Der Daten-Schutzmantel

Daten sind der Rohstoff der Informationsgesellschaft. Oft genug aber geraten Unternehmen, die nicht auf umfassende Datenanalysen verzichten wollen, in Konflikt mit dem Datenschutz. Paul Francis, Direktor am Max-Planck-Institut für Softwaresysteme in Kaiserslautern, sucht einen Ausgleich zwischen den gegenläufigen Interessen. Sein Unternehmen Aircloak spielt dabei eine wichtige Rolle.

TEXT CHRISTIAN J. MEIER

er Blick hinter die Kulissen des Internets ist ernüchternd. Ein kostenloses Programm namens Ghostery zeigt an, wer mein Surfverhalten verfolgt. Paul Francis vom Max-Planck-Institut für Softwaresysteme in Kaiserslautern hat es mir empfohlen, als wir uns ein paar Tage zuvor in einem Café nahe dem Institut gegenübersaßen.

Mit seinem verschlissenen Ampelmännchen-T-Shirt und einem karierten Hemd wirkt Francis wie ein legerer, etwas in die Jahre gekommener Computersonnyboy aus dem Silicon Valley. Es verwundert daher nicht, dass der Wissenschaftler nicht nur forscht, sondern gleichzeitig ein Start-up in Kaiserslautern betreibt. Sowohl seine Forschung als auch sein Unternehmen widmet er einem besseren Schutz der Privatsphäre von Internetnutzern.

Dabei betrachtet Paul Francis sein Start-up namens Aircloak als ein Forschungsinstrument. Den kommerziellen Erfolg der Firma sieht er als Gradmesser für die Fortschritte seiner Forschung. Was Francis mit seinem Start-up macht, ist eine Art Expedition in die reale Welt des Internets. Und die ist ein Dschungel, in dem Hunderte von Firmen eifrig Daten über Surfer sammeln. Diese Dienstleister haben sich darauf spezialisiert, die Onlinewege der Internetnutzer zu verfolgen. Die Daten verkaufen sie an Unternehmen, die damit etwa ihre Werbung optimieren können.

Davon bekomme ich, wieder zu Hause am Schreibtisch, schnell einen Eindruck: Sechs "Tracker" zeigt Ghostery nach dem Klicken auf einen Onlineartikel eines Nachrichtenmagazins. Nach den Besuchen von ein paar weiteren Seiten, etwa einer Suchmaschine für Flüge oder von Facebook, habe ich schon etwa zwanzig verschiedene solcher Tracker identifiziert.

DAS FALSCHE VERSPRECHEN DER **NUTZER-ANONYMITÄT**

Die Tracker liefern den Datensammlern die Information, wer welche Seite besucht. Zwar bleibt der Nutzer dabei eine Nummer, aber immer die gleiche Nummer: Es lässt sich verfolgen, welche Websites der Nutzer mit der Nummer X besucht. "Die Firmen legen bei jedem Besuch einen Datensatz an", erklärt Francis. Anhand der so entstehenden Datenbank lässt sich das Surfverhalten von Nutzer X untersuchen. Das ist für gezielte Werbung nutzbar, die X bestmöglich bei seinen Vorlieben packt.

"Es ist unglaublich", sagt der aus den USA stammende Informatiker und schüttelt den Kopf, bevor er erklärt, wie die gezielte Werbung funktioniert. "Nehmen wir an, Sie kommen auf eine Website, die Platz für eine Anzeige hat, und mehrere Firmen wollen Ihnen ihre Werbung zeigen", sagt er. "All diese Firmen machen dann bei Google ein Angebot. Der Meistbietende kommt zum Zug."

Wo ist das Problem?, möchte man einwenden. Die Daten sind ja anonymisiert. Niemand weiß, dass es, sagen wir, Paul Francis oder Christian J. Meier sind, die diese oder jene Websites gern besuchen. Es sind Hausnummer eins oder Hausnummer zwei. Die Privatsphäre bleibt gewahrt.

Doch so einfach sieht Francis die Sache nicht. Er spricht von einem falschen Versprechen der Nutzeranonymisierung, das da lautet: Wenn Daten erst einmal anonymisiert sind, kann niemand etwas über ein bestimmtes Individuum herausfinden.

