Prague, we have a problem – the challenges of building a research infrastructure for the social sciences in the digital era

Bodó, B.; van de Velde, R.N.

Publication date
2018

Prague, we have a problem - the challenges of building a research infrastructure for the social sciences in the digital era

Bodo, B[1], van de Velde, R. N.[2]

Abstract

The University of Amsterdam Research Priority Area on Personalized Communications is based on the cooperation of a legal research institute, the Institute for Information Law (IViR), and a quantitative communication science institute, the Amsterdam School of Communication Research (ASCoR). In the course of 2015-16 we at have outlined the blueprints of a monitoring infrastructure to track and trace how we interact with our algorithmically personalized digital information environment to inform the policy debates, and legal research (Bodo et al. 2017). We argued that we need to be able to collect every data that is exchanged between a user and the internet, for a representative group of respondents, if we want to understand what is taking place in the often algorithmically personalized digital information environments, and what effect digital intermediaries play on the individual and societal level. Our recognition that we need such an infrastructure came from the understanding that it would be impossible to conduct meaningful, evidence based debates about issues, such as filter bubbles, fake news, the effects of political microtargeting, the prevalence of native advertisement, the manipulation of our news diet by hostile state actors, algorithmic biases without actual, first-hand knowledge of what is happening on the screens of our fellow citizens (Bodo, Helberger & de Vreese, 2018). In order to be able to control platforms, algorithms, we need to have a comprehensive, granular, societally representative oversight of the insides of our individual information cocoons.

Traditional methods to create transparency on such a level seem to have failed. The corporations which collect all technically possible data on us do not disclose any of it by their own initiative. Statutory transparency obligations (even versus towards a public authority) are virtually nonexistent in the digital media domain. Classic research methods, such as surveys, focus groups are inadequate, (automated) data collection via scraping and sock puppet audits

[1] Institute for Information Law, UvA, bodo@uva.nl. Corresponding author. The authors are listed in alphabetical order, and contributed equally to the paper.
[2] Information Language Processing Systems group / Amsterdam School for Communication Research, UvA, R.N.vandeVelde@uva.nl

give no insight into the personalized aspects of communication (Kitchin, 2017; Sandvig, 2014), and how humans interact with -or are affected by- such services. Audience measurement companies (Nielsen or ComScore) don't yet measure the use digital services the same way they measure TV, radio, or newspaper consumption. Even if they did, knowing whether someone is on Facebook hardly tells you what they saw there. As a result, individuals, the society, public authorities and scholars alike lost a sense of our digital information environment, and the effects of information intermediaries. The crucial debates on fundamental issues such as manipulation, bias, algorithmic discrimination are mostly based on pure speculation and fear rather than empirical observation.

Goals

The project set out to get a grip on algorithmic influences in the Netherlands. The key goal was to obtain the ingoing and outgoing http(s) requests issued by their browsers for a panel of voluntary respondents during their daily internet routines over a longer period of time. This type of data could show more than just their usual clickstream of what websites they visited, and offer insights into the content they have seen, including advertisements, news items or prices; whether any of these are only shown to specific sub-populations; and the associated user interactions. Looking at the actual exposure for different parts of the population would show for example whether content targeting only generates opinion-congruent information -filter bubbles-, higher income respondents get different offers and face higher pricing -discriminatory pricing- and whether politically naïve respondents get misleading information unseen by others -fake news-. Achieving this goal would entail *intercepting* web traffic generated and consumed by the browsers used by these respondent[3].

Anticipated Challenges

It was clear at the outset that significant challenges would have to be bested. First, the key to understanding the influence of algorithms is to study them not for a niche part of the population (such as heavy internet users), but for all demographics. This entails recruiting and onboarding users with diverse levels of technical knowledge and efficacy, and often requires external partners in the form of panel companies and software developers.

Secondly, to monitor web traffic in a responsible manner, bespoke software was required. This software had to be designed to be able to intercept the communication streams, handle the noisy data observed in this traffic, ensure compliance with data protection and privacy regulation, and make collection transparent to respondents. The sensitive nature of

[3] Commonly referred to as a 'man-in-the-middle attack'.

much of this communication also necessitated the development of complex, resource intensive, and failure prone filtering technologies and access limitations.

Thirdly, privacy concerns not only featured in software design, but were perhaps even more pronounced in the legal and ethical conundrums faced. The capture of potentially highly sensitive personal data, enjoying extra protections under European law forced us to design an implement costly organizational and procedural safeguards above the technical architecture to ensure compliance, and balance the potential harm done to respondents with the data needs of the research.

Unanticipated Challenges

During the implementation of the technical design, and the planned organizational, institutional frameworks we have encountered a number of challenges which we severely underestimated, or completely failed to anticipate. In the following we briefly outline the most important unanticipated challenges.

