Richtlinie zum „Software-Sprint“

DeKUF – Decentralizing KUserFeedback

Schlussbericht

Zuwendungsempfänger:

Balazs, Dahlke, Stoffels GbR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen **01IS24S05** gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Die aktuelle Praxis der Speicherung persönlicher digitaler Daten in sogenannten „Clouds“, also auf Rechnern, die nicht unter der persönlichen Kontrolle der eigentlichen Dateneigentümer:innen stehen, birgt große Probleme. Auch wenn die Politik Rahmenbedingungen für die Speicherung und Verarbeitung dieser Daten setzt, kommt es immer wieder zu Missbrauch oder Diebstahl dieser Daten mit unangenehmen Folgen für die Dateneigentümer:innen, aber auch für die Gesellschaft als Ganzes.

Gleichzeitig sind personenbezogene Daten Quelle und Motor für Innovationen. Dies gilt sowohl für die Qualität persönlicher Dienstleistungen als auch für die Beantwortung gesamtgesellschaftlicher Fragen. Persönliche Daten müssen daher dem Einzelnen direkt und der Gesellschaft in anonymisierter Form zur Verfügung stehen.

Das Projekt DeKUF - als Teil des übergeordneten Projekts privact.org - versucht nun dieses Problem grundsätzlich zu lösen, indem persönliche digitale Daten nur unter der Kontrolle der Dateneigentümer:innen - idealerweise lokal - gespeichert werden. Der Hauptfokus des DeKUF-Projekts lag dabei auf der Demonstration der anonymen Auswertung von Daten, die bei vielen Nutzer:innen verteilt gespeichert sind.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Privatsphäre ist ein Grundpfeiler jeder freien und offenen Gesellschaft. Das Projekt privact.org will dieses Grundrecht allen Menschen unabhängig von ihren technischen Fähigkeiten auch im digitalen Leben ermöglichen. Dazu arbeiten wir an der Schaffung des dafür notwendigen Ökosystems.

Das DeKUF-Projekt hat es uns ermöglicht, wesentliche Grundlagen für den größeren Projektkontext zu legen. Konkret konnten wir zeigen, dass es möglich ist, verteilte und unter der Kontrolle der einzelnen Nutzer:innen liegende persönliche Daten auszuwerten, ohne die Anonymität des Einzelnen zu tangieren.

Im Rahmen des Gesamtprojektes wird also jeder Mensch - nicht nur in Deutschland oder der EU - in die Lage versetzt, zu verstehen und zu kontrollieren, welche Daten über ihn von welchen Organisationen zu welchem Zweck und für welchen Zeitraum gesammelt werden. Sie können diese Daten verwalten, ihre Einwilligung zur Datenspeicherung widerrufen und selbst bestimmen, in welche gesellschaftlichen Forschungsfragen ihre Daten einfließen sollen. All dies wird auf eine Art und Weise geschehen, die für jeden möglich und praktikabel ist. Das Projekt stärkt somit die Zivilgesellschaft als Ganzes und befähigt jede:n Einzelne:n, selbstbestimmt und sicher an der digitalen Welt teilzuhaben.

Ausführliche Darstellung der Ergebnisse

Wir haben alle gesetzten Meilensteine erreicht. Konkret haben wir die Bibliothek KUserFeedback [https://phabricator.kde.org/source/kuserfeedback/] erweitert. Diese ermöglicht es KDE-Anwendungen, Telemetriedaten zu sammeln und auszuwerten. Dabei haben wir den bisherigen Mechanismus, Rohdaten an einen KDE-Server zu senden, durch eine datenschutzfreundlichere Variante ersetzt. Die Rohdaten werden in einem lokalen Client auf den Rechnern der Anwender:innen gespeichert. Die Auswertung dieser verteilten Daten wird dann über einen Webservice angestoßen. Das „Federated Secure Aggregation Protocol“ garantiert dabei, dass die Privatsphäre der Anwender:innen geschützt bleibt, während das Ergebnis der Auswertung an KDE zurück gemeldet wird.