Lückenhafter Persönlichkeitsschutz: Durch die geschickte Kombination von Daten aus verschiedenen Quellen kann sich wie in unserem fiktiven Beispiel ein umfassendes individuelles Profil ergeben. Forscher des Max-Planck-Instituts für Softwaresysteme wollen das verhindern.



Der Weg in die Anonymität: Die Cloak garantiert, dass Analysten aus Datensätzen keine Individuen herausfiltern können. Hinter der Cloak, einem undurchdringlichen Schutzmantel für Informationen, werden Nutzerdaten aus einer oder mehreren Quellen verwaltet. Ehe sie dorthinfließen, erhält die Datenquelle eine Beglaubigung, dass die Informationen tatsächlich in die Cloak fließen und nicht zu einem Ort, der lediglich vorgibt, eine solche zu sein. Der Analyst erhält auf seine Anfragen anonymisierte Antworten, die hinter der Cloak ermittelt werden. Dabei wehrt das Unternehmen Aircloak, das die Cloak betreibt, auf Individuen zielende Anfragen ab. Ausschließlich seine Auditoren kontrollieren den Zugang zu den Daten.

Die Brisanz steigt außerdem dadurch, dass es neben den Firmen, die das Surfverhalten einer Person X kennen, andere Unternehmen gibt, die weitere Informationen über die Person besitzen: Die Bank kennt ihre finanzielle Situation, der Energieversorger ihren Energieverbrauch, die Kreditkartenfirma erfährt einiges über das Konsumverhalten von X. Der Mobilfunkanbieter weiß, wann sich X wo aufgehalten hat. "Oft verkaufen die Firmen Daten über ihre Kundschaft", erklärt Francis. Er wisse von Fällen aus den USA, in denen Banken anonymisierte Kundendaten an andere Organisationen gaben. Prinzipiell ist es also möglich, dass ein Käufer all diese Daten in einen Topf wirft und so ein umfassendes Bild über X gewinnen kann. Der Verbraucher wird gläsern. "Die Daten werden zwar zu harmlosen Zwecken gesammelt, aber je nach Käufer kann die Sache schwerwiegend werden", warnt Francis.

Ein Datenschutz-Guerillero ist der Informatiker trotz solcher Szenarien nicht. An sich verteidigt er die Analyse anonymisierter Nutzerdaten. Sie könne sehr nützlich sein, meint er und nennt ein Beispiel: "Im medizinischen Bereich verursacht Betrug Milliardenschäden", sagt Francis. In medizinischen Datenbanken könnte man Betrugsfällen auf die Spur kommen, zum Beispiel indem man die Verschreibungen untersucht. Gibt es Ärzte, die besonders viel verschreiben? Oder die Medikamente verschreiben, die sie nicht verschreiben sollten?

Doch Anonymisierung allein helfe nicht, die Privatheit des unbescholtenen Gros der Ärzte zu schützen. "Es ist schwierig, alle medizinischen Daten in eine große Datenbank zu packen, ohne die Privatheit zu gefährden", sagt Francis. Daher werde dieses Potenzial nicht genutzt.

Zwei spektakuläre Fälle aus der Vergangenheit zeigen, was Francis meint. Sie demonstrieren, dass mitunter sogar Institutionen, denen man fundiertes Know-how im Datenschutz zutraut, leicht zu überlisten sind.

KOMBINIERTE DATEN IDENTIFIZIEREN INDIVIDUEN

Ende der 1990er-Jahre veröffentlichte eine staatliche Agentur im US-Staat Massachusetts, welche die Krankenversicherungen staatlicher Angestellter verwaltete, Daten über die Versicherten, damit Forscher diese nutzen konnten. Die Agentur glaubte die Privatsphäre der Staatsbediensteten zu schützen, indem sie den Namen, die Sozialversicherungsnummer und andere "ausdrückliche Bezeichner" jeder Person aus den Daten entfernte. Auch der damalige Gouverneur von Massachusetts, William Weld, garantierte der Öffentlichkeit, dass die Privatheit der Versicherten damit geschützt sei.

