Information Research - Vol. 30 No. iConf (2025)

Investigating privacy risks in open government data: an exploratory case study

Daniel Carter and Caroline Stratton

DOI: https://doi.org/10.47989/ir30iConf47287

Abstract

Introduction. As governments increasingly release large datasets to the public, concerns have arisen regarding the potential risks associated with the inadvertent disclosure of personal information. However, there are few empirical case studies of this phenomena.

Method. Taking a red teaming approach, we collected public, nonprofit reporting documents from the IRS, as well as a list of professional athletes. Using standard data processing techniques, we attempt to identify and evaluate the personal information that can be attached to these notable individuals in order to propose a framework describing how features of information, people and processes influence the potential release of personal information.

Findings. We find that name commonness plays a large role in determining the feasibility of matching records. We also find that personal information, including addresses and business relationships, are accessible in the dataset.

Conclusion. Our exploratory case study suggests several features of information, people and processes that influence the release of private information held in open government data. While limited, it suggests the value of further work in the area.

Introduction

In recent years, the open government data (OGD) movement has gained significant traction, promoting transparency, accountability, participation and trust (Janssen et al., 2012). However, as governments increasingly release large datasets to the public, concerns have arisen regarding the potential risks associated with the inadvertent disclosure of personal information(Graham et al., 2016). This paper extends the literature on open government data by presenting an exploratory study that evaluates the risk related to a specific case of government data and proposes an initial framework that can guide future work.

Specifically, we focus here on the risk that personal information might be inadvertently released in nonprofit reporting documents such as the IRS form 990. The risk scenario that we attempt to better understand is one in which a dataset collected from such a source might be used to identify personal information such as addresses and contact information for notable individuals such as celebrities (in this case, NBA players) or a list of less notable people that might be relevant in a specific context (e.g., journalists, activists or teachers). While such contact information is available through other public sources, such as property records, we choose to work with nonprofit reporting documents because they have a national scope and, for risk scenarios that rely on targeting individuals with wealth, likely represent a higher density of targets.

The generalized information context that we attempt to understand, then, is a large dataset that is not valuable as a whole but is rather assumed to contain some valuable information that must be located and extracted. In conducting a case study to better understand this phenomenon, we assume that an adversary would approach this as an information retrieval task and model our work accordingly.

Based on our findings, we propose an initial framework that identifies key attributes of information, people, and processes that may impact the likelihood of harmful data release as a result of OGD initiatives. By examining these factors, we aim to contribute to the ongoing dialogue about balancing the benefits of open government data with the need to protect individual privacy and prevent potential misuse of personal information.

Literature review

The literature on OGD has raised the possibility that privacy is a salient concern for governments and may be in tension with the aim of transparency (Graham et al., 2016; Janssen & van den Hoven, 2015). In general, while OGD conforms to the conditions that Janssen et al. (2012, p. 258) stipulate as '*non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distribution*,' the removal of all personally identifiable information (PII) from OGD, both that of residents who interact with governments and that of government workers, is far from uniform. In some cases, the value of an open dataset lies in its identification of people and their associated characteristics. The three most popular datasets on New York City's open data portal as of September 2024, for example, list individuals' names and their vehicle identification numbers (For Hire Vehicles dataset), individuals' names and scores for a civil service exam (Civil Service List dataset), and individuals' names and addresses associated with building permits (DOB Job Application Filings dataset) (City of New York, n.d.).

While scholars have speculated about the relationship between privacy and OGD, we find little identification or exploration of specific cases in the literature. For most OGD research, the concept of individual privacy is left at the hypothetical level.

In the years since OGD has grown in popularity, other phenomena that jeopardize individual privacy, such as the growth of data broker companies (Crain, 2018) and the prevalence of large-scale corporate data breaches (e.g., Newman, 2024; Srinivasan, 2017), have perhaps overshadowed interest in risk and OGD. While the impacts of selling and stealing data are certainly grave, the risk

posed by PII in OGD is one of hiding in plain sight, with potentially valuable personal information available to anyone capable of finding it. An older example illustrating potential harms of PII in OGD comes from Eightmaps, a visualization that showed donors to a 2008 California ballot measure to stop single-sex couples from marrying (Stone, 2009). De-identifying individual data is among the strategies that governments might take to reduce individual risk to privacy. Techniques for re-identification, though, may be powerful enough to nullify such efforts. Liu et al. (2021), for example, raised the possibility of re-identification for a dataset of car trips from highway toll-collecting, with license plate numbers removed. Because the dataset also contained information about location of departure, destination, and travel times, the authors argued that re-identification might be possible for some entries.

