Data&Society

The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards[1]

Case Study 03.17.16

NATHANIEL POOR, PHD, ROEI DAVIDSON, PHD

When the exact data you wanted but couldn't get presents itself to you thanks to hackers, can you use the data?

Introduction

We study crowdfunding [3,4] as part of our research into cultural production and the changes and challenges faced by producers in the internet age. Crowdfunding, via a website, is when someone has an idea for a project but needs help with the funding, and so throws out the idea to the wider internet audience, the crowd. In return for financial backing, supporters of a project can get various rewards or the product, depending on the website, as there are many crowdfunding websites. We have found that a project founder's network is important for repeat crowdfunding, as people are hesitant to ask friends

[1] This case study was first written for the Council for Big Data, Ethics, and Society. Funding for this Council was provided by the National Science Foundation (#IIS-1413864). For more information on the Council, see: http://bdes.datasociety.net/

and family for financial support more than once. Given that today's crowdfunding is mostly done online, it creates a great deal of digital data that can be studied if one has access to that data.

Data scientists at crowdfunding companies have access to all the data for that company. Some fortunate academic researchers have affiliations or research agreements with crowdfunding and other social data producing companies, thereby giving them access to the data they want. Yet, most academic researchers have no such luck and need to be able to get the data on their own by reading it off all of the relevant web pages for the projects at a crowdfunding website—potentially tens of thousands of pages. The scale makes this impossible for a research assistant. Instead, the process has to be automated with a "scraper": a computer script that acts like a web browser and saves each page you want (it "scrapes" the data from the web). Scrapers are fast and tireless, but aren't particularly smart, because you need to be able to tell them exactly which pages you want or how to get the URLs for each page. Furthermore, they often violate the Terms of Service of corporate websites.

In August, 2015, we planned to scrape data from the website Patreon, a crowdfunding website that focuses on long-term funding ("Patreon" from "patron"), an area we are particularly interested in. However, we couldn't find a nice hook into the website and the projects for scraping. An email inquiry to the website, Patreon, went unanswered. There would be no way to be sure we had retrieved data on all the projects at the site, and therefore we would have had to rely on a convenience sample and we could not be certain of our conclusions based on such a sample.

A Gift?

We were limited to such a sample until October, 2015, when Patreon was hacked [8]. The entire site was made available—not just all the data on projects, which we wanted, but apparently also private messages and the code that runs the site as well as email addresses and passwords. All of it, approximately 15GB of information. This was such a gift! Except, initially, we disagreed about whether it was appropriate to use it. The method the data was gathered under was not legal under US law, and included information meant to be private. Nevertheless, some of the data, before the hack, was already public, and this was the data we were interested in.

One of us felt that the data was now public, like a newspaper archive, and we could safely use it. The other pointed out that there were multiple ethical criteria that using the data would not meet. We discuss these issues below.

Guidelines: Cases and Literature

The proliferation of data and big data have caused challenges for both journalists and social scientists [2,9]. Many problems, sometimes overlooked, have arisen, such as the status of informed consent in big data [9], where data can be used for a variety of studies after collection for purposes not originally envisioned by the individuals providing the data.

In one prominent example, journalists published classified American security documents that Edward Snowden had released without authorization. Members of the Association of Computing Machinery (ACM) debated the ethics of Snowden's actions, and those of journalist Glenn Greenwald [1]. As they pointed out, Snowden's actions were clearly illegal and a breach of both workplace ethics and rules. Some pointed out how any whistleblower will be violating at least some ethical rules and perhaps laws as well, in the hope of stopping a greater wrongdoing. This example does not fully apply here, as there was no greater good served by hacking the website. And, we can accomplish our research in other ways via other websites or collection methods.

Another recent example is the hacking into citizens' cell phones by employees at Rupert Murdoch's News Corporation in the UK [6]. Both journalistic ethics and the law were violated, and both people's privacy and their expectation of privacy were infringed upon. In this case, the expectation of privacy was clear, as the hacked data was all private. The parallel illegality and invasion of privacy related to the Patreon data cannot be overlooked. Both the phone hacking and the use of the data in published news articles were a problem [6], suggesting that our use of hacked data, even with care about public and private information within that data, would be problematic.

