The **Data Cloak**

Data is the raw material of the modern information society. All too often, however, companies that require comprehensive data analyses risk breaching data protection guidelines. **Paul Francis**, Director at the **Max Planck Institute for Software Systems** in Kaiserslautern, seeks to strike a balance between these conflicting interests. His company, Aircloak, plays an important role in this endeavor.

TEXT **CHRISTIAN J. MEIER**

Taking a behind-the-scenes look at the Internet is a real eye-opener. A free program called Ghostery shows me who is tracking my online browsing behavior. Paul Francis at the Max Planck Institute for Software Systems in Kaiserslautern recommended it to me while we were sitting in a café close to the institute a few days earlier.

In his plaid shirt and his faded T-shirt depicting the iconic little traffic light man from Eastern Germany, Francis looks like an aging computer kid from Silicon Valley. It therefore comes as no surprise that the scientist not only conducts research, but also runs a start-up in Kaiserslautern. Both his research and his company are dedicated to more effectively protecting the privacy of Internet users.

Paul Francis considers his start-up, named Aircloak, to be a research tool. He sees his company's commercial success as an instrument to gauge the progress of his research. What Francis is doing with his start-up company is similar to going on an expedition into the real world of the Internet. And that world is a jungle where hundreds of companies are busy collecting data on web users. These service providers specialize in tracing the "paths" Internet users take when they are online. This data is then sold to other companies, which in turn use the information, for example, to optimize their advertising strategy.

Back at my desk, I could soon see what Francis was talking about: Ghostery detected six "trackers" after I clicked on an online article from a news magazine. After visiting a few more websites, such as an online flight comparison portal and Facebook, I could now identify about twenty different trackers.

THE FALSE PROMISE OF USER ANONYMITY

The trackers provide the data collectors with information about who visits which website. While the user is identified by means of a number, that number never changes, meaning that it's possible to track which websites the user who was assigned the number X has been visiting. "The companies create a dataset for each visit," explains Francis. Put together, these datasets form a database that allows analysts to study the browsing habits of user X. This information can then be used to display targeted online advertising that most closely matches X's personal interests.

"It's unbelievable," says the computer scientist from the US, shaking his head before going on to explain how targeted advertising works. "Let's assume you're visiting a website that has space for advertising, and several companies want you to see their ads," he says. "All of these companies then make Google an offer. The highest bidder wins."

You may wonder where the problem is. After all, the data is anonymized. Nobody knows that, say, Paul Francis or Christian J. Meier visited this or that particular website. It's house number one or house number two. The user's privacy remains intact.

But it isn't that simple, Francis points out. He calls it the false promise of user anonymity, namely the belief that once the data is anonymized, nobody can find out anything about a particular individual.

Adding to the controversy is the fact that, apart from the companies that know the browsing habits of person X,

Gaps in the privacy sphere: By cleverly combining data obtained from different sources, it becomes possible to create a comprehensive personal profile of an individual, such as the one shown in our made-up example. This is precisely what the researchers at the Max Planck Institute for Software Systems seek to prevent.

Photo: Caro

NAME: MAX MUSTERMANN
AGE: 58
PLACE OF RESIDENCE:
OBERPFAFFENHOFEN

OCCUPATION:
HEAD OF DEVELOPMENT OF A
MEDIUM-SIZED COMPANY

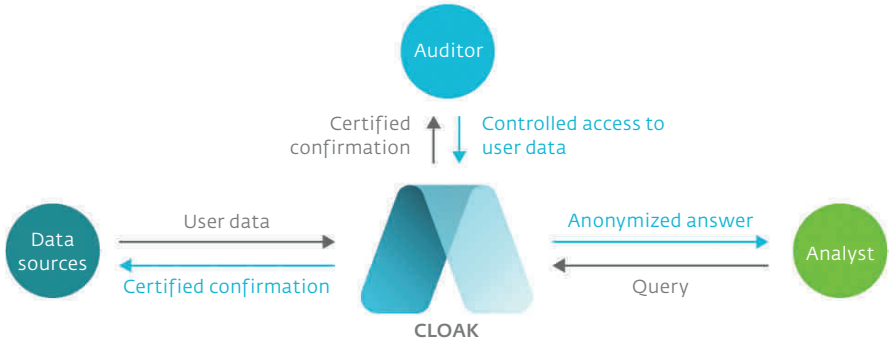
ANNUAL SALARY: 110,000 EUROS

MARITAL STATUS:
MARRIED, THREE CHILDREN

HOBBIES:
PARAGLIDING, COLLECTING BEER COASTERS

MOST COMMON ONLINE QUERIES:
PARAGLIDING TAKE-OFF RAMP, MIGRAINE, SEX
TIPS, TRADING BEER COASTERS, BREWERY,
LINGERIE, VARICOSE VEINS