The potential for exploitation of PII found in OGD would seem to be against the ethos of the OGD movement, which has emphasized the transparency benefits and potential gains in efficiency for government and other stakeholders; however, one of the most important proclamations regarding OGD, Barack Obama's 2013 Executive Order (2013), declares that these data can and should lead to economic benefit, because enterprising people can '*develop a vast range of useful new products and businesses using these public information resources*.' A tension emerges here, then, between the commercial value of OGD and the potential for that value to be derived through the exploitation of PII found in government information. In this short paper, we explore risk around PII in OGD through a specific information retrieval task, building toward future work by asking:

What are the salient features of public information, people and processes that put personal information at risk?

Methods

To explore our research question, we adopted a red teaming approach (Mansfield-Devine, 2018) designed to understand how the release of personal information might be perpetrated by someone with a financial or personal incentive. We began with a dataset composed of Form 990-N submissions made publicly available from the IRS. Form 990-N is an abbreviated version of the standard form 990 which is used by nonprofit organizations with gross receipts of less than $50,000 to make annual reports. The abbreviated form 990-N data is made available in a single, delimited text file, as opposed to the full form 990 data, which would require more sophisticated search and join procedures. While we assume that there are differences in these datasets related to the information made available as well as other features such as the prevalence of professional document preparers, we chose the abbreviated set in order to facilitate an initial, exploratory study and because the ease of acquiring and working with the data might more closely align with the workflows of adversaries.

Our *form 990-N dataset* consists of 1.39 million records, representing annual reports from 2007 through 2023. The relevant information included in the data includes the principal officer's name and address as well as the address listed for the organization. We note that the full form 990 includes additional information that might be relevant to the topic, such as information related to revenue and board member compensation.

In order to explore the presence of information related to notable individuals in the dataset, we use a list of NBA player names obtained from basketball-reference.com, which includes all players who have ever played in the league. The dataset, which we refer to here as the *target dataset*, includes 5,738 players. Because the name of the nonprofit's principal officer is the only feature that can be matched, we do not include other features, such as date of birth, in our *target dataset*.

In order to explore the extent to which notable individuals can be identified in our *form 990-N dataset*, we first reduce the dataset to include only records that match a name in the *target dataset*. The resulting *matched dataset* includes 5,769 form 990-N records that might correspond to a

notable individual. However, some of these records will obviously represent false positives, or organizations with a principal officer who shares a name with a notable individual.

Because we are interested in features of the information as well as of the processes used to identify relevant personal information within it, we employ an evaluation process that balances thoroughness with feasibility, as follows:

1. **Ranking by Commonness:** We first ranked all the names in our dataset based on their commonness, from least common to most common using Hadley Wickham's babynames dataset (2022), which includes all first names used at least five times in the US, as provided by the Social Security Administration. This ranking allowed us to prioritize the evaluation of names that were more likely to correspond to notable individuals.

2. **Cutoff Determination:** The information contained in the public forms collected is insufficient to automatically verify if information is relevant (i.e., it corresponds to a notable individual); therefore, a more time-consuming, manual process must be employed. Because we assume that the value of information about players is randomly distributed (i.e., there is as much likelihood of value in the information regarding a notable individual that is easy to identify as one that is difficult to identify), we also assume that an efficient search strategy, such as would be adopted by an adversary, would focus efforts on data that is easier to verify and has a greater chance of corresponding to a notable individual. We discuss findings related to determining such a search cutoff below.

3. **Analysis:** For each record we evaluated, we determined the following:

    1. Relevance: Using information available online, are we able to determine if the named principal officer is a notable individual from our *target dataset*? We use a conservative approach and only record a match if we are able to locate explicit confirming information.

