Introduction to National Internet Observatory

Álvaro Feal, Jeffrey Gleason, Pranav Goel, Jason Radford, Kai-Cheng Yang,
John Basl, Michelle Meyer, David Choffnes, Christo Wilson, and David Lazer

National Internet Observatory, Northeastern University, USA

Abstract

The National Internet Observatory (NIO) aims to help researchers study online behavior. Participants install a browser extension and/or mobile apps to donate their online activity data along with comprehensive survey responses. The infrastructure will offer approved researchers access to a suite of structured, parsed content data for selected domains to enable analyses and understanding of Internet use in the US. This is all conducted within a robust research ethics framework, emphasizing ongoing informed consent and multiple layers, technical and legal, of interventions to protect the values at stake in data collection, data access, and research. This paper provides a brief overview of the NIO infrastructure, the data collected, the participants, and the researcher intake process.

Introduction

Individuals are increasingly spending a significant portion of their lives online. Currently, more than half of the world's population has access to the Internet,[1] with notably higher percentages in developed countries (e.g., 89% of US adults; Pew Research Center (2024)). Dominant platforms like Google, Facebook, and YouTube are used by the majority of US adults (Smith and Anderson 2018),[2] while several other platforms, such as Twitter and Instagram, serve a smaller yet sizable portion of the population (Smith and Anderson 2018).[3] The Internet plays a pivotal role in connecting people and serves as a primary medium for obtaining and disseminating information. The shift towards more extensive online engagement worldwide presents unique opportunities for observing human behavior on an unprecedented scale. Moreover, the Internet facilitates the seamless monitoring of human activities that platforms leverage to infer user characteristics based on behavioral data stored in extensive databases.

However, virtually none of this data is available for academic research. The vast amount of data created about online human behavior is siloed, proprietary, with problematic scientific foundations, and often collected in ways that are on ethically shaky grounds. While useful research has been conducted on the Internet and society in social sciences, digital humanities, and computer science, the current major instruments available for open science face severe limitations:

- **Bespoke data collection:** Much of the research on digital traces is akin to the search of a drunk in the parking lot for their keys, under the lamp post; rather than in the dark where they were dropped. The lamp post here includes research on Twitter (Grinberg et al. 2019; Vosoughi, Roy, and Aral 2018), Reddit (Haralabopoulos, Anagnostopoulos, and Zeadally 2015), and Wikipedia (Kumar, West, and Leskovec 2016) data. The resulting data sets, however, are often not publicly available (e.g., Twitter's terms of service forbid the sharing of tweets); are siloed; are rarely well documented and thus not replicable (Gaffney and Matias 2018); and are often decontextualized from other information about individuals.

  Moreover, the collection of such data has become increasingly challenging and restricted. This data collection often relied on official Application Programming Interfaces (APIs) provided by major platforms, including social media. However, these APIs have been progressively restricted over recent years, marking the transition into a "Post-API" era, with Facebook shutting down key APIs in 2018 after the Cambridge Analytica scandal (Freelon 2018; Bruns 2021), and another effective closure of the Twitter/X API in the spring of 2023—a data collection tool that has been at the heart of vast amounts of social media research (Murtfeldt et al. 2024).

- **Data from vendors:** There are a number of options from vendors, including Nielsen, Comscore, and YouGov, all of which maintain panels of subjects and monitor their online activities. While all of these have, on occasion, been used by academics, there are a number of serious problems. Comscore and Nielsen are optimized for commercial clients, with pricing structures utterly incompatible with academic budgets, although some ad hoc discounts have been provided for specific projects. They have several features that make them problematic for scientific use: all of them are closed-source, often without longitudinal access to a large sample. Because these systems are entirely proprietary, it is impossible to assess

[1] internetlivestats.com/internet-users

[2] statista.com/topics/1001/google; pewinternet.org/2018/03/01/social-media-use-in-2018; youtube.com/intl/en-GB/yt/about/press

[3] statista.com/topics/1882/instagram

many dimensions of scientific validity, from sample to instrumentation quality.

