic. Noise could be filtered out using meta information associated to each message, for instance, content related to presentation styles, requests for encrypted connections,

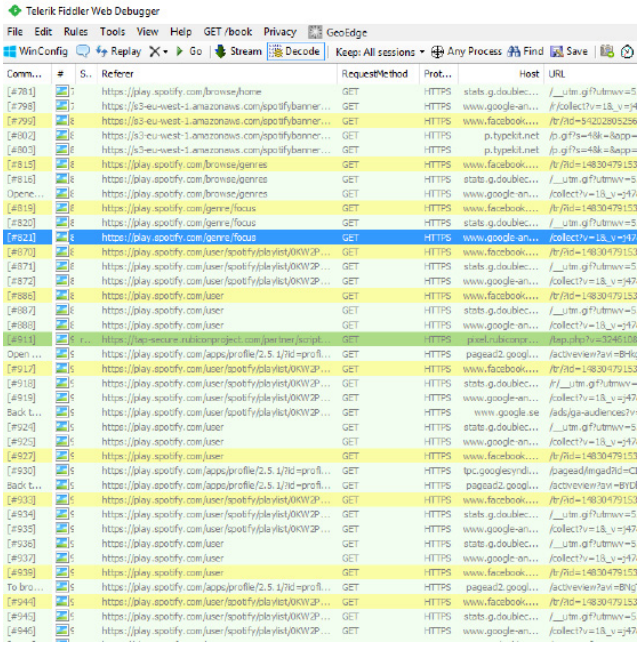

Figure 3. Sample of captured data. Every line is a request and response from the local machine to a remote machine.

or failed requests.

An hour long session with the Spotify desktop client—logged in using a Facebook account—resulted in no less than 2,391 collected requests (and associated responses). Of these requests, 1,025 were irrelevant for this research, with 1,366 requests remaining. Not all of the latter, however, were ad-related. We got 279 Web documents (html files), 691 images, 209 source code files (JavaScript), 56 text-data files (text, xml or json), 21 files, 54 redirects and 56 of unspecified type (most of them having so called "Not Changed" response code). These requests originate from 17 sites in nine different domains, and targets 71 different hosts in 41 different domains.

The semi-manual workflow used in this part of the experiment started by specifying a usage scenario and context that designated what tracks to play and actions to perform in Spotify. Before we executed our scenario we started the data capture; besides capturing Web traffic, we also recorded the entire desktop session using the open source Open Broadcaster Software. During the session we added comments at specific points to indicate when certain actions started, or when certain updates occurred in the user interface (for instance when a new track started or the leaderboard updates). This made it easier to select and analyze a subset of the traffic that corresponded to specific actions or updates.

When the usage scenario had ended (i.e. the user was logged out and the recording stopped) the data was exported to Excel for noise elimination and encoding (e.g. content type encoding, domain names, cookie identification, redirects). This was also the step where site names and domain names were translated into actors and companies, with the help of online resources such as the Ghostery database, the Thalamus database, and Cookiepedia. Ideally, this site-to-actor linking should be automated, especially since the online databases also give information of where companies are positioned within the ads ecosystem, which was vital for our investigation. The next step was to use a graph visualization tool such as Gephi to explore and analyze the data—both as a whole, or in part for specific user scenarios. The graph below shows the accumulated data exchange between sites during a 60 minute session. As is evident, a lot of noise still remains to be filtered out especially regarding content types, and parties not related to advertisements. The size of a node is proportional to number of requests to that site.

It is possible to get at cleaner graph if one looks at domains (figure 4.) instead of specific sites. For instance, "4721227.fls.doubleclick.net" and "stats.g.doubleclick.net" both belong to the same domain, called "doubleclick.net". Figure 5, in turn, goes a step further by displaying the companies that operated within each domain. In fact, the graphs become even more interesting if one only selects requests that are related to a specific user action or a system event. For instance, if one

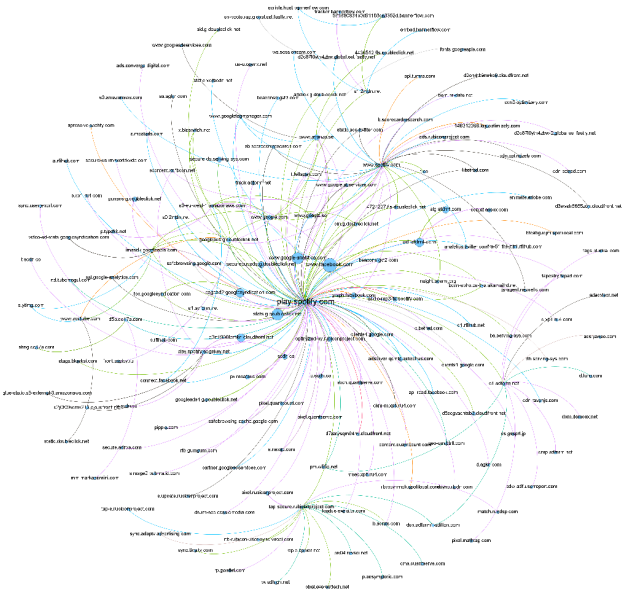

Figure 4. Graph of involved internet sites during a 60 minute session.