    2. Personal information: We evaluate if the information from the form 990-N record contains personal details such as a non-commercial address.

    3. Evidence of professional preparation: We evaluate if there is evidence that a professional entity such as a law firm or business manager prepared the form.

    4. Additional notes: We make note of any aspects of the form information that does not correspond to our analysis plan or warrants further attention.

Findings

We observe that both the *target dataset* and the *form 990-N dataset* exhibit an exponential distribution of name commonness (figures 1 and 2). We also observe that the *merged dataset* exhibits a markedly different distribution, with a greater proportion of results found to have higher name commonness (figure 3). Calculating the ratio of the proportion of values from the *merged dataset* to the *form 990-N dataset*, we find a roughly linear increasing relationship (figure 4), suggesting that the number of results that do not correspond to an individual in the *target dataset* will increase with name commonness. While this is not a surprising result, it has notable consequences for the general search problem described, as it suggests that an adversary would likely focus efforts on less common names.

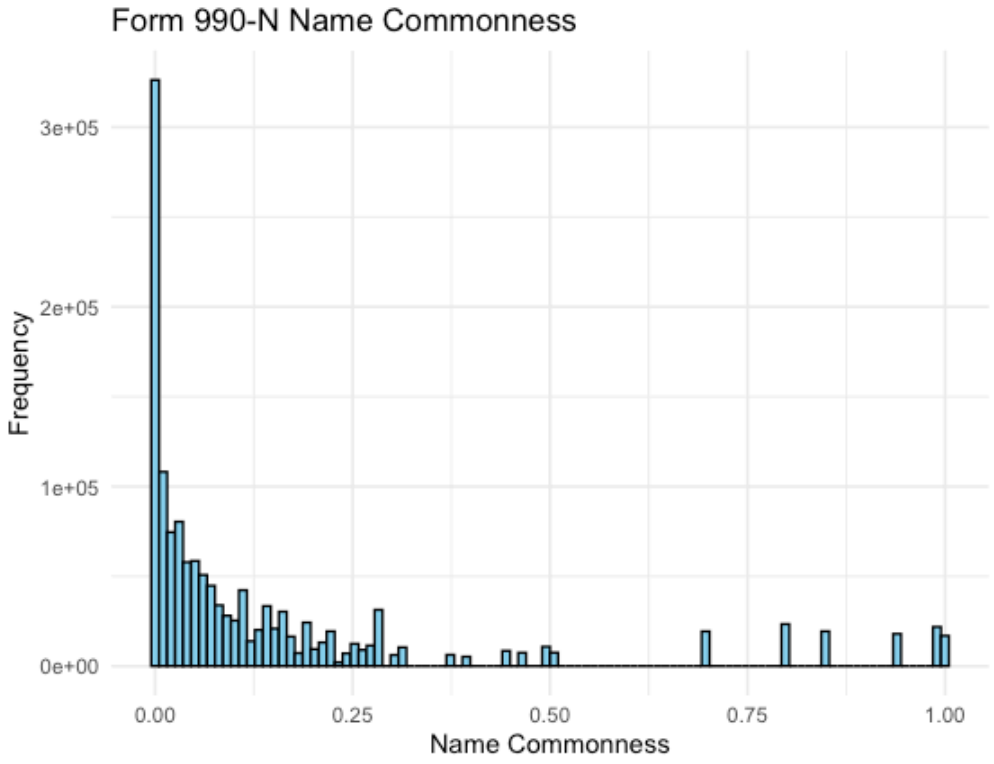

Figure 1. Distribution of name commonness in the Form 990-N dataset.

