The Issue 32 incident – An update

Published
08 March 2016

Tags
ghtorrent
legal
openaccess

Many of you are aware of the GHTorrent issue 32. To sum up the discussion in a couple of lines, various developers included in GHTorrent wanted their email removed from it (which I did) and then wanted all emails to be excluded from the dataset (which I refused to do). The reasons behind the requests where *privacy* and the *right to do what ever one wants* with their personal data (email in many jurisdictions is considered personal data). What caused the whole thread was that researchers used GHTorrent as a source of emails for *research surveys* which were sent to thousands of developers, which made some developers nervous with respect the use of their personal data.

As a response to the issue, I had temporarily shut down access to GHTorrent until the situation is cleared up. In the mean time, I have consulted with my employer's legal department and several experts, mostly on the legal and data protection domain. The issue even caught the interest of respected ICT laywer Arnoud Engelfriet and caused interesting meta-discussion on a legal forum (Dutch).

Here is what I learned.

Personal data

Personal data are defined as "any information relating to an identified or identifiable natural person". From the information that GHTorrent collects, emails and real names can be considered as *personal data*.

As GHTorrent is processing personal data (e.g. linking users with their actions), it must comply with data protection legislation. This includes:

- controlling access to and distribution of personal data
- informing subjects about how their personal data is used
- enable subjects that are not in agreement with use of their personal data from the project to *opt-out* of the collection
- enable subjects to have their data completely removed (right to be forgotten)

Now, here is a catch: GHTorrent started as a project in Europe but has since moved to the US (on the East-US Azure datacenter). Which legislative domain should the project comply to? Most lawyers I've asked said that being on the safe side is the best way to go, which means adhering the union of provisions of both European and US data protection laws.

Privacy

GHTorrent is not required by law to create an *opt-in* mechanism. Personal data has been shared by subjects publicly and as GHTorrent is not breaking the terms of use of GitHub, it can download and process them, if it complies with data protection laws.

Moreover, GHTorrent has no liability on what its users are going to do with the data; as a top consultant in our University's legal department stated it, "a store owner is not responsible if you buy a hammer and kill someone".

Ethical considerations

The root of the problem lies in the domain of ethics; are we researchers allowed to use databases such as GHTorrent to target specific developers with our research surveys? Can we profile developers using their public activity traces? Can we rank developers based on how good code they write? Can we recommend work to developers (e.g. solving issues), based on their expertise on specific project areas, programming techniques or other projects? There are no easy answers. Personally, I have run surveys twice and even provided guidance on how to do it at scale.

One safeguard that we can employ is *anonymity*: we can do research with publicly available traces BUT we should make sure that the results (when shared) do not reveal the true identity of developer. In turn, this might have negative implications in the public replicability of the results or their re-use. To address this, replication packages can be offered in a tiered manner: i) anonymized data with public access, and ii) data that may reveal the identities of people under agreements of non-disclosure.

A course of action

To comply with legal requirements, GHTorrent will do the following:

- No real names or emails will be shared in the MySQL dumps or the online MySQL access services. All other database fields will be left untouched. Researchers interested to mine GitHub users can do so using the `login` field as reference. Getting the email is a GitHub API call, but it will be the responsibility of the researcher to go the extra mile. GHTorrent itself will continue to use emails internally, to ensure consistency of the data.
- GHTorrent will offer to interested researchers an additional download that links logins to personal data (essentially, a CSV file with 3 fields: `login`, `email`, `name`). To obtain access to it, researchers will need to agree to the terms of use and publicly state the indented use, by submitting a pull request here.
- A FAQ with specific privacy-related questions/answers will be created and shared on GitHub. People can submit their questions in this page.
- An opt-out process will be created; developers who do not want to be tracked will be able to do so by submitting a form on the GHTorrent web site. This will have the effect of their `name` and `email` fields being replaced with random strings.
- The research community will need to work on a code on conduct wrt the use of personal data for research. Underground processes have started already; we should co-ordinate those efforts and will hopefully come up with a document during this year's MSR conference.

The changes (esp the opt-out process) will need some time to be implemented; in the mean time, GHTorrent will re-enable online access to the datasets (excluding email addresses and real names in MySQL of course), but will not offer downloads until all the measures above have been put into place.

Personal reflection

In the heat of the discussion, I have made a couple of statements that i) were wrong ii) made the discussion hotter. Initially, I was wrong in that copyrighted content shared on the web without a license is in the "public domain". The opposite is actually true: copyrighted content on the web with no accompanying license means that all rights are reserved by the copyright holder (usually the content creator). Moreover, I said that "GHTorrent tracks public data", while emails and names are in fact personal data, as they identify a person uniquely. Finally, while as a researcher I witnessed and documented the negative effect of online conflict on GitHub, I still managed to fall into the same trap.

I would like to thank Efthimia Aivaloglou, Arie van Deursen, Arnoud Engelfriet, Sean Lang, Leon Moonen, Mario van der Toorn, Bogdan Vasilescu, and Alexis Zavras for advice and support those hot last two weeks.

← Previous | Archive | Next →