Even instances of sharing and aggregating public data can cause an uproar, as shown when a New York state newspaper published names and addresses of gun owners in their readership area, retrieved via a freedom of information request [7]. Making public information more widely public is not always viewed as appropriate by those to whom the data relates, and public data was what we had hoped to use.

Journalists use data and information in circumstances where authorities and significant portions of the public don't want the data released, such as with Wikileaks, and Edward Snowden. This is a professionally accepted practice if done for the public good in a careful manner, much as we hoped to do. However, journalists also have robust professional norms and well-established ethics codes that the relatively young field of data science lacks. Although cases like Snowden are contentious, there is widespread acceptance that journalists have some responsibility to the public good which gives them latitude for professional judgment. Without that history, establishing a peer-group consensus and public goodwill about the right action in data science research is a challenge.

Similar to controversial journalism, hacked data allows researchers access to data ultimately useful to the public that companies are unwilling to share. Independent computer security researcher Mark Burnett had long collected user IDs and passwords from logins found in illegal data dumps released by hackers. Many students and academics asked him for his collection over time, and he decided to release his collection of 10 million logins publicly in order to facilitate research [5]. Companies with a large user base are typically unwilling to share login data with anyone because it exposes their users to harm. Burnett stripped the domain name from the logins, arguably providing more user security than the original data dumps. He claimed that no plausible further harm would come from releasing a consolidated dataset because all of the data was already available, yet it could be used by security researchers and professionals to improve security in the long run. Burnett claimed that the intent to improve knowledge should insulate him from the arguable illegality of his action, saying "It is beyond all reason that any researcher, student, or journalist have to be afraid of law enforcement agencies that are supposed to be protecting us instead of trying to find ways to use the laws against us" [5]. A year later, another group of security researchers would release a method for releasing large login datasets that are sufficiently anonymized without harming the integrity and utility of data, and persuaded Yahoo! to release password frequency data for 70 million users [10].

For further guidance, we read over the ethics statement of the Association of Internet Researchers (http://aoir.org/ethics/), but it did not address this type of situation directly. A query to the mailing list resulted in a lively discussion where no consensus was reached, although most participants cited reservations. A few said they wouldn't use it, but none said they would use the data without reservations. The most mentioned issues, both related to each other, were respect for the users and their consent (or permission) for our use of the data. One also mentioned how our use of the data might be construed as condoning illegal hacking.

These two issues both reflect the underlying, and important, issue of treating human beings not as anonymous, dehumanized research "subjects" or as a line of numbers in a file, but as human beings with feelings, autonomy, and agency in their lives. This agency applies to their actions, the data they create, and their ownership of that data. Discussing the paradox users may encounter between knowingly using a system that stores and makes available all their comments in a global manner, such as many online forums, and their expectations of privacy, Walther [11] directs our attention to "justice toward human subjects". The question is one of audience expectation regarding the data's use.

There are some paradoxical differences between online data and other, more traditional data types used for decades by media scholars, such as newspapers and magazines. Researchers have viewed journalistic content as fair game for use in research because it is considered public. Yet it is not easily gathered, as one must have access to either restricted databases available through a university library or access to the physical microfilm and microfiche. Online data, in contrast, is easily found by anyone

with an internet connection. Yet, as boyd and Marwick [2] have stated, in relation to the internet, "just because content is publicly accessible doesn't mean that it was meant to be consumed by just anyone.

Discussion

After having reflected on the experiences of journalists' uses of hacked data and the debates surrounding such use, we list several arguments for and against its use in the hope of spurring additional constructive debate:

Arguments in Favor of Use

1. Data is public, like a newspaper.
2. We hope to serve the public good via our work.
3. This is the data we want, but we can't get it via other methods.

Arguments Against Use

1. Researchers have a limited capability to distinguish between public and private information within the hacked data.
2. May see private data when cleaning the data.
3. Perhaps legitimizing criminal activity.
使合法化
4. Violating users' expectation of privacy.
5. Using people's data without consent.
6. We want this data, but we don't need it. Other data can be ethically collected and used.

These arguments *against* using the data, we feel, are much stronger than the arguments for using the data. Thus, in the end, we did not use the data copied and released by the hackers. Considering other cases and academic guidelines, we felt it would not be appropriate. Altogether, despite our hoping to do some good with the data and despite our hope to only use parts of it that were originally public, we felt the negatives outweighed the positives, especially when we could gather all or most of the same data in a more legal and more accepted manner.