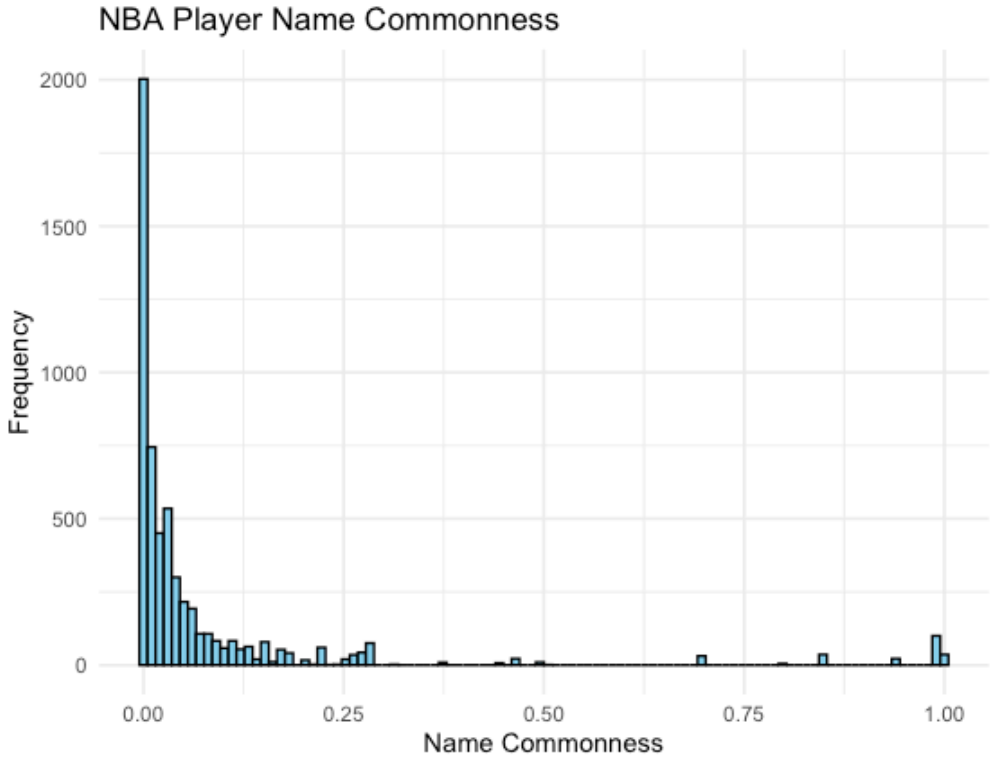

Figure 2. Distribution of name commonness in the target dataset.

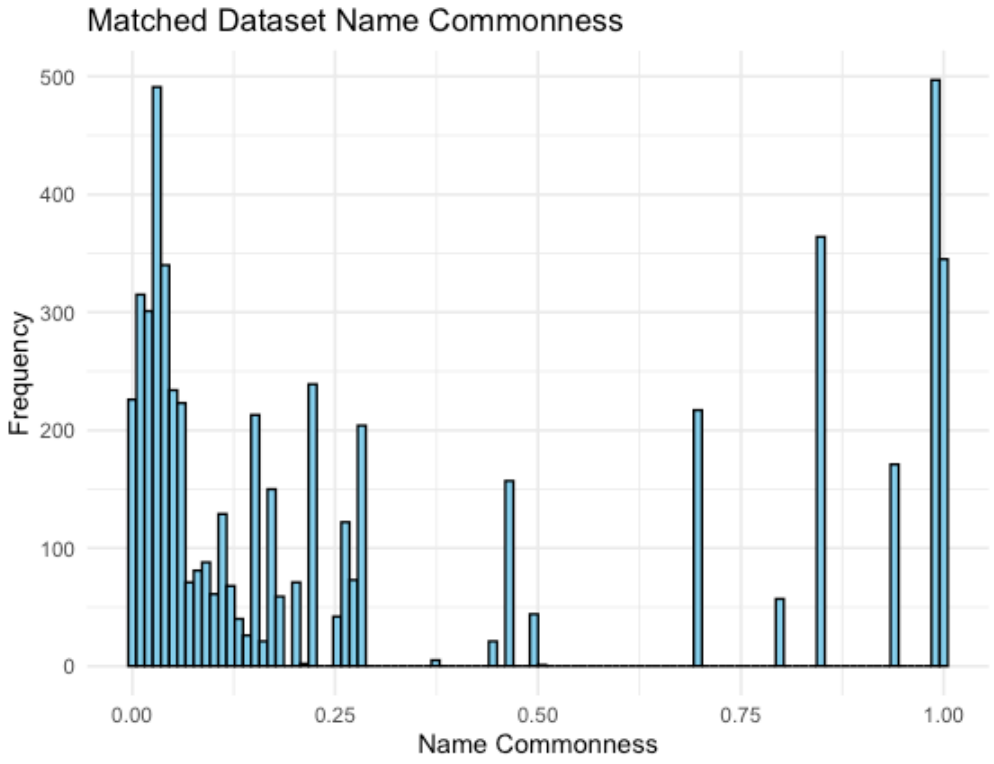

Figure 3. Distribution of name commonness in the matched dataset.