Additionally, the ethical practices of these vendors are either non-transparent or problematic. Comscore, in particular, has faced criticism for questionable ethical practices, including allegations of installing monitoring software on computers without user consent (Lerer 2006). The security practices of these vendors are often unclear, and some available details suggest potential security risks for panelists. YouGov, for instance, maintains its Pulse panel, which tracks online behavior, using proprietary technology from Wakoopa. Its security procedures are completely opaque, and its website describes the technology with "Brands who want to get up-close and personal with their target audience deserve the best seat in the house... Enjoy the view."[4]

- **Partnerships with industry:** Industry has access to a myriad of data about human behavior and has occasionally partnered with academics (Kramer, Guillory, and Hancock 2014; Muchnik, Aral, and Taylor 2013). However, these collaborations face challenges (Jasny et al. 2017). Firstly, companies are commercial entities with no interest in releasing information that may hurt their public image. This was evident when Facebook restricted external collaborations following the backlash from the publication of the Kramer, Guillory, and Hancock (2014) study on emotional contagion. Secondly, while industry data sets are extensive, they are typically siloed, providing in-depth behavior observations within specific platforms but lacking a broader cross-platform perspective. Social Science One, an emergent effort to facilitate access to Facebook data (King and Persily 2020), potentially offers a more robust model of academic-industry collaboration but does not resolve the need for a research-oriented platform for cross-platform data collection.

Other common limitations of the current instruments include *i)* the focus on production behavior (content created and shared by users), as opposed to consumption behavior or exposure to content created by others; and *ii)* an emphasis on a limited set of platforms, such as Twitter, which is often chosen due to the relative ease of data availability.

In short, there is a pressing need for a scientifically rigorous infrastructure for academics to study behavior on the Internet comprehensively. The lack of such an infrastructure not only hampers scientific advancement but also has wider societal implications, given the Internet's integral role in contemporary life. Indeed, the demand for such infrastructures has grown in recent years, underscored by numerous calls for the creation of a "data commons" to facilitate research into Internet usage and other areas of human behavior (See 2018 Social Science Research Council Report To Secure Knowledge: Social Science Partnerships for the Common Good;[5] also, Lazer et al. (2009); Resnick, Adar, and Lampe (2015); Watts (2017)). These calls have appeared in both academic and government reports, but few efforts have been made to realize these infrastructures to date. Among the efforts to build them is the University of Zurich's Digital Democracy Lab, a panel-based infrastructure limited to collecting social media data. As discussed above, the Comscore, Nielsen, and YouGov Pulse online monitoring panels have also been used; however, the applications have been limited due to their prohibitive costs.

The National Internet Observatory (NIO) aims to help answer these calls by serving as an open, large-scale, secure, and privacy-preserving observatory of online behavior to enable academic research without relying on bespoke data collection, proprietary sources, or partnerships with industry.[6]

Infrastructure & Data Collection

A visual summary of the main part of NIO's infrastructure is shown in Fig. 1. There are two mechanisms for data collection: browser extension-based and mobile-based systems. Below, we introduce both and then provide an overview of our infrastructure's privacy and ethics mechanisms.

Desktop/laptop client

NIO leverages a browser extension to collect data about participants' online behavior. The extension is compatible with Chromium-based browsers (i.e., Google Chrome, Microsoft Edge, Brave) and collects the following data types:
Browsing activity. This includes the sequence of visited URLs, duration and focus of each visit, transitions between tabs and windows, and page navigation, allowing the full reconstruction of users' web browsing trajectories.
HTML snapshots. For a subset of important websites, we collect a full snapshot of the non-private content shown to users. We then use open-source and custom-built HTML parsers to extract structured data from these snapshots (e.g. Google Search results (Robertson and Wilson 2020)) to facilitate downstream analyses.
Browser state. We collect data about browser cookies and Chrome's new privacy initiatives, such as the Topics API.[7]

Mobile clients

NIO has mobile clients for Android and iOS. Some of the collected data types are common to both systems, while others are unique to Android because it provides access to APIs unavailable on iOS. Specifically, the clients collect the following data:
Network communications. Both of our clients rely on an on-device VPN that intercepts traffic from all apps (Le et al. 2015), allowing us to collect header information (TCP/UDP and IP) before forwarding the traffic to its intended destination. These headers allow us to gather information such as which IPs (and domains) are being contacted by individual apps or the amount of data sent. We could also create fingerprints of certain network behaviors (i.e., loading a specific news article) to detect user behaviors in our historical data. Note that we do not decrypt any of the traffic nor do we record packet payloads (i.e., content).

