Aircloak Analytics: Anonymized User Data without Data Loss

An Aircloak White Paper

Companies need to protect the user data they store for business analytics. Traditional data protection, however, is costly and often fails. The alternative – user data anonymization – is hard to get right, and requires throwing away valuable user data. Aircloak's cloaked computing technology provides strong anonymization without user data loss. With Aircloak, companies can now enjoy the benefits of business analytics over complete user data without risk of user data leaks.

Data Protection and the Broken Promise of Data Anonymization

User data drives business analytics. Collecting and sharing user data, however, can be a legal and public relations nightmare. Protecting user data from leaks while allowing access to those who need it for business intelligence is difficult, expensive, and error prone. Obtaining data-protection or data-privacy certifications can alleviate user concerns over privacy, but is a costly, lengthy, and never-ending process. Accidental data leakage can require user notification, and leads to a loss of user confidence and possible legal action.

Data anonymization was supposed to solve these problems ... anonymization has turned out to be much harder than originally thought.

Data anonymization was supposed to solve these problems. The promise of anonymization is that, once user data is anonymized, nobody with access to that data can learn anything about an individual user. Once anonymized, user data is safe to share. Unfortunately, anonymization has turned out to be much harder than initially thought. As several companies have discovered the hard way, it is not simply a matter of removing Personally Identifying Information (PII)[i]. In what is called the "curse of high dimensionality", privacy researchers have discovered, in essence, that the more user data one has, the more user data one must throw away to anonymize it[ii].

Anonymity through Cloaked Computing

Aircloak takes a different approach to anonymization. Based on research done at the Max Planck Institute for Software Systems in Germany, Aircloak restores the promise of anonymization. The basic idea is simple: instead of anonymizing the data, Aircloak keeps the data and anonymizes the answers, a much easier task. In order to fully protect the data, Aircloak exploits cryptographic hardware to prevent any direct access whatsoever to that data.

Aircloak's approach, called "Cloaked Computing", leverages Trusted Platform Module (TPM) hardware standardized by the Trusted Computing Group[iii] (TCG). The data protection offered by Cloaked Computing is so strong that absolutely nobody, not even system administrators, has access to raw user data. As a result, none of the typical problems that undermine data protection – leaked or weak passwords, incorrect firewall settings, misconfigured access control tables, or overly-specific queries – can leak data under cloaked computing. From the hardware up, Aircloak analytics is immune to accidental data leakage as well as exposure to corporate "peepers".

A New Standard for Trust and Transparency

Anybody can examine our code ... Over time, this results in extremely secure software.

Aircloak doesn't simply *claim* to run cloaked computers. Aircloak *proves* that it runs cloaked computers through a TPM feature called attestation. This provides cryptographic proof that the software we claim to run is what we actually run. Anybody can examine our software and validate that it prevents leakage of user data. Our software is also examined and given a privacy seal by a privacy certification firm. Because a certification company cannot, however, discover all software bugs that may lead to security violations, Aircloak takes an additional step: we publicly publish our code, and offer a bounty to anyone who can discover security flaws. Over time, this results in extremely secure software.

Business Case 1: Smart-Meter Power Company

A large electric utility has recently begun offering smart meters for homes and small businesses. They use smart meters from a variety of vendors, each of which individually gathers detailed energy usage data. The company also has its own user data, including billing records, and also purchases data from other sources, for instance from credit rating agencies. They would like to gather all this data together for business intelligence, but because of heightened consumer sensitivity to smart-meter privacy issues, the smart-meter vendors are unwilling to release their data to the company. Furthermore, the company is concerned about negative publicity regarding gathering data from other sources. The solution is to have each data source directly input its data to Aircloak. This frees them from the cost and liability of handling all the user data, as well as satisfies the concerns of the smart meter vendors.

Business Case 2: Consumer Financial Software

A software company makes a consumer financial software product designed to run on PCs and mobile devices. Each user's financial data is stored on his or her own PC, where users can manage their finances and prepare tax forms. Users can also input purchase information on their mobile devices, which syncs with the PC. Pickup of the mobile feature has been sporadic, and the software company would like to gather user data to understand why. Because of the sensitive nature of the data (location, finances, purchases), they are concerned about a range of legal, technical, and public relations issues. They considered getting a privacy certification, but besides being very costly, it would take well over a year to complete. Aircloak provides an immediate turnkey solution, allowing the company to leverage both Aircloak's certification as well as the Aircloak brand.