Some cases of using data (or not) will be clear, other cases will not be. In the spirit of making lemonade out of lemons, we hope our case highlights some of the difficulties and considerations academics may encounter when contemplating the use of data.

Questions

1. Does the illegal nature of the data collection and the release of private data taint the data in the release that was already publicly available?
2. Users of Patreon initially had an expectation of privacy, but that privacy no longer exists. Do researchers need to respect the intent or the reality?
3. Scholars and journalists share some functions in dealing with information and making it accessible to the public, but are the ethical considerations the same? If not, why not?
4. Researchers will nearly always claim that their research will have a net public benefit and thus their methods are justified. Who gets to decide is that is accurate in any given case?

References

1. A. Adams. 2014. Report of a debate on Snowden's actions by ACM members. *SIGCAS Computers & Society* 44, 3: 5-7.
2. danah boyd, Kate Crawford. 2011. *Six provocations for big data*. Retrieved from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1926431
3. Roei Davidson, Nathaniel Poor. 2015. The barriers facing artists' use of crowdfunding platforms: Personality, emotional labor, and going to the well one too many times. *New Media & Society* 17, 2: 289-307.
4. Roei Davidson, Nathaniel Poor. 2016. Why sugar daddies are only good for Bar-Mitzvahs: Exploring the limits on repeat crowdfunding. *Information, Communication, and Society* 19, 1: 127-139.
5. Stuart Dredge. 2015. Security researcher publishes 10m usernames and passwords online. *The Guardian*, 11th February. Retrieved February 22, 2016 from: http://www.theguardian.com/technology/2015/feb/11/security-researcher-publishes-usernames-passwords-online-mark-burnett
6. Natalie Fenton. 2012. Telling tales: Press, politics, power, and the public interest. *Television & New Media* 13, 1: 3-6.
7. Jim Fitzgerald. 2013. Journal News removes controversial handgun permit information from website. *Associated Press*. Retrieved December 16, 2015, from http://www.huffingtonpost.com/2013/01/18/journal-news-handgun-removes-information_n_2507774.html
8. Dan Goodin. 2015. Gigabytes of user data from hack of Patreon donations site dumped online. *Ars Technica*. Retrieved December 9, 2015 from http://arstechnica.com/security/2015/10/gigabytes-of-user-data-from-hack-of-patreon-donations-site-dumped-online/
9. Seth C. Lewis, Oscar Westlund. 2015. Big data and journalism. *Digital Journalism* 3, 3: 447-466.
10. Byron Spice. 2016. Carnegie Mellon, Stanford Researchers Devise Method to Share Password Data Safely: Yahoo! Releases Password Statistics of 70 Million Users For Cybersecurity Studies. *Carnegie Mellon News*, 22nd February. Retrieved from February 22, 2016 from: http://www.cmu.edu/news/stories/archives/2016/february/sharing-password-data.html
11. Joseph Walther. 2002. Research ethics in internet-enabled research: Human subjects issues and methodological myopia. *Ethics and Information Technology* 4, 3: 205-216.

The Data & Society Research Institute Program on Ethics in "Big Data" Research will investigate the potential benefits and challenges put forward in this primer. Through partnerships, collaboration, original research, and technology development, the program seeks cooperation across sectors to innovate and implement thoughtful, balanced, and evidence-based responses to our current and future data-centered issues.

Data & Society is a research institute in New York City that is focused on social, cultural, and ethical issues arising from data-centric technological development. To provide frameworks that can help address emergent tensions, D&S is committed to identifying issues at the intersection of technology and society, providing research that can ground public debates, and building a network of researchers and practitioners that can offer insight and direction. To advance public understanding of the issues, D&S brings together diverse constituencies, hosts events, does directed research, creates policy frameworks, and builds demonstration projects that grapple with the challenges and opportunities of a data-saturated world.

Authors:
Nathaniel Poor: natpoor@gmail.com
Roei Davidson: roei@com.haifa.ac.il

For additional case studies:
Jacob Metcalf
bdes@datasociety.net
Data & Society Research Institute
36 West 20th Street, 11th Floor New York, NY 10011
Tel. 646-832-2038
datasociety.net