Clean Insights
Privacy-Preserving Measurement

THE GUARDIAN PROJECT
https://guardianproject.info
Guardian Project creates easy to use secure apps, open-source software libraries, and customized mobile devices that can be used around the world by any person looking to protect their communications and personal data from unjust intrusion, interception and monitoring.
Success?

Not sure :(  

We have millions of active users, but no acceptable way to measure their satisfaction or our impact.
Decision makers, developers, and data scientists need to understand their products’ effectiveness and their users’ happiness.

This must not come at the cost of privacy, security and trust.
Doing it Wrong
All your permissions and data belong to us!

Meitu details why it needs all those permissions.

Chinese photo editing app Meitu made landfall in the U.S. recently, with the free app shooting up the Play Store rankings over the course of the week. The app adds anime-style filters to photos, and the final results end up being equal parts wonderful and weird.

Meitu also went into detail over the permissions it requires:

- **MAC address/IMEI number**: In some cases, Meitu cannot get both info at the same time and in some cases different devices even have the same IMEI number, so we combine these two details into one unique ID to track user devices.
- **LAN IP address** is used to prevent business fraud.
- **SIM card country code** is used for a rough location detection.
- **GPS and network location** are used for detecting countries and regions for Geo-based operation and advertisement placement.
- **Phone carrier info** is used as a standard tracking channel for analytics, just like the other third-party analytics tools (e.g., Flurry).
- **RUN_AT_START**: because the Google service (including GCM) is not available in mainland China, Meitu uses a third-party push notification service called Getui (www.getui.com).

That's certainly a lot to put up with for a photo filter app. If you're satisfied with Meitu's explanation, the app is available for free from the Play Store.
Weaponized users though insecure analytics

In the attack on GitHub and GreatFire.org, the GC intercepted traffic sent to Baidu infrastructure servers that host commonly used analytics, social, or advertising scripts.
REPORT: TESLA'S FATAL CRASH CAN'T BE BLAMED ON SOFTWARE ERRORS

“Blackbox” exonerates corporation
Tesla publishes Model S driving logs that show The New York Times’ blatant lies

By Sebastian Anthony on February 14, 2013 at 8:17 am | 143 Comments

0 shares facebook twitter googleplus stumbleupon

Following Elon Musk’s initial denouncement of The New York Times for publishing a fake review of the Tesla Model S electric car, he has now published the actual logs recorded by the car — and boy are they damning. In short, the NYT’s John Broder lied through his teeth to smear electric vehicles in general, and the Model S in specific.

The basic premise of John Broder’s story for The Times was that the car lied about its self-reported estimated remaining range; when it said there was 79 miles left in the battery, there was in actual fact only 60. Eventually after a few such cases of the car’s battery range being way below what it claimed, the NYT concluded that the car was a chimera.

“A Most Peculiar Test Drive

Elon Musk, Chairman, Product Architect & CEO February 13, 2013

“Blackbox” incriminates the user
Cops use pacemaker data to charge homeowner with arson, insurance fraud

Police called pacemaker data an "excellent investigative tool" that provided "key pieces of evidence" to charge a man with arson and insurance fraud.

Middletown Police said this was the first time it had used data from a heart device to make an arrest, but the pacemaker data proved to be an "excellent investigative tool;" the data from the pacemaker didn’t correspond with Compton’s version of what happened. The retrieved data helped to indict Compton.
Be careful what you read

WARNING

Although the Santa Cruz Library makes every effort to protect your privacy, under the federal USA PATRIOT Act (Public Law 107-56), records of the books and other materials you borrow from this library may be obtained by federal agents. That federal law prohibits library workers from informing you if federal agents have obtained records about you. Questions about this policy should be directed to Attorney General John Ashcroft, Department of Justice, Washington, D.C. 20530.

NEWS ITEM:
THANKS TO CONGRESS, THE USA PATRIOT ACT ALLOWS THE FBI TO TRACK READING HABITS OF DANGEROUS SUSPECTS....

DO YOU MIND?

SORRY MISS, JUST DOING OUR JOB...
Existing Work
Differential privacy (Dwork 2006)

A mechanism $\mathcal{K}$ gives $\varepsilon$-differential privacy if for all values of DB, DB' differing in a single element, and all $S$ in $\text{Range}(\mathcal{K})$

$$\frac{\Pr[ \mathcal{K}(\text{DB}) \text{ in } S]}{\Pr[ \mathcal{K}(\text{DB'}) \text{ in } S]} \leq e^\varepsilon \sim (1+\varepsilon)$$
Apple Thinks Differentially

Learning Popular Emojis with Privacy

Anonymous
Not associated with Apple ID
Randomized identifier
Not linked to other Apple services
Not shared with third parties
You're in control

Aggregate via Differential Privacy

Learn from crowd while protecting individual privacy
Strong mathematical guarantees
iOS and macOS
Learning statistics with privacy, aided by the flip of a coin
October 30, 2014

Cross-posted on the Research Blog and the Chromium Blog

At Google, we are constantly trying to improve the techniques we use to protect our users' security and privacy. One such project, RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response), provides a new state-of-the-art, privacy-preserving way to learn software statistics that we can use to better safeguard our users' security, find bugs, and improve the overall user experience.