1. Extremely low signup rates. We hoped to move beyond the limitations often used convenience panels recruited on the internet, by working with a well-established social science research panel in the Netherlands. We hoped that the fact that we recruit from a panel where participants are used to participating in research would not only help us in assessing the representativity of our study, but would result in relatively high signup rates.
In reality recruitment proved to be one of the most important obstacles. We approached in total 6097 potential respondents. Only 40% of them satisfied the technical requirements of being the sole users of an eligible browser. Only a third of the eligible respondents wanted to actually participate in the study, and donate data for us, and many have failed to successfully get on board. This meant we only managed to convert 9.41% of all potential respondents into data generating users. Our active respondents -not surprisingly- tend to be younger, male, highly educated; an interesting, but by all means not representative subset of the Dutch society.

2. Software design tradeoffs The need for bespoke software was anticipated, although the extent of design difficulties were not. Building a reliable infrastructure that includes client-side code (browser plugins), scale-able backends (for data capture and anonymization) with production level uptimes, and storage solutions that allow for analysis is considerably more complex than it seems. External development is hard to obtain with skills that span the requirements for all the moving parts of these infrastructures and hard-choices have to be made by researchers about balancing reliability of data collection with security and investment in testing, as well as monitoring versus timely launching the service.

Rolling out the software to respondents and keeping browser performance tolerable is a second unanticipated challenge. Using the browser as the latch-on point for data collection proved harder because of unforeseen external requirements (such as the installation of proxy

certificates necessary for us to decrypt https traffic). This entailed additional development time to improve the ease-of-installation. Regardless of these efforts, the relative inexperience of many internet users with browser plugins and certificates meant decreases in effective respondent recruitment, and a considerable churn over time (see Fig. 1)

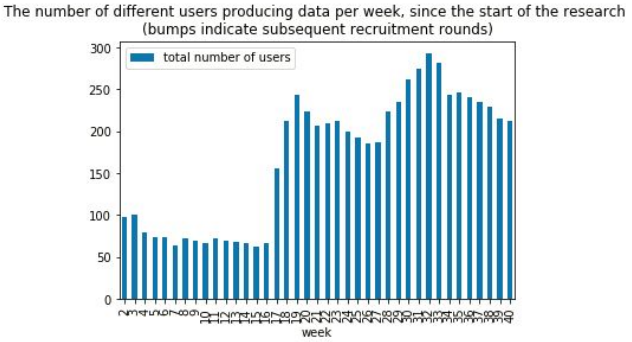

Fig 1. Weekly active users over the course of 40 weeks of data collection. Bumps correspond to recruitment rounds, followed by churns in subsequent weeks.

3. Robustness in the face of messy data Browser are known to be extremely fault-tolerant. The upside is that browsers are able to show content on badly formatted and buggy pages. The downside to any kind of monitoring tool is that content can be widely diverse in terms of structure, quality and formatting. Even explicitly declared content types (such as text, image, script) are often false. The extremely heterogeneous nature of content (un)intentionally obfuscates what websites are showing to respondents.

Not only is content entropy a key challenge to analysis, even the simple act of storage may run into problems. As *a priori* design choices relating to the size and formatting of content may fail when confronted with unforeseen, often undeclared, content types. This can result in data loss in the data-collection pipeline, or worse, break functionality on websites for respondents. As content changes over time due to changes in websites, constant updates to the data-collection pipeline are required to maintain operation.

Preferably, bugs resulting in diminished quality-of-service for respondents or data-loss for researchers are caught in testing, not in the wild. However, the vast size of domains studied and the heterogeneity of respondent and browser behaviour on these domains make test prohibitively expensive. As such, bugs are often encountered post-hoc and the additional costs in terms of development time, data lost and respondent dropouts are inescapable.

4. Too little help from friends The successful implementation of such a large scale project required the coordination of a wide and complex web of partners, including the hardware infrastructure provider, the panel provider, designers, researchers, developers, legal compliance, etc. Such coordination requires resources, skills and expertise, which is not usually available in the domains of legal or communications research. In addition, none of the partners had prior experience with similar approaches, creating friction and extra obstacles on the way. Coaxing partners out of their standard-operating procedures is a costly but essential part in such research. Legal uncertainty, coupled with a general tendency to avoid risk made the implementation especially burdensome.

5. Legal compliance The legal compliance proved to be an especially challenging issue. The EU has a highly sophisticated privacy and data protection framework, which sets strict conditions on both data collection and processing. Navigating this landscape was difficult in itself, balancing the legal obligations with the practical needs of research, the available resources that we could devote to ensuring compliance resulted in seriously limiting our research ambitions to ensure compliance.

And even with such focus on compliance there were unresolvable legal dilemmas, such as collecting data from services which prohibit their users to share their data with third parties. Given that Facebook has such a clause in its Terms of Service, compliance would have prevented us to monitor the most important platform shaping our societies. The decision on which side to err, was the hardest one of all.

6. The implications of data protection Data protection is vital to the privacy of respondents and trust of external partners. Unfortunately, security often comes at the expense of convenience. Making data *safely* accessible to researchers implies setting up an additional infrastructure in which access to raw data which may include highly sensitive personal information can be curtailed and controlled. Researcher access thus requires additional development efforts and must adapt their normally analysis routines to fit the affordances of the data-access platform.