[4] wakoopa.com

[5] ssrc.org/wp-content/uploads/2021/08/SSRCToSecureKnowledge.pdf

[6] nationalinternetobservatory.org

[7] privacysandbox.com/intl/en_us/open-web

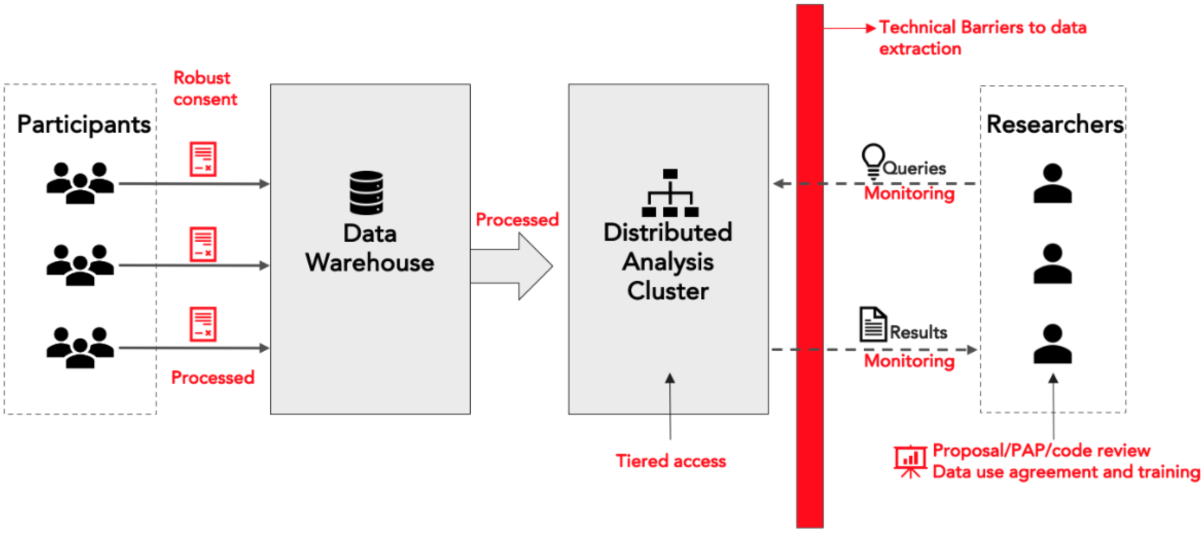

Figure 1: A visual summary of NIO's infrastructure.

App usage. Both Android and iOS provide APIs to access app usage statistics. We query this data every 15 minutes and calculate the number of times that each individual app is opened and how long the app stays in the foreground (i.e., actively visible on the phone). For iOS, these statistics are aggregated per app type (e.g., Utility and Games).

App content (Android only). Contrary to the web, where it is possible to obtain a snapshot of the content being displayed to the user, mobile operating systems do not provide such an API. However, Android has an Accessibility API designed to support functionalities such as screen readers, which allow people with reading impediments to access the text visible on the screen. The NIO Android app leverages this API to build custom-made parsers to obtain certain parts of the content displayed to users through apps. Although not all apps make this data available through the Accessibility API, for certain apps, we can obtain specific parts of the text shown to users that are not considered private (e.g., posts shown on the user's Twitter/X timeline).

Privacy and ethics

NIO's commitment to ethical data collection is oriented around a set of core ends, which include respecting participant autonomy, minimizing privacy risks to participants and bystanders, and promoting transparency (Meyer et al. 2023). NIO's ethics team includes two ethicists with specializations in bioethics and AI and data ethics, alongside team members with expertise in cybersecurity and privacy.

Our informed consent procedure advances the state of the art for consent and promotes participant autonomy with an e-consent process that requires participants to take a comprehension quiz before they are allowed to join NIO. This quiz helps ensure that our research participants genuinely understand the ways in which their data might be used and the risks to them for their participation in NIO. Access to NIO data is limited, and researchers will undergo training designed to sensitize them to the nuances of ethical issues associated with research enabled by NIO. Research projects will undergo ethics review to help researchers mitigate ethical risks prior to conducting research. Researchers and their institutions are bound by legal agreements and a code of conduct that aligns them with NIO's core values. This code includes, for example, the requirement to alert NIO whenever a research participant is identifiable from the data a researcher has access to.

NIO also implements various technical tools to ensure privacy and data security. We use state-of-the-art security procedures from data collection to storage and analysis, including TLS encryption between participants and our servers and file encryption for at-rest data. Administrators and researchers utilizing the platform are authenticated using unforgeable hardware two-factor authentication (YubiKeys). Web servers, databases, and long-term file storage are held on physically distinct servers, which are installed in a secure data center with keycard access controls. Standard tools like network and application firewalls, logging, and intrusion detection have been deployed to mitigate and detect security issues. No data is and will be stored in public clouds.

We are deploying a defense-in-depth strategy against "insider" attacks on sensitive data to guard the security of the data and privacy of participants. Standard access controls are used to limit access to sensitive datasets, and all file accesses and database queries are tracked to identify anomalies. Strict bandwidth limits are implemented to prevent data leaks. We are also developing a set of differential privacy methods to prevent the re-identification of participants.

Participants

We are building a panel of thousands of participants who regularly contribute data from their mobile device(s) and/or web browsers. All participants must be older than 18, reside in the US, and have at least one eligible device (we do not offer devices to participants). We support English and Spanish-speaking participants. Participants are offered regular surveys to remain engaged and provide ongoing data about their behaviors, opinions, and beliefs to be used for research.