selects the requests occurring during one single update of Spotify's banner ads, one can create a graph specific for this update, and even schematically show the sequence of how that graph evolved. By using this straightforward, and not overly-complex workflow, one can thus get a number of insights into what is actually going on behind the scenes in the context of a specified Spotify usage scenario.

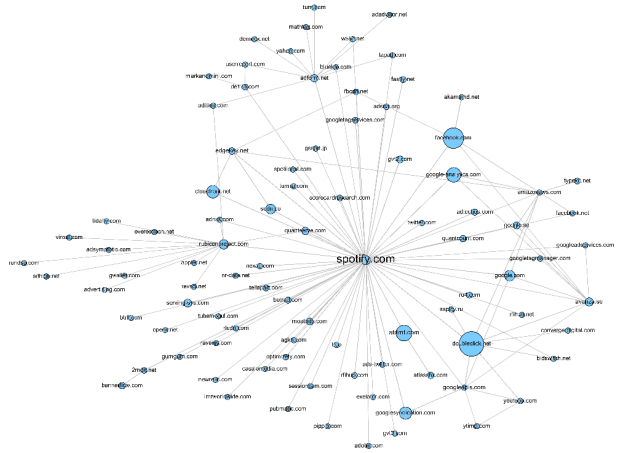

Figure 5. Domain graph of involved internet sites during a 60 minute session.

As is apparent, the collected data is rich, which enables one to explore involved parties and messages sent between parties. With this data one could, for example,

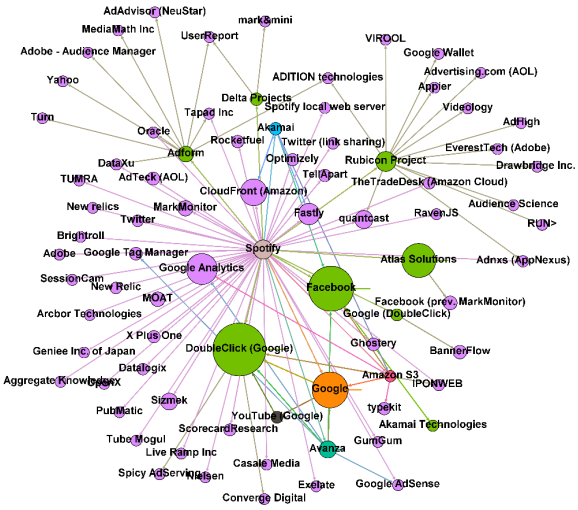

Figure 6. Graph of involved companies during a 60 minute session. Nodes colored by the Gephi Closeness Centrality algorithm, and edges by type of received content. Redirects and content type JavaScript has been removed from this graph.

relate parties and messages to the portions of the user interface being affected by the interaction. The workflow is rather time-consuming, though, and requires both technical knowledge around tools and Web protocols, as well as domain knowledge of the entire ads ecosystem. It is also helpful if one is knowledgeable about the actors involved—all the way from the advertiser to the targets being exposed to the ads. For this purpose, it is possible to create more automated chains

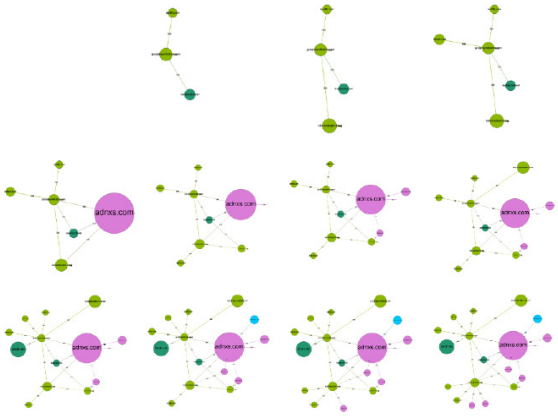

Figure 7. Graphs showing the evolving network of a Spotify Leaderboard as update.The networks are in sequence from left to right and top to bottom, with the final network last.

of tools, especially as there are some public crowd sourced initiatives, an example being the Thalamus database.