Dazu haben wir im Projekt Meilensteine in zwei Bereichen bearbeitet: KDE und Web-Infrastruktur.

Für KDE haben wir einen lokalen Client gebaut, der Datenpunkte über D-Bus empfangen kann. Dazu haben wir KUserFeedback prototypisch so erweitert, dass es erste Datenpunkte über D-Bus an den Client sendet. Diese Datenpunkte müssen derzeit einem sehr einfachen Datenschema entsprechen: dem „Zählen von Vorkommnissen“. Der Client kann zusätzlich Anfragen beantworten, die vom Webserver über das

Federated Secure Aggregation Protocols (siehe unten) initiiert werden. Schließlich können die Anwender:innen im Client sehen, welche Daten darin gespeichert und welche Queries beantwortet wurden.

Für den Webserver wurde eine Benutzeroberfläche implementiert, die es KDE-Entwicklern ermöglicht, die in KUserFeedback erfassten und verteilt liegenden Daten abzufragen. Dabei kann konfiguriert werden, welche Datenpunkte ausgewertet werden sollen und die entsprechenden Auswertungsgruppen pro Datenpunkt definiert werden. Zusätzlich kann die gewünschte Mindestanzahl an Teilnehmern festgelegt werden und in wie viele Untergruppen diese aufgeteilt werden sollen.

Die so definierten Queries können durch die Anwender:innen auf der Weboberfläche gestartet werden. Der Webserver übernimmt dann die Durchführung unter Beachtung des „Federated Secure Aggregation Protocol“ und stellt nach erfolgreicher Durchführung die Ergebnisse für die Anwender:innen dar.

Das „Federated Secure Aggregation Protocol“ war das Herzstück und der Hauptteil unserer Arbeit. Das Protokoll beschreibt einen Kommunikationsprozess, bei dem viele Clients mit dem Server so zusammenarbeiten, dass die Daten durch den Einsatz von asynchroner und homomorpher Verschlüsselung die lokalen Clients nie in „roher“ Form verlassen.

Zunächst bestimmt der Server zufällig einen der beteiligten Clients als „Aggregation Client“. Dieser erzeugt dann ein asynchrones Schlüsselpaar zur Verschlüsselung der Nachrichten. Der öffentliche Schlüssel wird an den Server gesendet. Der Server erzeugt nun ein homomorphes (Paillier) Schlüsselpaar und sendet das eigentliche Query sowie die beiden öffentlichen Schlüssel an alle Clients. Diese beantworten das Query, verschlüsseln die Daten mit dem homomorphen öffentlichen Schlüssel und die gesamte Nachricht mit dem asynchronen öffentlichen Schlüssel. Die so doppelt verschlüsselte Nachricht wird über den Webserver an den Aggregation Client weitergeleitet. Dieser kann die Nachricht mit dem privaten asynchronen Schlüssel entschlüsseln und die immer noch homomorph verschlüsselten Daten addieren. Das aggregierte Ergebnis - immer noch homomorph verschlüsselt - wird vom Aggregation Client an den Server weitergeleitet. Dieser besitzt den privaten Schlüssel und kann nun das homomorph verschlüsselte aggregierte Ergebnis entschlüsseln.

Durch dieses Protokoll ist die Privatsphäre der Anwender:innen stets geschützt, da die Rohdaten niemals unverschlüsselt den lokalen Client verlassen und die Kombination der beiden Schlüsselsysteme dafür sorgt, dass die Daten stets nur in aggregierter und damit nicht mehr individuell identifizierbarer Form entschlüsselt werden.

Die Entwicklung des Protokolls war Forschungsarbeit und daher von vielen Richtungswechseln und Tests begleitet. Mehr dazu im Abschnitt „Kurze Darstellung der Arbeiten, die nicht zu einer Lösung geführt haben“.