Er hatte nicht mit der pfiffigen Informatikstudentin Latanya Sweeney gerechnet. Für zwanzig Dollar kaufte sie sich das Wählerverzeichnis der Stadt Cambridge bei Boston, wo Weld wohnte. Darin standen Namen, Adressen, Postleitzahl, Geburtsdatum und Geschlecht jedes Wählers. Mit Leichtigkeit konnte sie so den Gouverneur in den

von der Agentur herausgegebenen Versichertendaten finden: Nur sechs Einwohner von Cambridge in den vermeintlich anonymisierten Krankenversicherungsdaten teilten seinen Geburtstag, drei davon waren Männer, von denen nur einer seine Postleitzahl hatte – der Gouverneur selbst. Öffentlichkeitswirksam sandte Sweeney dem Gouverneur seine Akte, samt der darin enthaltenen Diagnosen und Verschreibungen.

Ein paar Jahre später, 2006, veröffentlichte der Onlinedienst AOL zwei Millionen Suchanfragen von 650000 Nutzern. Forscher freuten sich über diese Möglichkeit, das Internetverhalten sehr vieler Nutzer anhand einer so riesigen Datenmenge untersuchen zu können. AOL anonymisierte die Daten: Das Unternehmen entfernte Nutzernamen, IP-Adressen, die den Computern zugeordnet sind, und andere Informationen, die eine direkte Identifizierung von Nutzern ermöglichten. Jeder Nutzer wurde allerdings mit einer eindeutigen Nummer versehen, damit die Daten für die Forschung wertvoll blieben.

Diesmal waren es zwei Journalisten der New York Times, die AOL zeigten, dass diese Anonymisierung keinen perfekten Schutz der Privatsphäre bot. In Anfragen des Nutzers 4417749 fanden sie Hinweise auf dessen Identität. Es gibt nämlich nicht viele Nutzer, die gleichzeitig einen Landschaftsgärtner in "Lilburn, GA", und nach einem zum Verkauf stehenden Haus in "Shadow Lake, Georgia", suchten. Die Journalisten identifizierten eine gewisse Thelma Arnold hinter den Anfragen. Arnold bestätigte, die Suchphrasen eingegeben zu haben, darunter auch Peinliches wie "Hund, der auf alles uriniert".

Die Moral dieser Geschichten: Ein listiger, vielleicht sogar böswilliger Analyst kann verschiedene Informationen über Personen kombinieren. Indem er verschiedene Datensätze wie Filter nutzt, kann er wie bei der Rasterfahndung Individuen identifizieren und aussagekräftige Profile von ihnen anlegen.

Paul Francis verweist anhand der Beispiele auf einen Zielkonflikt beim Umgang mit den Daten: Für Analysten sind Daten umso interessanter, je mehr sie über das Individuum aussagen. Für Werbeleute zum Beispiel ist nicht nur Vermittler im Datenkonflikt: Paul Francis (rechts) und Sebastian Probst Eide entwickeln Konzepte, um Unternehmen aussagekräftige statistische Informationen zu liefern und dabei persönliche Daten vor Missbrauch zu schützen.

das Geschlecht einer Person wichtig, sondern auch Fragen wie: Lebt sie in einem Double-income-no-kids-Haushalt? Gehört sie einer bestimmten Szene an? In welcher Wohngegend lebt sie?

Doch die Präzision kostet etwas: Das Risiko steigt, dass sich ein Leck in der Privatsphäre auftut. Um die Privatsphäre zu schützen, sollten Daten also möglichst wenig über eine Einzelperson verraten. "Je besser jedoch die Privatheit geschützt ist, desto weniger nützlich sind die Daten", erklärt der Forscher.

Paul Francis möchte das gebrochene Versprechen der Nutzeranonymisierung wiederherstellen und den Unternehmen gleichzeitig aussagekräftige Daten zur Verfügung stellen. Doch er sagt auch: "Das Problem lässt sich nicht wirklich lösen, man kann es nur Schritt für Schritt entschärfen." Zwischen Analysten und Datenschützern finde ein ähnlicher Wettlauf statt wie zwischen Programmierern von Computerviren und Virenschützern. Letztere hängen immer einen Schritt hinterher. Wie Virenschützer müssen auch Datenschützer die konkreten Tricks der Gegenseite analysieren, um dafür konkrete Gegenmaßnahmen zu finden.