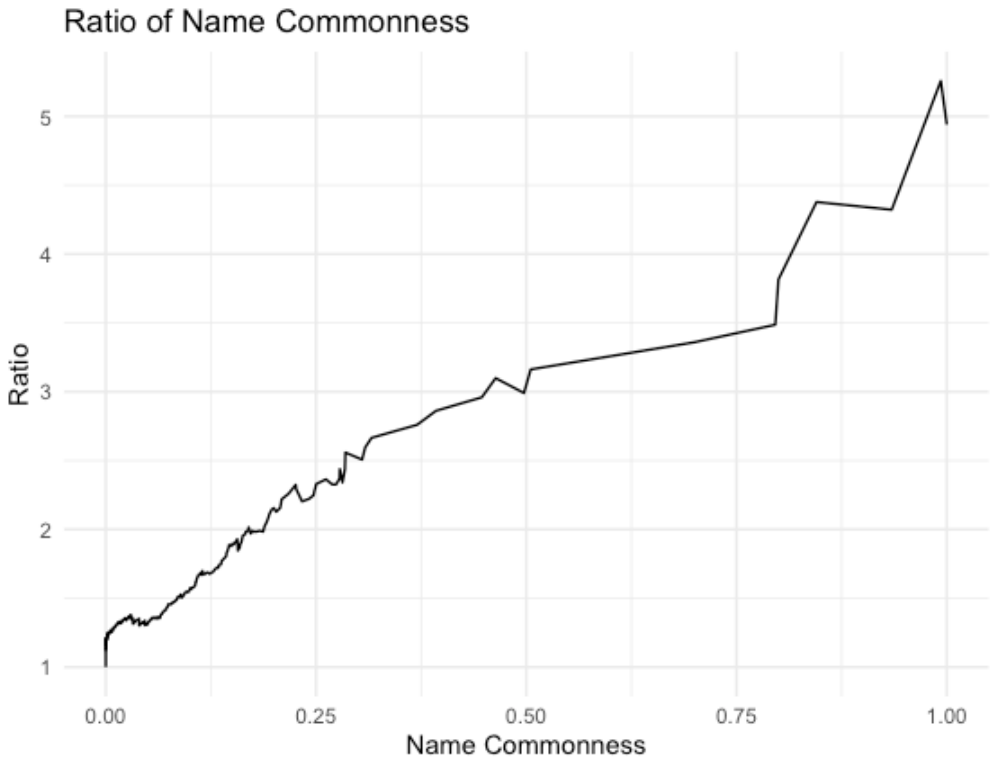

Figure 4. Ratio of the proportion of records, by name commonness, in the merged dataset and the Form 990-N dataset.

Because the purpose of this study is not to locate as much personal information as possible but to identify features of the information, people and processes involved that influence the risk of personal information release, we do not attempt to determine a precise cutoff (based, for example, in the ratio presented in figure 4) — indeed, given the highly contextual nature of the search task, determining such a limit is not realistic in this context. Instead, we evaluated records until we reached a point at which more than 100 consecutive records could not be matched to an individual from the *target dataset*, resulting in the inspection of 384 records. Figure 5 depicts the observed decrease in the rate of matches.

Of the evaluated records, we found that 13% matched an individual in the *target dataset*, with the rate of matches decreasing, as expected, as names became more common. Notably, we find that this rate decreases sharply, with just over 5% of records evaluated before matches became quite rare.

