Faster homomorphic encryption for private data processing

A Data Management Plan created using DMPonline.be

Creator: Robin Geelen https://orcid.org/0000-0003-4684-3532

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / **URL:** 1162123N

ID: 192231

Start date: 01-11-2022

End date: 31-10-2024

Project abstract:

Homomorphic encryption allows computations on encrypted data without revealing its content. Although it has a wide range of applications in privacy preserving data processing, current techniques are impractical due to extreme overheads compared to unencrypted computations. The PhD aims to reduce the computational cost of homomorphic encryption at three levels: algorithmic improvements, software and hardware implementations and applications.

Last modified: 12-01-2023

Faster homomorphic encryption for private data processing
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA
Software	Open source software and scripts	Generate new data	Digital	Software	.c, .py, .m, ...	<100GB
Project deliverables	LaTeX and Microsoft Word files	Generate new data	Digital	Other	.tex, .docx, .pdf	<1GB
Visuals	Figures, graphs and media	Generate new data	Digital	Compiled/aggregated data	.jpg, .png, ...	<100GB
Data sets	Machine learning data sets used for evaluation and measurements	Generate new data and reuse existing data	Digital	Observational, experimental and simulation data	.csv	<10TB

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Machine learning data sets are publicly available data sets used for evaluation of research results: examples include MNIST and CIFAR.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, dual use

No personal data will be collected, but there is the possibility of dual use or misuse. Cybersecurity technologies and machine learning have the potential for military applications and massive unethical data processing. However, this project focuses exclusively on defensive technologies and theoretical evaluation of algorithms and protocols. Moreover, only publicly available data sets will be used for evaluation of research results.

My ethical approval reference number is D-20221213.c.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Algorithms and protocols are self-explanatory. Comments will be integrated in the computer code to improve readability. The source code will be accompanied by README files that explain the structure, the compilation order and the overall architecture.
Visual data will be kept with the project deliverables, and both are self-explanatory.
Documentation of data sets can be found online. Measurement results will be documented with a text file.

Our research group has a detailed roadmap on how to store research data so that it can easily be reused. See the section about data storage and backup below.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

See the section about documentation above.

3. Data storage & back-up during the research project

Where will the data be stored?

Data used to produce project results will be stored in a GitLab repository with access control. The GitLab repository is hosted by a server at our research department. In a later phase, and after the publication of the research paper concerned, the relevant data will be made available in a public GitHub repository.

How will the data be backed up?

We make sure to replicate the data sets among at least two department computers.
The GitLab repository is hosted by a server at our research department and gets a daily backup.
GitHub is a well-established, commercial repository service provider, which we assume to have industry-standard data availability measures in place (such that repositories will not be lost unless deleted by authorized GitHub users).

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

There is sufficient storage & backup capacity during the project. We have many terabytes of storage space available on the department machines that we use to replicate our data sets. The department's GitLab instance and GitHub are offered services and are thus ensured to have enough spare capacity.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Access to our department machines is managed by access to the university network and standard-practice user permissions on the machine (only users with the right set of permissions can access the data, we control the management of these users).
The GitLab repository has access control and is maintained by a DMP officer who keeps the structure in the repository in place and manages the access control. Unauthorized GitHub modifications are prevented through their industry-standard user management.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The costs of the GitLab maintenance are covered by our research group. These costs are integrated in the general IT costs. The use of GitHub is free of charge.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All the research data will be stored on the GitLab repository for at least 10 years after the end of the project. A selection of the research data used to produce project results will be pushed to GitHub regularly. This will happen after the publication of a project paper. The GitHub will remain active for at least 10 years after the end of the project as well.

Where will these data be archived (stored and curated for the long-term)?

To the extent feasible in terms of the volume of collected data, the data will be stored in the GitLab and GitHub repositories. If we consider it necessary to purchase archival storage capacity for the entirety of the research project's artifacts (e.g., due to non-reproducibility of some data sets and limitations on the GitLab and GitHub services), we will do so appropriately and in time before the project concludes. Options include the storage services offered by KU Leuven or dedicated storage hardware to be kept long-term at our department.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The costs of the GitLab maintenance are covered by our research group. These costs are integrated in the general IT costs. The use of GitHub is free of charge.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Other, please specify:

Relevant data used to produce project results will be made available on GitHub. Relevant data can consist of software or hardware scripts and code, algorithms, protocols, figures and others.

If access is restricted, please specify who will be able to access the data and under what conditions.

There are no restrictions in place.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Aspects of dual-use

To mitigate the risk of dual use and misuse, only publicly available data sets will be used for evaluation of research results. There are no confidentiality concerns related to this data.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All source code, protocols, and algorithms used for published articles will be made available as open source software on GitHub.

When will the data be made available?

To the extent feasible due to volume of the collected data sets, we will make our data sets publicly available on GitHub upon publication of the corresponding article. In case a data set can be reproduced, we will also publish clear instructions/scripts for how to do so (so that a large data set may also be recreated instead of downloaded).

Which data usage licenses are you going to provide? If none, please explain why.

We will publish software and hardware implementations under Apache License.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

No costs are to be expected.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The PhD student, Robin Geelen, funded by this grant, is responsible for data documentation and metadata.

Who will manage data storage and backup during the research project?

In the first place, the PhD student, Robin Geelen, funded by this grant, is responsible for data storage and backup. The DMP officer maintaining the GitLab repositories will share in the responsibility for data storage and backup.

Who will manage data preservation and sharing?

The PhD student, Robin Geelen, funded by this grant, is responsible for data preservation and sharing.

Who will update and implement this DMP?

The PhD student, Robin Geelen, funded by this grant, is responsible for updating and implementing this DMP.