The requirement to anonymize data before and during storage also amounts to significant challenges. The extremely heterogeneous nature of data, and the myriad of opportunities in which respondents can (un)willingly disclose sensitive and/or identifying information to websites, make comprehensive anonymization practically impossible. Even best-effort anonymization requires a tremendous amount of time and personnel. As a result, sharing data with external researchers is heavily limited to curated, abstract representations of the data. Anonymized data as such hardly supports the goals of the project.

Similarly, publishing data together with publications is often impossible in any form relevant to reproduction. The key challenge in operationalizing important social and legal concepts is the processing from the raw data to a numeric representation used in further analysis. It is this step that is key to understanding -and reproducing- the papers. Unfortunately, such data cannot, for privacy reasons, be published with papers.

7. Institutionalizing research Perhaps most importantly from a project management perspective are the challenges faced in engaging with this type of research. Social and legal disciplines are unaccustomed to dealing with software development, maintaining ongoing data-collection infrastructures and analyzing online data. The first resulting challenge is the limited availability of researchers with any substantial coding experience. Even when outsourcing software development and system administration in the best of cases, coding experience is required to understand the trade-offs faced, the technical implications of choices, the required translation of technical possibilities to solutions that support the actual research needs, and the assessment of feasibility and risks.

When it comes to research, the main challenge is the relatively underdeveloped social-empirical methodology to process and analyse these data streams. The volume, velocity and variety of big data [McAfee] sound nice on grant proposals, but lead to great challenges when it comes to: 1) scaleable, time-sensitive and robust operationalizations of concepts such as 'personalization', 2) *meaningful* differences in content and 3) even simply best practices in exploring the data. Part of the problem here is that it requires coding expertise to be able to handle the data in light of the constant need to translate between technical, legal and social perspectives on the data. An example can be found in the need to operationalize personalization in news content. Classic content-analysis approaches would count the salience of a topic in newspapers and correlate this with readership to gauge exposure. But in web data, differences in exposure may happen for a myriad of mostly random reasons (different time logging in, different browser, different search terms) and topics cannot be defined across all variations of shorter and longer texts shared on social media platforms. Without even limited technical knowledge, researchers cannot meaningfully engage in multi-disciplinary research because they lack any feeling for what is trivial or impossible in operationalizing and thus building or testing theories.

The role of universities as institutions that foster research provide another double-edged sword with regard to such a project. The curiosity enshrined in the university as an institute about the empirical facts underlying theories of online personalization, fragmentation and misinformation drives the development of this data-collection infrastructure. Other institutional properties are less forgiving. Universities hire personnel for the long term, however if future research skills are incorrectly anticipated, inadequate or missing skills will hinder the development and application of adequate, future proof research methods. The need to invest resources in activities normally foreign to universities, makes budgeting software development and system administration for data collection hard to do. The need to step -at least partially- outside normal research traditions also raises the stakes for young researchers aiming to make a career for themselves.

What have we learned?

It is clear that we need long-term, solid commitment to an infrastructure which provides the necessary insight into our digital information environment. Having access to such information is a must, not just for academia, but for every member of society, for the policy discourse, for the public in general.

Our journey proved that the challenges associated with building such an infrastructure are commensurate with the size of the societal problem we face. While there are a number of isolated, small scale efforts to shed light into our digital communications, it is obvious, that without focused, wide scale efforts, even limited transparency is hard to achieve. We need to build consortia that bridge disciplinary as well as national boundaries, to create the scope and depth of expertise required to overcome the aforementioned technical, organizational, institutional, and paradigmatic challenges. It seems that academia as of yet, is not the best

prepared to solve the these challenges, but it is of utmost importance that any such infrastructure is first and foremost public. We cannot, but we also should not wait for the market to build such an infrastructure. The incentive structures in the market push information capture to be fragmented, of limited availability for inspection and biased to protect corporate interests. Such insights are and will remain far too valuable to leave its provision exclusively to the market or to the observed online services themselves. Sadly, as we described, such an effort needs to involve taking risks, and pushing the the existing boundaries of scholarly research. Luckily this is exactly what science ought to do.

References

Bodó, Balázs, Natali Helberger, Kristina Irion, Frederik J. Borgesius Zuiderveen, Judith Moller, Bob van de Velde, Nadine Bol, Bram van Es, and Claes H. de Vreese. 2017. “Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents.” Yale Journal of Law & Technology 19: 133.

Bodó, B. & Helberger, N. & de Vreese, C. (2017). Political micro-targeting: a Manchurian candidate or just a dark horse?. Internet Policy Review, 6(4). DOI: 10.14763/2017.4.776

Kitchin, Rob. 2017. “Thinking Critically about and Researching Algorithms.” Information Communication and Society 20 (1): 14–29. doi:10.1080/1369118X.2016.1154087.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms.” Data and Discrimination: Converting Critical Concerns into Productive Inquiry.