Business Case 3: Health Data Analytics

An organization operates a large regional Health Information Exchange (HIE). They want to make their massive health records database available for research analysis purposes, both inside and outside their member organizations. NIH standards for de-identification remove valuable location and timing data from the database. Furthermore, the organization is concerned that even with this de-identification, users may be identified by analysts. By using Aircloak, the organization's database maintains its complete value, while at the same time providing much stronger anonymization than is possible through de-identification alone.

System Architecture

The Aircloak Analytics system consists of the following components:

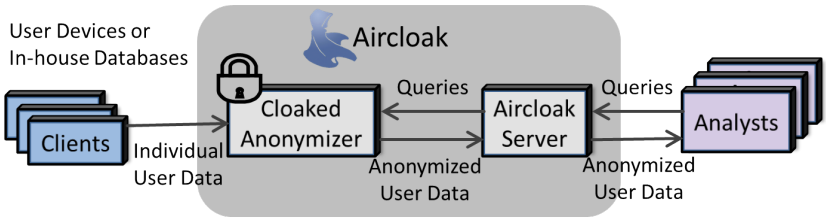

Figure 1: Aircloak Analytics System

- *Clients* that contain user data and operate Aircloak client software. Clients can be user devices or an existing corporate database.
- *Analysts* that make queries over user data. The analyst may or may not be the company that generated the user data.
- A *Cloaked Anonymizer* that receives individual user data, protects it from any exposure, and generates anonymized answers in response to queries.
- An *Aircloak Server* that receives queries from analysts, stores anonymized answers, and makes the answers available to analysts.

Cloaked Anonymizer

The key to the Aircloak Analytics system is the *Cloaked Anonymizer* (Figure 2). The Cloaked Anonymizer allows us to store and process individual user data while keeping it anonymous. The Cloaked Anonymizer has four primary data protection and anonymization features:

- User data sandboxing in a hardened OS
- Answer anonymization
- TPM-sealed user data
- TPM-based remote attestation

Hardened OS: The Cloaked Anonymizer runs within a Linux OS that is hardened against external access, including access by Aircloak. For instance, login is disabled, memory cannot be read externally, and software cannot be installed or updated without a memory-wiping system reboot. The system operates with an extremely limited set of inputs and outputs and, once booted, operates outside the control of any operator, including Aircloak.

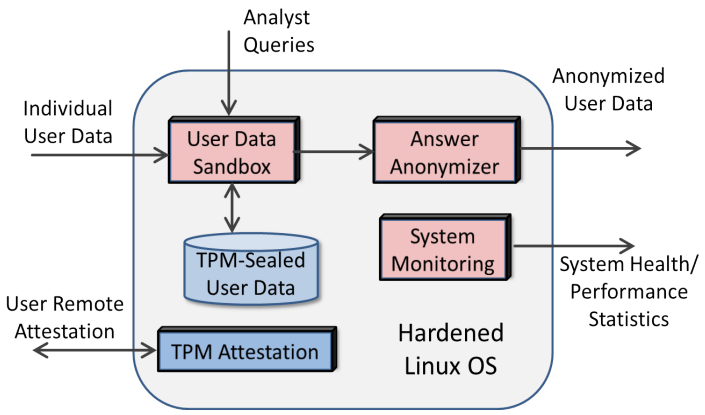

Figure 2: Cloaked Anonymizer Design

Individual user data may enter the system, but is sandboxed such that it can only leave the system via an anonymization function. User data is stored only encrypted on TPM-sealed disks. Either way, individual user data is protected from exposure.

The Cloaked Anonymizer prevents accidentally disclosing information about a single user

Answer Anonymization: The Answer Anonymization function utilizes a unique combination of K-anonymization, random noise addition, and repeat-answer suppression. The K-anonymization prevents answers pertaining to small number of users or user attributes from being released. An analyst can learn that the answer to a query is small, but not how small. The random noise addition adds uncertainty to answers, in particular overcoming some of the weaknesses of K-anonymity (for instance, where the analyst knows that there are either K or K+1 users in the answer, and would otherwise use the existence or non-existence of an answer to determine which). Repeat-answer suppression makes it very hard for an analyst to use repeated queries to cancel out the random noise added to each answer.

Even with these protections, there remains a theoretical though remote possibility that a determined analyst could break anonymization for a very small number of users. To prevent this, Aircloak monitors analyst queries for behaviors that might indicate such an attack.

TPM-sealed User Data: Sealing is a TCG-standard method of encrypting and storing data on disk in such a way that only certain software configurations can decrypt and read the data. The TPM chip will refuse to decrypt the sealed data unless specific software is running on the computer. In the case of the Cloaked Anonymizer, the TPM chip will decrypt sealed data only if the software stack exactly matches the one that sealed the data in the first place. In other words, *any changes* to the Cloaked Anonymizer software will render stored user data as unreadable.