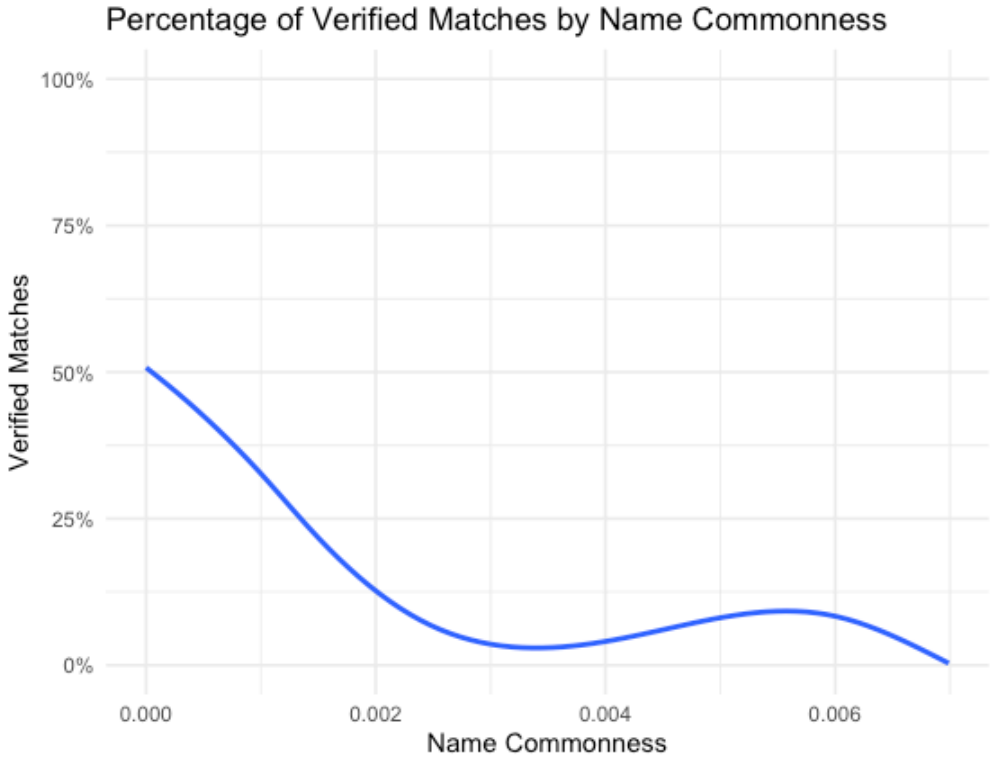

Figure 5. Percent of verified matched by name commonness

Of the verified matches, 52% included a non-commercial address, which we assume is likely the individual's residence. Notably, this information can be linked to other personal information, such as the value of an individual's home, which, while public, can be difficult to locate because property records are held at the county level.

36% of the matched records show evidence of the use of a professional service to submit the information. Observed professional services included both law firms and agencies that represent athletes. In all such cases, a commercial address was present instead of a residential address.

Discussion

Based on the results of our exploratory case study, we propose the following framework, which highlights features of the documents, people and processes involved in open government data.

Information characteristics

At a basic level, we observe that the fields included in the released public data play a role in determining the risk of personal information disclosure. While it's obvious that the inclusion of details such as social security numbers or health information would increase the risk related to disclosure, we also note that the absence of information that is commonly used to link records (e.g., date of birth) plays a marked role in reducing risk and makes the verification process more laborious.

Individual characteristics

We observe two features of individuals that likely impact the release of personal information through OGD processes: information literacy and name commonness.

Regarding information literacy, we observed several records in which fields were filled with information other than what was requested. This was most frequent with the form's Website field, which was sometimes used by individuals to enter an email address. In these cases, a lack of information literacy results in the release of personal information and provides another opportunity to link the record to additional information. The inclusion of personal addresses might be considered another example of a lack of information literacy. As we noted, professional preparation services, such as law firms, always use a commercial address, likely signaling that these entities believe there is an advantage to obscuring personal information. The inclusion of such information, then, signals an individual's lack of information literacy and understanding of how submitted information is, or is not, made public.

We also observe that less common names have a notable effect on the difficulty of verifying information in the specific search context described. This is especially obvious in our *target dataset* of NBA players, as a substantial number of foreign players come to the United States for this reason. As we discuss above, higher name commonality quickly makes verifying information impractical without automatic processing, making individuals with less common or foreign names much more vulnerable.