Um praxistaugliche Mittel zu entwickeln, die zwischen den gegenläufigen Interessen von Datenschutz und Datennutzung ausgleichen, verfolgt Paul Francis einen Ansatz, der sich von den Lösungsvorschlägen vieler Informatiker grundlegend unterscheidet.

Bisher behandelten Informatiker das Problem allein als ein informationstechnisches oder informationstheoretisches. "Auf diese Weise vernachlässigt man aber viele Aspekte des Problems, das neben technischen auch rechtliche, wirtschaftliche und psychologische Seiten hat", kritisiert Francis. "Es wurden Hunderte von akademischen Publikationen geschrieben, aber kaum eine dieser Lösungen wird in der Praxis benutzt", sagt er. Der Grund: Die Industrie akzeptiere keine Lösung, die eine Datenanalyse wesentlich teurer oder weniger präzise mache.





Auch Francis hat bis vor wenigen Jahren auf rein akademische Weise Techniken entwickelt, die Lecks in der Privatsphäre stopfen sollten. "Dann hatte ich das Gefühl, dass die Technik reif genug war, um ein Start-up zu gründen", sagt er. Das Unternehmen sollte den Praxistest für die Forschungserkenntnisse bringen. Inzwischen gibt es Aircloak seit eineinhalb Jahren. Neben Francis besteht das Team aus fünf jungen Computerspezialisten, die alle praktische Erfahrung mitbringen, etwa aus dem Team von Google+ oder aus dem Kampf gegen Malware und Hacker.

Aircloak will eine Privatsphäre ohne Leck schaffen und dabei alle Aspekte des Datenschutzes berücksichtigen, neben den technischen also auch die juristischen, ökonomischen und psychologischen.

Francis spricht daher mit Datenschutzexperten ebenso wie mit Unternehmern. Erfahrungen darüber, wie Firmen ticken, hat er als Forschungsleiter bei zwei Start-ups im Silicon Valley gesammelt. So kann der Forscher etwa die Sorgen einer Firma verstehen, die Finanzsoftware für die PCs und Mobilgeräte von Verbrauchern herstellt und wissen möchte, warum die Software auf Mobilgeräten wenig genutzt wird. Dazu würde sie gern Nutzerdaten sammeln und auswerten. Doch wegen der sensiblen Natur der Daten - sie beinhalten den Aufenthaltsort, die Finanzen und die Einkäufe eines Nutzers - macht sich die Firma Sorgen um technische und rechtliche Probleme sowie um die mögliche Wirkung des Vorhabens auf die Öffentlichkeit. Daher lässt sie diese Form der Marktforschung lieber bleiben.

Diese Sorgen will Aircloak seinen Kunden durch das Cloaked Computing nehmen. Die Erfindung des Unternehmens erklärt Felix Bauer, Forscher des Max-Planck-Instituts für Softwaresysteme und Mitgründer von Aircloak: "Die Daten werden noch auf dem Computer



oder dem Mobilgerät des Nutzers verschlüsselt", sagt der Physiker. "Dann werden sie in unser zentrales System geschickt." Dieses System, eine sogenannte Cloak, was zu Deutsch Umhang oder Deckmantel heißt, ist nach außen hin abgesichert, sodass niemand darauf zugreifen kann. "Nur innerhalb dieses Systems können die Daten entschlüsselt und analysiert werden", erklärt Bauer.

Die Cloak ist mehr als eine Firewall, mit der sich etwa Unternehmen gegen Onlineattacken von außen schützen. "Es ist eine Art Blackbox", erklärt Francis. Es gebe keine Nutzernamen und Passwörter, keinen Weg, von außen einzudringen. Ein Chip garantiere diese Sicherheit ähnlich wie ein Trusted Platform Modul, das an einen bestimmten PC gebunden ist und diesen umfassend gegen äußere Angriffe schützt. Manipulationen sind praktisch unmöglich: "Jede Änderung der Software, die wir vornehmen, muss von einer dritten Partei genehmigt werden."