HEALTH ISSUES:
MIGRAINES, VARICOSE VEINS



The process of anonymizing data: The cloak guarantees that analysts can't filter out individuals from data sets. Beneath the impenetrable cloak that protects the information, user data from one or more sources is managed. Before the data reaches the cloak, the data source receives a certified confirmation verifying that the information will be sent to the actual cloak, and not to another location purporting to be the cloak. The analyst who made the query will then receive anonymized answers that are composed underneath the cloak. During this process, Aircloak (the company that operates the cloak) blocks any queries aimed at obtaining information about particular individuals. The company's auditors have exclusive control over who is granted access to the data and who isn't.

there are other companies out there that possess further information about that same individual: the bank knows their financial situation, the utility company knows how much electricity they consume, the credit card company has insight into X's consumer behavior, and the mobile phone provider knows where X has been and when. "Companies often sell data pertaining to their customers," Francis explains. He has heard of cases in the US in which banks passed on anonymized customer data to other organizations. This means it is possible in principle for a buyer to piece all of this information together to obtain a comprehensive picture of X. The user becomes transparent. "The data is collected for harmless purposes, but depending on who buys that information, things can take a very serious turn," warns Francis.

Despite such scenarios, the computer scientist is no staunch data protection guerilla. He defends the concept of analyzing anonymized user data as such. He believes it can be very helpful, and gives an example: "In the field of medicine, fraud leads to financial losses worth billions of euros," says Francis. Medical databases could be used to expose fraud cases, for example by examining prescriptions. Are there any doctors that write an above-average number of prescriptions? Or are there any that are perhaps prescribing medication they shouldn't be prescribing?

Yet anonymization itself wouldn't help protect the privacy of the majority of doctors who have a clean record. "It's difficult to cram all of the medical data into one large database without jeopardizing people's privacy," says Francis. That is why this potential isn't being used.

Two spectacular cases that have come to light in the past confirm what Francis is talking about. They show that even institutions you would believe to be well-versed in matters of data protection can easily be outwitted.

COMBINED DATA IDENTIFIES INDIVIDUALS

In the late 1990s, a government agency in the US state of Massachusetts that was responsible for managing the health insurance policies of government employees published data about the policyholders in order to make it available to researchers. The agency believed it was protecting the privacy of the government employees by removing the name, social security number and other "unique identifiers" of each person listed in the data set. Even the governor of Massachusetts at the time, William Weld, assured the public that this would protect the privacy of the policyholders.

He hadn't reckoned with Latanya Sweeney, a bright computer science student. For the sum of twenty dollars, she bought the electoral roll of Weld's hometown of Cambridge, near Boston. The register contained the name, address, zip code, date of birth and gender of every single voter. This made it easy for her to find the governor among the list of insurance policyholders published by the agency: only six Cambridge citizens listed in the anonymized health insurance data shared the same birthday, three of whom were men, and only one of whom lived in the same zip code – the governor himself. As a publicity stunt, Sweeney sent the governor his file, including the medical diagnoses and prescriptions it contained.

Several years later, in 2006, the online services company AOL published two million search requests placed by 650,000 users. Researchers rejoiced at the opportunity to examine the Internet behavior of a large number of users by analyzing such a huge amount of data. AOL anonymized the data: the company removed user names, IP addresses (which are allocated to each computer), and other information that would make it possible to directly identify individual users. However, every single user was allocated a unique number to ensure that the data remained valuable for the researchers.

This time, it was two journalists from the New York Times who showed AOL that this form of anonymization didn't adequately protect people's privacy. The queries of user 4417749 contained clues as to her identity. After all, there aren't that many users who would search for both a landscape gardener in "Lilburn, GA" and a house for sale in "Shadow Lake, Georgia". The journalists identified this particular user as Thelma Arnold. Arnold confirmed that she had entered these search terms into her web browser, including embarrassing queries such as "dog that urinates on everything".

The moral of the story: a cunning, perhaps even malicious analyst could piece together different pieces of information about a person. By using different data sets as filters, the analyst could apply a method similar to a grid search to identify individuals and create comprehensive profiles on each of them.