Process characteristics

Related to the process through which information is submitted, we observe that individuals who use a business intermediary such as an agent or lawyer appear much less likely to include personal information such as a home address. Instead, the address of the business appears in the public information; while this masks some of the individual's information, it can reveal other information such as business relationships that might be valuable in themselves.

Conclusion

This paper presents a needed empirical study of the risks associated with OGD. Taking a red teaming approach, we sought to better understand the features of information, people and processes that influence the risk of personal information disclosure. While we note challenges to the information retrieval task described above, notably in relation to name commonness and linking data, we also note that, with fairly little effort and using only public data, we were able to obtain the residential addresses of 50 NBA players, as well as other, non-public information such as the agency they contract with. The value of this information is highly context-dependent, but our study does suggest that a process similar to ours could be profitable to an adversary in some contexts.

Because this is an exploratory study, it has considerable limitations.

At a very broad level, we are constrained by the general difficulty of studying behavior (the exploitation of public information) that would likely be considered unethical and therefore difficult to access through conventional qualitative methods (see, for example, Carter et al., 2021). Our

approach here, red teaming, is adopted from computer science contexts in which the goal is not to describe a process occurring in the world but to imagine and circumvent such a process. While it is beyond the scope of this short case study to further explore the methodological and epistemological implications of adopting red teaming in the context of social science research, we believe that it has potential value, especially in the early, exploratory stages of research when there is a need for the creation of initial frameworks that can be validated and modified by further work.

While we believe that the data and processes, we adopt have features that likely generalize to other contexts, a small, single case is necessarily limited. Studies that, for example, survey public datasets and classify the kinds of personal information found in them could help guide such work. Additional case studies that purposefully address a wide range of contexts could validate and extend the framework presented above.

About the authors

Daniel Carter is an Associate Professor in the school of journalism and mass communication at Texas State University. They can be contacted at dcarter@txstate.edu.

Caroline Stratton is a social scientist who studies technology, society, and policy topics.

References

Carter, D., Acker, A., & Sholler, D. (2021). Investigative approaches to researching information technology companies. Journal of the Association for Information Science and Technology, 72(6), 655–666. https://doi.org/10.1002/asi.24446

City of New York. (n.d.). NYC Open Data. NYC Open Data. Retrieved September 13, 2024, from http://nycod-wpengine.com/

Crain, M. (2018). The limits of transparency: Data brokers and commodification. New Media & Society, 20(1), 88–104. https://doi.org/10.1177/1461444816657096

Graham, F. S., Gooden, S. T., & Martin, K. J. (2016). Navigating the Transparency–Privacy Paradox in Public Sector Data Sharing. The American Review of Public Administration, 46(5), 569–591. https://doi.org/10.1177/0275074014561116

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management, 29(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Janssen, M., & van den Hoven, J. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? Government Information Quarterly, 32(4), 363–368. https://doi.org/10.1016/j.giq.2015.11.007

Liu, C.-Y., Li, W.-P., & Tu, Y.-P. (2021). Privacy Perils of Open Data and Data Sharing: A Case Study of Taiwan's Open Data Policy and Practices. Washington International Law Journal, 30(3). https://digitalcommons.law.uw.edu/wilj/vol30/iss3/8

Mansfield-Devine, S. (2018). The best form of defence–the benefits of red teaming. Computer Fraud & Security, 2018(10), 8–12.

Newman, L. H. (2024, August 16). The Slow-Burn Nightmare of the National Public Data Breach. Wired. https://www.wired.com/story/national-public-data-breach-leak/

Obama, B. (2013, May 9). Executive Order—Making Open and Machine Readable the New Default for Government Information. The White House, Office of the Press Secretary.

Srinivasan, S. (2017, October). Data Breach at Equifax. Harvard Business School. https://www.hbs.edu/faculty/Pages/item.aspx?num=53509

Stone, B. (2009, February 7). Prop 8 Donor Web Site Shows Disclosure Law Is 2-Edged Sword. The New York Times. https://www.nytimes.com/2009/02/08/business/08stream.html

Wickham, H. (2022). Package 'babynames' [Computer software]. https://cran.r-project.org/web/packages/babynames/babynames.pdf