Because the Cloaked Anonymizer software prevents individual user data from leaving the system without being anonymized, and because the TPM prevents any software except the same Cloaked Anonymizer software from reading encrypted user data from disk, individual user data stays protected.

TPM-based Remote Attestation: Of course analysts and their users require assurance that Aircloak is indeed running a Cloaked Anonymizer as specified. We provide this assurance through another feature of the TPM chip called *software attestation.* Through software attestation, the TPM chip securely verifies (or "attests") what software is running. TCG standards allow for *remote attestation* of software. This means that one computer can verify what software is running on another computer over the network. Aircloak uses remote attestation to allow clients to attest the software running on the Cloaked Anonymizer.

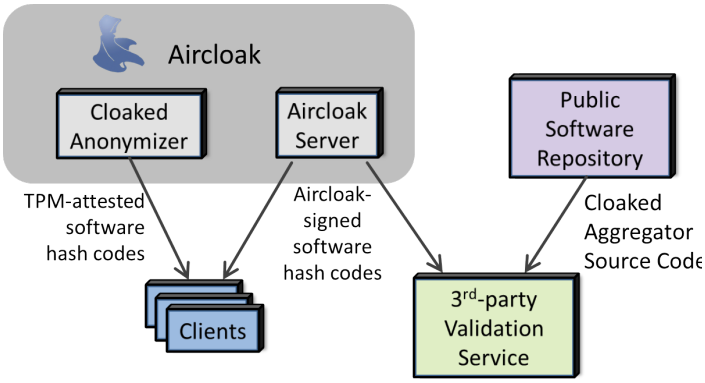

Figure 3: Open and Attested System

Open, Attested, and Crowd-source Validated

Figure 3 illustrates the overall procedure for Aircloak's open and attested system operation. Aircloak openly publishes all of the source code running on the Cloaked Anonymizer in a public repository. Aircloak also publishes the software binaries, and the information needed by a third party to generate the same binaries from the source code. Separately, Aircloak publishes the TPM-generated attestation codes for the software binaries. These attestation codes

are the hash values of the software binaries. Finally, Aircloak signs the published attestation codes.

With this in place, any interested third party, such as a government agency, industry watchdog group like the EFF, or an academic organization, is able to:

1. Validate that the published and signed attestation codes do indeed represent binaries compiled from the Cloaked Anonymizer source code, and
2. Validate that the Cloaked Anonymizer source code does indeed protect individual user data from exposure.

As described above, Aircloak clients do TPM-based remote attestation on the Cloaked Anonymizer. Specifically what this attestation tells the client is that "*some valid TPM chip* attests that *the software binaries* with *the reported attestation hash codes* are running on the remote machine". Separately, clients download the signed attestation software hash codes from Aircloak. If the signed hash codes match those from the TPM, the user is assured that Aircloak has published the source code that the Cloaked Anonymizer is running.

Build Trust, Reduce Costs and Risks, Get Great Analytics

Our TPM-based architecture eliminates the need for many costly data-protection tasks ...

Aircloak's user data analytics service is far more cost effective than traditional approaches, for several reasons. Our TPM-based architecture eliminates the need for many costly data-protection tasks such as training and monitoring system administrators, maintaining access control lists, and securing data backups, and compiling massive documentation for certification by third parties. And of course we provide the usual benefits of outsourcing highly specialized services—amortization of highly skilled professionals, equipment, operational costs, and legal costs.

With its protected access to raw user data, Aircloak promises high-quality user intelligence. At the same time, through its unprecedented transparency and superior data protection, Aircloak creates confidence among users that their data is fully secure. By using Aircloak Analytics, companies demonstrate that they are serious about user privacy and can ultimately achieve something invaluable: user trust.

[i] America Online, the state of Massachusetts, and Netflix: see section B.1 of Paul Ohm, endnote ii below.
[ii] Paul Ohm, *"Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization"*, 57 UCLA L. Rev. 1701, http://www.uclalawreview.org/?p=1353
[iii] www.trustedcomputinggroup.org

We are sponsored by:

EUROPÄISCHE UNION

Gefördert durch:
Bundesministerium für Wirtschaft und Technologie
aufgrund eines Beschlusses des Deutschen Bundestages

ESF Europäischer Sozialfonds für Deutschland

eXIST Existenzgründungen aus der Wissenschaft

aircloak

Max Planck Insitute for Software Systems
Paul-Ehrlich-Str. 26
67663 Kaiserslautern
Germany

Phone: +49 631 9303 9600
solutions@aircloak.com