Darüber hinaus haben wir durch die Schulungen der Open Knowledge Foundation und Superbloom wertvolle Hinweise für den Aufbau und die Weiterführung des Projektes erhalten. Besonders hervorheben möchten wir die Themen Namensfindung, Aufbau einer Community sowie Möglichkeiten und Strategien für weiteres Sponsoring. Auch wenn diese während des laufenden Projekts noch nicht zum Tragen kommen, helfen sie uns enorm, das Projekt auch nach Auslaufen der Förderung weiterzuführen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Das Projekt DeKUF ist ein Grundlagenprojekt innerhalb des größeren Projekts privact.org. Ein unmittelbarer Nutzen ergibt sich aus der prototypischen Lösung, die im Rahmen des Projektes erstellt wurde, noch nicht.

Ganz konkret wird jedoch gemeinsam mit dem KDE Projekt daran gearbeitet, dass die prototypische Lösung für KDE zu einem ersten Produkt weiterentwickelt und upstream ausgerollt wird. Damit können die KDE Nutzer:innen direkt in die Weiterentwicklung des KDE Projekts eingebunden werden, ohne Details ihrer Privatsphäre preisgeben zu müssen. Davon profitiert insbesondere die UX des KDE Projektes, die dann mit wesentlich relevanteren Daten die Weiterentwicklung vorantreiben kann. Und davon profitieren wiederum die Nutzer:innen von KDE.

Diese Integration wie auch die weiteren Pläne sind nur durch die Lizenzierung unter der GPL v3 möglich, da so die prototypischen Lösungen in andere Projekte einfließen können.

Diese weiteren Pläne konzentrieren sich zunächst darauf, die Lösung auch anderen Freien Software Projekten zur Verfügung zu stellen und mit diesen gemeinsam das eigentliche Produkt weiterzuentwickeln und für verschiedene Plattformen zur Verfügung zu stellen. Im Gegensatz zum Backend des Internets, das heute von Freier Software dominiert wird, spielt Freie Software im Frontend eine eher untergeordnete Rolle. Ein Grund für diese Situation ist der Respekt der Privatsphäre der Nutzer:innen durch Freie Software-Projekte und der daraus resultierende Mangel an Daten und Zugängen für eine nutzer:innenzentrierte Entwicklung (UX). Diese Lücke können wir mit der im DeKUF entwickelten und von privact.org weitergeführten Technologie schließen. In der Folge erhoffen wir uns eine deutliche Verbesserung der UX vieler Freier Software Projekte. Dies wird den vielfältigen Anwender:innen dieser Projekte zu Gute kommen.

Längerfristig werden wir die Technologie und das mit privact.org zu entwickelnde Ökosystem auch kommerziellen Diensten zur Verfügung stellen. Damit schaffen wir einen Rahmen, in dem Unternehmen zum Beispiel die Anforderungen des EU-Datenschutzgesetzes erfüllen können. Wenn wir diesen Schritt geschafft haben, sind wir am Ziel, allen Menschen einen selbstbestimmten Umgang mit ihren persönlichen Daten im digitalen Leben zu ermöglichen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Die Hauptforschungsarbeit im DeKUF-Projekt bestand darin, ein Protokoll für die Analyse verteilter Daten zu finden. Das Protokoll sollte die Rohdaten und damit die Privatsphäre schützen, aber Aussagen über die persönlichen Daten einer größeren Stichprobe von Anwender:innen ermöglichen. Dazu wollten wir die Prinzipien der Föderierung und Verschlüsselung nutzen.

Zunächst haben wir versucht, auf den Ansätzen der Datenaggregation aus dem Bereich des föderierten Lernens aufzubauen. Siehe zum Beispiel: Secure Aggregation with Flower [https://flower.ai/docs/examples/flower-secure-aggregation.html]. Wir mussten jedoch feststellen, dass dieser Ansatz sehr stark darauf ausgelegt ist, mit Trainingsvektoren des maschinellen Lernens zu arbeiten, die typischerweise Gewichte zwischen 0 und 1 repräsentieren. Dies entspricht nicht den Eigenschaften der Daten, mit denen wir arbeiten wollen. Außerdem ist das Lernen in diesem Ansatz föderiert, die Aggregation der Daten jedoch nicht.