We adopt a two-pronged approach to participant recruitment to address different research needs and counteract inherent sampling biases. Most participants will be recruited through non-probability recruitment channels, specifically existing participant pools such as Forthright Panels, Pure Spectrum, and Verasight, and paid online advertising through Meta, Google, and Reddit. This strategy aims to amass a large participant pool, anticipated to reach around 5,000 by the fall of 2024. The rationale behind this approach is to gather a sufficiently large sample that can capture rare, long-tailed phenomena that occur only in small percentages within the population. While cost-effective and capable of quickly scaling the participant base, non-probability samples can introduce biases that may skew the results and affect their generalizability.

To mitigate the limitations of non-probability samples, we also plan to establish a smaller, probability-based panel consisting of about 1,000 to 2,000 participants. This panel aims to collect the same types of browser and mobile phone data as the non-probability panel but from a statistically representative sample of the population. This method allows researchers to model and understand the biases present in the non-probability data. For instance, if a discrepancy in AI-generated results in Google searches is observed, researchers can use the probability-based panel to determine whether such patterns are genuinely prevalent across the broader population or are specific to the initial non-probability sample.

Researcher intake

We will provide researchers with access to various levels of detailed data. Collected data will be processed by NIO scripts, producing aggregated and pseudonymized datasets. We will publicly release some of these anonymized and aggregated datasets, showing top search terms and websites.

Researchers requiring more detailed data will need to sign data use agreements, which will contain a detailed overview of their project, the people requiring access, the duration of access, and the data sources and kinds of data needed for their research. If approved, the researchers will be onboarded and obtain credentials for remote access to a secure Spark computing cluster. In cases where research hinges on raw data access, authorized researchers will be given read-only access via the Spark cluster to only the corpus sections critical to their specific research project.

References

Bruns, A. 2021. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Disinformation and Data Lockdown on Social Platforms*, 14–36.

Freelon, D. 2018. Computational research in the post-API age. *Political Communication*, 35(4): 665–668.

Gaffney, D.; and Matias, J. N. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one*, 13(7): e0200162.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425): 374–378.

Haralabopoulos, G.; Anagnostopoulos, I.; and Zeadally, S. 2015. Lifespan and propagation of information in On-line Social Networks: A case study based on Reddit. *Journal of network and computer applications*, 56: 88–100.

Jasny, B.; Wigginton, N.; McNutt, M.; Bubela, T.; Buck, S.; Cook-Deegan, R.; Gardner, T.; Hanson, B.; Hustad, C.; Kiermer, V.; et al. 2017. Fostering reproducibility in industry-academia research. *Science*, 357(6353): 759–761.

King, G.; and Persily, N. 2020. A new model for industry–academic partnerships. *PS: Political Science & Politics*, 53(4): 703–709.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America*, 111(24): 8788.

Kumar, S.; West, R.; and Leskovec, J. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, 591–602.

Lazer, D.; Brewer, D.; Christakis, N.; Fowler, J.; and King, G. 2009. Life in the network: the coming age of computational social. *Science*, 323(5915): 721–723.

Le, A.; Varmarken, J.; Langhoff, S.; Shuba, A.; Gjoka, M.; and Markopoulou, A. 2015. AntMonitor: A system for monitoring from mobile devices. In *Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data*, 15–20.

Lerer, L. 2006. How Much Privacy? Retrieved February 18, 2019, from https://www.forbes.com/2006/12/07/internet-security-research-tech_cx_ll_1208comscore.

Meyer, M. N.; Basl, J.; Choffnes, D.; Wilson, C.; and Lazer, D. M. J. 2023. Enhancing the Ethics of User-Sourced Online Data Collection and Sharing. *Nature Computational Science*, 1–5. Publisher: Nature Publishing Group.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science*, 341(6146): 647–651.

Murtfeldt, R.; Alterman, N.; Kahveci, I.; and West, J. D. 2024. RIP Twitter API: A eulogy to its vast research contributions. *preprint arXiv:2404.07340*.

Pew Research Center. 2024. Internet, Broadband Fact Sheet.

Resnick, P.; Adar, E.; and Lampe, C. 2015. What social media data we are missing and how to get it. *The ANNALS of the American Academy of Political and Social Science*, 659(1): 192–206.

Robertson, R. E.; and Wilson, C. 2020. WebSearcher: Tools for Auditing Web Search. In *Proceedings of the 2020 Computation+ Journalism Symposium (Boston, MA, USA)(C+ J 2020)*.

Smith, A.; and Anderson, M. 2018. Social Media Use in 2018.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Watts, D. J. 2017. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1): 0015.