Wenn ein Unternehmen etwas über seine Nutzer wissen möchte, dann stellt es eine Anfrage an die Cloak, zum Beispiel: Wie viele meiner Nutzer sind weiblich? Die Cloak verarbeitet die Daten entsprechend und schickt die anonymisierten Daten zurück.

Cloaked Computing unterscheide sich vom derzeit üblichen Umgang mit Daten, erklärt Francis. Bislang werden die Daten meist schon in anonymisierter Form in die Datenbank der Unternehmen gelegt, die das Surfverhalten analysieren oder sonstige Daten sammeln. Bei Aircloak hingegen gelangen die verschlüsselten, aber noch nicht anonymisierten Informationen in die Datenbank. So verlieren sie nicht an Qualität für den

Kunden. Wegen der Cloak sind sie dennoch sicher. Die Anfrage des Kunden wird mithilfe der Rohdaten beantwortet, enthält also das Maximum an Information. Erst die Antwort wird anonymisiert und an den Kunden weitergegeben.

Wenn die Datenbank nicht von den Unternehmen betrieben wird, die an den Daten interessiert sind, und wenn sie gleichzeitig durch eine Cloak geschützt ist, könne weniger private Information in unbefugte Hände gelangen als aus Datenbanken der einschlägigen Unternehmen, in denen die Information bereits anonymisiert liegt. Allerdings garantiert auch das Cloaked Computing keine absolute Sicherheit. Denn listige Analysten können durch die geschickte Kombination von Anfragen Informationen über Einzelpersonen herausfinden.

ZUFÄLLIGE SCHWANKUNGEN IN DEN ANTWORTEN

Um das zu verhindern, beobachtet Aircloak die Anfragen von Analysten und sucht nach Hinweisen auf einen solchen Angriff. Angenommen, eine Datenbank enthält Angaben zum Einkommen von Personen, gibt auf Anfragen aber nur das gesamte Einkommen einer ganzen Gruppe von Nutzern oder andere statistische Antworten zur Gehaltsverteilung heraus.

Ein seriöser Analyst stellt vielleicht folgende Anfrage: Gib mir die Altersverteilung der Nutzer, die ein bestimmtes Einkommen haben. Das Ergebnis wäre ein Diagramm, das jeweils die Anzahl der Nutzer mit einem Monatseinkommen von, sagen wir, über 4000 Euro in den Altersgruppen von 20 bis 30, 30 bis 40 und so fort angibt.

Ein unseriöser Analyst will aber das Gehalt einer Person X herausfinden. Vorausgesetzt, der Angreifer kann die Person X anhand von Postleitzahl, Geburtsdatum und Geschlecht identifizieren, so könnte er zuerst nach dem Gesamteinkommen aller Personen mit der gleichen Postleitzahl fragen. Er stellt dann eine zweite Anfrage: Gib mir das Gesamteinkommen aller Personen mit dieser Postleitzahl außer X. Um das Gehalt von X zu ermitteln, muss er dann nur noch beide Antworten voneinander subtrahieren.

Um solch einen Bruch des Datenschutzes zu verhindern, fügen Aircloak und andere einschlägige Unternehmen, die derlei Analysen betreiben, den Antworten eine leichte zufällige Schwankung hinzu. Dann weicht die Differenz der beiden Gesamteinkommen deutlich von der tatsächlichen Differenz ab. Der Angreifer erhält keine wertvolle Information.

Für einen seriösen Analysten, der nach der Altersverteilung in einer Gehaltsklasse fragt, bliebe die Antwort dagegen wertvoll, obwohl die Antwort auf seine Frage leicht von den tatsächlichen Zahlen abweicht. Wenn es statt 203 Personen in einer Altersgruppe 206 oder 202 sind, kann er immer noch erkennen, welche Altersgruppe in einer Gehaltsklasse wie vertreten ist.

Ein Angreifer könnte seine Anfrage durch zusätzliche Kriterien auch auf bestimmte Individuen eingrenzen. Um derlei Trickserei zu erschweren, hat Francis' Team ein einfaches Mittel ersonnen. "Es gibt eine untere Schwelle", erklärt der Informatiker. Die Antwort wird mit der zufälligen Schwankung versehen, und wenn das Ergebnis unterhalb dieser Schwelle liegt, gibt das System keine Antwort, es sagt zum Beispiel: Sorry, der Wert ist zu klein. Das System verwehrt dem Anfragenden somit die Methode der Rasterfahndung, bei der ein Datensatz nach dem Ausschlussverfahren immer weiter eingegrenzt wird.