Paul Francis uses these examples to emphasize the conflict of objectives that exists with regard to how data is managed: the more a set of data reveals about an individual, the more interest-

Graphic: Aircloak

Mediating the data conflict: Paul Francis (right) and Sebastian Probst Eide develop concepts to provide companies with informative statistics while at the same time preventing the misuse of personal data.

ing it becomes for analysts. Advertising experts, for example, are interested not only in a person's gender, but also in numerous other questions: Does this person live in a double-income-no-kids household? Does he or she prefer a particular social scene? In which area does he or she live?

Yet such precision has its price: there is an increased risk of private data leaking out. In order to keep a person's privacy safe, the data should reveal as little as possible about that individual. "But the higher the security level of a person's privacy, the less useful the data becomes," the researcher explains.

Paul Francis wants to fix the broken promise of user anonymization while at the same also making informative data available to companies. However, he concedes that "the problem can't be solved completely, only defused step by step." The battle being waged between analysts and data privacy advocates is similar to that between programmers who develop computer viruses and those who try to protect systems by coding anti-virus software. The latter are always one step behind. Just like programmers seeking to ward off viruses, data protection groups must analyze their opponents' modus operandi to come up with effective countermeasures.

In order to develop practical resources for balancing out the conflict of interests between data privacy and data use, Paul Francis has adopted an approach that is fundamentally different from the solutions many IT experts propose.

To date, computer scientists have analyzed the situation solely from the point of view of information technology or information theory. "But that way, you ignore many other aspects of the problem, which has not just technical, but also legal, economic and psychological facets," says Francis, criticizing the approach. "Hundreds of academic papers have been published, yet hardly any of these solutions are being implemented in practice," he says. The

Photo: Thomas Hartmann

Putting data protection to the test: Felix Bauer presents the Aircloak concept at the CeBIT trade fair in Hanover. One of the aims the researchers at the Max Planck Institute are pursuing is to make this technology attractive for companies wishing to use information about their customers.

reason: industry won't accept a solution that makes data analysis significantly more expensive or less precise.

Up until a few years ago, Francis, too, tried patching the privacy leaks by purely academic means. "Then I felt that the technology had since become sophisticated enough to create a start-up," he says. The company was designed to test the research findings. Aircloak has now been around for a year and a half. In addition to Francis himself, the team consists of five young computer specialists, all of whom have experience in the field, for example having worked for Google+ or battled malware and hackers.

Aircloak's objective is to create a privacy sphere without leaks, taking into account all aspects relevant to data protection: technical, legal, economic and psychological.

For that reason, Francis talks to both data privacy experts and entrepreneurs. As the research manager at two start-ups in Silicon Valley, he learned what makes companies tick. Thus, he can understand, for example, the concerns of a company that develops financial software for consumer PCs and mobile devices and wants to know why the software is hardly being used on mobile devices. It therefore wants to collect and evaluate user data. Yet due to the sensitive nature of this data, which often contains a user's current location, financial situation and purchase history, the company is concerned about technical and legal issues, as well as about the public's reaction to the intended analysis. As a result, the company decides not to pursue this line of market research.

Aircloak seeks to assuage its customers' concerns by means of cloaked computing. Felix Bauer, researcher at the Max Planck Institute for Software Systems and co-founder of Aircloak, explains how the invention works: "The data is encrypted while still on the user's computer or mobile device," says the physicist. "Then it is sent to our central system." This system, known as a cloak, is shielded against the outside, preventing any form of unauthorized access. "The data can be decrypted and analyzed only within the system," Bauer explains.

The cloak is more than the kind of firewall companies or individuals use to protect themselves against online attacks. "It's sort of like a black box," explains Francis. It doesn't contain any user names or passwords, and there is no way to access it from outside. This level of security is guaranteed by a chip, similar to a trusted platform module that is bound to a particular PC, protecting it against any type of outside attack. Manipulating the system is practically impossible: "Any changes we make to the software must be authorized by a third party."

When a company wants to know something about its users, it requests information from the cloak. Example: How many of my users are female? The cloak then processes the data accordingly and sends the anonymized data back to the company.

Cloaked computing manages data in a different way than conventional methods, explains Francis. Up until now, data is already anonymized in most cases before being stored in the database of the company that is analyzing the browsing behavior or collecting other data. In the case of Aircloak, however, the information sent to the database has already been encrypted, but not yet anonymized. This is done to ensure that the data retains the quality the customer needs. And thanks to the cloak, it remains secure. The customer's request for information is answered using raw data, meaning it contains the maximum level of information. The first element to be anonymized and passed on to the customer is the answer itself.

When the database isn't operated by the companies that are interested in the data, and when it is protected by a cloak, the likelihood of personal information falling into the wrong hands is much lower than if the data were stored in the companies' own databases and had already been anonymized. However, even cloaked computing can't guarantee absolute security, due to the fact that cunning analysts can find out information about individuals by cleverly combining data requests.