Diese Faktoren haben dazu geführt, dass wir diesen Ansatz nach eingehender Evaluierung verworfen haben. Stattdessen haben wir ein eigenes Protokoll entwickelt, das bekannte und erforschte Mechanismen so kombiniert, dass wir den gewünschten Effekt erzielen: Das Federated Secure Aggregation Protocol.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Für das Gesamtprojekt betreiben wir die Website www.privact.org. Hier können sich Interessierte umfassend über den aktuellen Stand informieren. Auf der Webseite finden sich auch Links zu unserem Diskussionsforum talk.privact.org sowie zu unserem Code Repository gitlab.com/privact.

Eine wissenschaftliche Veröffentlichung über das von uns entwickelte „Federated Secure Aggregation Protocol“, also das Herzstück der dezentralen und anonymitätswahrenden Auswertung der persönlichen Daten der Anwender:innen, ist in Planung.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Das Projekt war in weiten Teilen als Forschungsprojekt angelegt. Die zentrale Forschungsfrage war, wie die anonymisierte Auswertung von verteilten personenbezogenen Daten realisiert werden kann. Das Ganze wurde in das konkrete

Projekt KUserFeedback eingebettet, um den Forschungsergebnissen einen direkt verwertbaren Rahmen zu geben.

Im Großen und Ganzen hat unsere ursprüngliche Planung dafür recht gut funktioniert. Allerdings haben wir uns in Absprache mit dem Projekt KUserFeedback dazu entschlossen, etwas weniger Aufwand in die konkrete Einbindung des Projektes zu stecken, um insbesondere den Auswertungsserver und die Förderung der Auswertung noch weiter zu verbessern. Der Hintergedanke dabei war, dass das KUserFeedback-Projekt die umfassendere Integration unserer Technologie selbst vornehmen wollte und stärker von den Weiterentwicklungen in den anderen Bereichen profitieren sollte.

Infolge dieser Entscheidung kam es zu einer leichten Verschiebung des ursprünglich geplanten Arbeitsaufwandes von Felix Dahlke zu Richard Stoffels, der im Projekt vorrangig für die Entwicklung des Auswertungsservers und der föderierten Auswertung zuständig war.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Während der Projektlaufzeit gab es keine Entwicklungen oder Ergebnisse von anderen Stellen, die unsere Arbeit beeinflusst hätten.

Allerdings hat Mozilla etwa zeitgleich mit dem Abschluss unserer Arbeit „Privacy friendly Attribution for Advertising Measurement“ veröffentlicht: https://support.mozilla.org/de/kb/datenschutzfreundliche-attribution-werbe-messung.

Mozilla verwendet dabei sehr ähnliche Mechanismen, wie wir sie in unserem endgültigen Federated Secure Aggregation Protocol vorschlagen. Der Hauptunterschied liegt in der Form der Föderierung bei der Datenauswertung. In unserem System werden zufällig einzelne Clients ausgewählt, die die Aggregation der verschlüsselten Daten durchführen und diese dann an einen zentralen Server senden, der die aggregierten Daten entschlüsseln kann. Das System von Mozilla verwendet dagegen verschiedene Server, die nur gemeinsam die Daten aggregieren und entschlüsseln können. Zusätzlich verwendet Mozilla das Prinzip der „differential privacy“ [https://en.wikipedia.org/wiki/Differential_privacy], das eine Ungenauigkeit in die Daten einbringt und auf das wir aufgrund der von uns verwendeten Föderierung und anderer Mechanismen verzichten können.

Mozilla plant, ihr „Distributed Aggregation Protocol for Privacy Preserving Measurement“ als Web-Standard zu etablieren [https://datatracker.ietf.org/doc/html/draft-ietf-ppm-dap].

Wir freuen uns sehr, dass Mozilla und andere zu einer sehr ähnlichen Lösung wie wir gekommen sind, um das grundlegende Problem zu lösen. Das schafft Vertrauen in die

Validität der von uns gefundenen Lösung. Wir werden den Kontakt zu Mozilla suchen, um im Idealfall die weitere Entwicklung mit ihnen abzustimmen. Wir werden weiterhin versuchen, unsere Gründe für die clientseitige Föderierung in den Prozess der Webstandardisierung einzubringen.