"Sie mögen einwenden, dass die Idee, eine untere Schwelle für herausgegebene Antworten einzuführen, nicht sonderlich originell ist", gibt der Forscher zu. "Das stimmt. Aber niemand hat diese Idee zuvor analysiert. Sogar diese simple Idee ist zu schwierig, um sie theoretisch zu analysieren. Natürlich löst unser Ansatz das Problem nicht perfekt. Aber es ist etwas, was man tun kann,"

DIE SYSTEMFORSCHUNG ÄHNELT **EINER ENTDECKUNGSFAHRT**

Allerdings werden Hacker jedem Stein, den man ihnen in den Weg wirft, auszuweichen suchen und neue Angriffsmöglichkeiten aushecken. "Um den Einfluss der künstlich hinzugefügten zufälligen Schwankungen zu eliminieren, könnte der Analyst etwa die gleiche Anfrage immer und immer wieder stellen. Der Mittelwert der Zahlen, die er dabei herausbekommt, wird nahe am wahren Wert liegen", sagt Francis. Freilich könne man verhindern, dass dieselbe Anfrage mehrmals gestellt werde.

Doch die Anfragen lassen sich auch unterschiedlich formulieren. Statt der Postleitzahl kann der Analyst beispielsweise die geografischen Koordinaten Länge und Breite verwenden. Das wäre die gleiche Anfrage. Auch zur Abwehr solcher Versuche haben die Forscher kürzlich eine Methode entwickelt, die aus patentrechtlichen Gründen jedoch noch nicht publik werden soll.

Bei aller Praxisnähe stellt die Forschung von Francis' Team Grundlagenforschung der Informatik dar. "Das sehr komplexe System, mit dem wir es zu tun haben, erfordert einen hohen Aufwand an ingenieurwissenschaftlichem Know-how und informeller Analyse." Informatiker sprechen hierbei von Systemforschung. "Wir beziehen ja auch noch Wirtschaft, Politik und Soziologie in unser Denken ein, sodass das System sogar noch komplexer wird als bisher", sagt Francis.

Ein wesentlicher Teil der Systemforschung ähnelt einer abenteuerlichen Entdeckungsfahrt. Das Team von Paul Francis gleicht der Crew eines Schiffes, das durch ein Seegebiet voller Riffe schippert. Kaum hat es am Bug ein Leck geflickt, kracht es am Heck, und sie müssen dort ein neues Loch stopfen. Doch der Kaiserslauterer hat sichtlich Spaß an diesem Wettstreit. Er wird es den Angreifern alles andere als leicht machen. ◀

AUF DEN PUNKT GEBRACHT

- Spezialisierte Unternehmen analysieren das Surfverhalten von Internetnutzern. Andere Firmen verfügen über vielfältige weitere Daten von Personen. Wer die Daten zusammenführt, erhält oft umfassende Profile mit teilweise sehr privaten Details von Einzelpersonen.
- Analysten, die möglichst viel Information aus Daten gewinnen wollen, und Datenschützer befinden sich in einem ständigen Wettlauf.
- Das Start-up-Unternehmen Aircloak will, basierend auf Erkenntnissen von Max-Planck-Forschern, eine Privatsphäre ohne Leck schaffen und berücksichtigt dabei die technischen, juristischen, ökonomischen und psychologischen Aspekte des Datenschutzes.

GLOSSAR

Cloak: Dieser Schutzmantel sichert nicht-anonymisierte Daten hermetisch vor unerlaubten Zugriffen von außen.

Cloaked Computing: Hinter der Cloak werden Daten in nicht-anonymisierter Form analysiert, um statistische Fragen mit größtmöglichem Informationsgehalt zu beantworten. Das Ergebnis wird anonymisiert und an den Fragenden geschickt.

Tracker: Mit dieser Software verfolgen einschlägige Unternehmen das Surfverhalten von Internetnutzern. Ein Tracker registriert, welche Webseiten von einem bestimmten Computer besucht werden.

<u>.....</u>