RANDOM FLUCTUATIONS ARE ADDED TO THE ANSWERS

In order to thwart such attempts, Aircloak monitors the requests placed by analysts and searches for any signs of an attack. Let's assume that a database stores information about the income of individuals, but responds to a request by showing only the total income of a whole group of users, or other statistics based on income distribution.

A reputable analyst might make a request like this: Show me the age distribution of users in a particular income bracket. The result would be presented

Photo: CeBIT

as a diagram depicting the number of users with a monthly income of, say, 4,000 euros in the age group from 20 to 30, 30 to 40, and so on.

An analyst with ulterior motives, however, wants to find out how much person X earns. Provided the attacker can identify person X by means of their zip code, date of birth and gender, they would proceed by first requesting the total income of all persons living in that same zip code, apart from X. In order to determine the income of X, all they need to do now is subtract one answer from the other.

In order to prevent such a breach of data privacy, Aircloak and other companies that conduct such analyses add a minor random fluctuation to the answer so that the difference between the two overall incomes deviates significantly from the actual difference. As a result, the attacker doesn't obtain any valuable information.

For the reputable analyst seeking information on the age distribution of a particular income bracket, the answer remains valuable despite the fact that the answer to the question deviates slightly from the actual figures. Even if the data analysis comes up with 206 or 202 individuals in a particular age group instead of 203, it is still clear which age group is represented in an income bracket.

An attacker could narrow down a query to obtain information about particular individuals by adding further criteria. Francis' team has devised a simple method to impede such trickery. "There is a bottom threshold," the computer scientist explains. A random fluctuation is added to the answer, and if the result falls below this threshold, the system won't provide an answer. Instead, it will say: "Sorry, this value is too low," for example. The system thus denies the inquirer the possibility of using the grid search method, in which the data set is narrowed down over and over.

"You could argue that the idea of introducing a bottom threshold for the answers provided isn't exactly original," the researcher admits. "That's true. But no one has analyzed this idea before now. Even this simple idea is too complex to be analyzed using a theoretical method. Of course our approach isn't the perfect solution to the problem, but at least it's a step in the right direction."

SYSTEMS RESEARCH IS LIKE GOING ON AN EXPLORATORY MISSION

Hackers, however, will try to find a way around every obstacle put in their path, and hatch new plans of attack. "In order to eliminate the influence of the artificially added random fluctuations, an analyst could, for example, make the same query over and over again. The resulting average would then closely approximate the actual value," says Francis. Of course you could prevent someone from repeatedly posing the same query.

Yet a query can simply be formulated in different ways. Instead of the zip code, for example, the analyst could use geographical coordinates. The two queries would be identical. The researchers recently developed a method to counter such attempts as well, but aren't making it public yet due to a pending patent.

Despite the fact that the solutions developed by Francis and his team are designed to be implemented in real life, their work also constitutes basic research in the field of information science. "This highly complex system that we are dealing with requires a substantial amount of engineering know-how and informal analysis." Information scientists call it systems research. "After all, we also include economic, political and sociological factors in our thought processes so that the system will become even more complex than it used to be," says Francis.

Conducting systems research is essentially like embarking on a mission to explore new territory. The team headed by Paul Francis is like the crew of a ship, sailing through rough waters full of hidden reefs. As soon as the sailors fix a leak in the bow, the stern of the ship crashes into another reef and they have to run over to plug a new hole. But the researcher from Kaiserslautern clearly enjoys this race. He certainly won't make it easy for the attackers. ◀

TO THE POINT

- **Certain companies specialize in analyzing the online browsing habits of Internet users. Other companies possess a wide range of other data on people. By combining all of this information, it is possible to create comprehensive profiles of individuals, sometimes even containing very personal data.**
- **A seemingly endless battle is being waged between analysts on the one hand, who seek to glean as much information as possible from data, and data privacy activists on the other.**
- **Based on the findings of the researchers at the Max Planck Institute, the start-up company Aircloak aims to create a privacy sphere without leaks, taking into account the technical, legal, economic and psychological aspects of data protection.**

GLOSSARY

Cloak: Hermetically seals off non-anonymized data and prevents unauthorized access from outside.

Cloaked computing: The data "hidden" beneath the cloak remains non-anonymized and is analyzed so that statistical queries can be answered using as much information content as possible. The result is then anonymized and sent to the person who made the query.

Tracker: Software certain companies use to track the online browsing habits of Internet users. Trackers register which websites were accessed on a particular computer.