Conclusion

This article has provided a brief description of some digital methods used in studying digital advertising technologies and the key stakeholders involved in ad tech infrastructure. We aimed to balance a more systematical media industries approach (Holt and Perren 2009) with the requirements of object-adequate methods for digital data collection. The aim of this research was not so much to generate representative results or models for other usage scenarios. Rather, it aimed to pinpoint how digital tools can be used in critical research without violating user rights or exposing sensitive company secrets. Ethical guidelines issued by the AOIR-Association of Internet Researchers or the Councel for Big Data, Ethics, and Society tend to focus on human subjects in Internet research. Major companies, however, are studied in different ways. Platforms that now act as quasi-monopolies for distributing cultural goods and services need to be open for "audit testing," and such testing must be made possible despite the often restrictive Terms of Services these platforms define (Sandvig 2017).

Patrick Vonderau is Professor at the Department for Media Studies at Stockholm University. His most recent book publication is the co-authored *Spotify Teardown*, forthcoming from MIT Press in 2018. He is a co-editor of *Montage AV*, a co-editor of *Media Industries Journal*, and a co-founder of NECS – European Network for Cinema and Media Studies. E-mail: patrick.vonderau@ims.su.se

Roger Mähler has a degree in computer science and currently holds a position as lead developer at Humlab, Umeå University. Mähler has a long background working as a software architect and systems developer both in academia as well as in the private sector. He participates in several research projects where his main focus is on software development and text analysis. E-mail: roger.mahler@umu.se

Notes

i Digital tools were used during one week in late August 2016. All these tools are publicly available. The data collection has ended and did not involve user data or sensitive company information. With the public and academic interest in mind, we appreciate Spotify's forbearance with any trespassings of Terms of Service that our data collection may have involved.

References

Angwin, Julia et. al. (2016): 'Facebook Doesn't Tell Users Everything It Really Knows About Them,' ProPublica, December 27. Accessed August 19, 2017. www.propublica.org.

Grasseger, Hannes and Mikael Krogerus (2017): 'The Data that Turned the World Upside Down,' Vice Magazine, January 30. Accessed August 19, 2017. www.vice.com.

Greenberg, David M. et al (2016): 'The Song Is You: Attribute Dimensions Reflect Personality,' Social Psychology and Personality Science 7(6): 597–605.

Hölck, Katharina (2016): Beyond the Single Platform: An Assessment of the Functioning and Regulatory Challenges of Multi-Layered Platform Systems in the Media and Communications Sector. Brussels: Vrije Universiteit Brussel.

Holt, Jennifer and Alisa Perren (eds.) (2009): Media Industries: History, Theory, and Method. Chichester: Wiley-Blackwell.

Madrigal, Alexis C. (2017): 'Google and Facebook Have Failed Us,' The Atlantic, October 2. Accessed October 2, 2017. www.theatlantic.com.

Marres, Noortje (2017): Digital Sociology. The Reinvention of Social Research. Cambridge: Polity Press.

Meyer, Robinson (2017): 'Could Facebook Have Caught Its "Jew Hater" Ad Targeting?,' The Atlantic, September 15. Accessed October 2, 2017. www.theatlantic.com.

Nudd, Tim (2016): 'Spotify Crunches User Data in Fun Ways for This New Global Outdoor Ad Campaign,' Adweek, November 29, 2016. Accessed October 24, 2016. www.adweek.com.

Rogers, Richard (2013): Digital Methods. Cambridge: MIT Press.

Sandvig, Christian and Eszter Hargittai (eds.) (2015): Digital Research Confidential: The Secrets of Studying Behavior Online. Cambridge, MA: The MIT Press.

Sandvig, Christian (2017): 'Sandvig v. Sessions: Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online,' www.aclu.org/cases/sandvig-v-sessions-challenge-cfaa-prohibition-uncovering-racial-discrimination-online, September 12. Accessed August 19, 2017.

Skeggs, Beverley and Simon Yuill (2016): 'Capital experimentation with person/a formation: how Facebook's monetization refigures the relationship between property, personhood and protest,' Information, Communication & Society 19 (3), pp. 380-396.

Spotify (2015): The Spotify for Brands Team, 'Hello private marketplaces. Spotify here.' 5 November 2015, brandsnews.spotify.com. Accessed August 19, 2017.

Terdiman, Daniel (2015): 'Spotify Exec: "We Collect An Enormous Amount of Data",' Venturebeat, February 24. Accessed August 19, 2017. www.venturebeat.com.

Vonderau, Patrick (2017): "The Spotify Effect: Digital Distribution and Financial Growth", *Television and New Media* (forthcoming).