Building on the concept of randomized response, RAPPOR enables learning statistics about the behavior of users' software while guaranteeing client privacy. The guarantees of differential privacy, which are widely accepted as being the strongest form of privacy, have almost never been used in practice despite intense research in academia. RAPPOR introduces a practical method to achieve those guarantees.

To understand RAPPOR, consider the following example. Let's say you wanted to count how many of your online friends were dogs, while respecting the maxim that, on the Internet, nobody should know you're a dog. To do this, you could ask each friend to answer the question "Are you a dog?" in the following way. Each friend should flip a
Solutions do exist...

APPLE’S ‘DIFFERENTIAL PRIVACY’ IS ABOUT COLLECTING YOUR DATA—BUT NOT YOUR DATA

...but are not readily available to most
Welcome!
What would you like to know about the Tor network?

Users
Where Tor users are from and how they connect to Tor.

Servers
How many relays and bridges are online and what we know about them.

Traffic
How much traffic the Tor network can handle and how much traffic there is.

Performance
How fast and reliable the Tor network is.

Onion Services
How many onion services there are and how much traffic they pull.

Applications
How many Tor applications, like Tor Browser, have been downloaded or updated.

Users
We estimate the number of users by analyzing the requests induced by Tor clients. These papers detail on how we count users and how we count bridge users.

Relay users  Bridge users by country  Bridge users by transport  Bridge users by country and transport  Bridge users by IP version
Top-10 countries by relay users  Top-10 countries by possible censorship events  Top-10 countries by bridge users

Directly connecting users
This graph shows the estimated number of directly connecting clients, that is, it excludes clients connecting via bridges. These estimates are derived from the number of directory requests counted on directory authorities and mirrors. Relays resolve client IP addresses to country codes, so that graphs are available for most countries. Furthermore, it is possible to display indications of censorship events as obtained from an anonymized censorship-detection system (for more details, see this technical report). For further details see these questions and answers about user statistics.

Tor’s Anonymous Metrics
**Matomo** is leading open-source analytics systems, with 100% data ownership, user privacy protections, and extensibility.

Acra catches exceptions, retrieves lots of context data and send them to the backend of your choice.

Best of all, it is **FREE** and **OPEN SOURCE**.
Expressed Needs
What Developers of Secure Messaging Apps Want to Know

- Is the network latency reduced since last week?
- How many average conversations do you users have open?
- How many people are typically in a group chat?
- Is the battery usage better or worse with the new version?
- Do users like the change in the user interface?
Feedback from our broad user survey

A first step would be to not have identifiable information on well-behaving users.

I feel very uncomfortable, but the mis-use over our system is a problem that poses a threat to the whole project.

We (a funder) are frequently asked for additional metrics demonstrating the impact of the projects we support.

The biggest thing we generally use analytics for is just tracking the health of our service in terms of overall user growth and location.

I would love if we could have something self-hosted, secure and useful that we could use to make our app a better tool but we couldn't find anything that would meet those requirements.

We don't know which features are popular. Therefore, it is hard to expand the service.
Consumers Want Coupons Without Being Geo-Located
We want developers to have a means to understand how to improve, but to do so in a way that respects privacy and security.
Threat Modeling
Vulnerable Assets

- Unique user / hardware identifiers
- Internet addresses
- Biometric data
- Geolocation data
- Social graphs
- Behavioral logs
- Political preferences
- And on, and on....
Attack Vectors

- Client/device temporary cache
- Server storage database
- Developer code libraries
- Network transports
- The Algorithm
Potential Mitigations

- FLOSS code
- Hardened network transports
- Data minimization
- Client-side processing
- Differential Privacy, Randomized Response, Randomized Controlled Trials and more...
The “Clean Insights” Pitch
Most companies treat data like gold...
... but we believe it is a more toxic element!
Who it serves

Developers

Data Scientists  Executives
Knowledge and Privacy!

Clean Insights Process

Toxic Data and Liability
Typical analytics services vacuum up every interaction, transmitting the raw data to a centralized cloud data warehouse, often in an insecure manner.
Clean Insights pushes data processing to the edge, selectively collecting and sharing through secure channels, to a self-hosted backend server.
Three Tenets of Clean Insights

**Hardened Security**
- Certificate Pinning
- TLS Best Practices
- Onion Routing

**Powerful Privacy Toolbox**
- Data Batching
- Smart Thresholds
- No Perma Cookies

**Advanced Anonymity**
- Differential Privacy
- Randomized Response
- Machine Learning
Empowering & Engaging User interactions

FAIL STATE DIAGNOSTIC REQUEST

We noticed something is going wrong.

Mind if we log some diagnostic data to help figure out what is going on?

YES  NO  NEVER!

LOYALTY OPT-IN REQUEST

Clearly, you like using this app. Want to help us make it even better?

YES  NO  NEVER!

MEASUREMENT OPT-IN REQUEST

There have been reports of network issues.

Can we run a quick measurement to check the quality of your connection?

YES  NO  NEVER!

PASSEIVE MEASUREMENT WITH FEEDBACK

App Network Health 70 / 100 % ....

Your Connection 82nd Rank GOLD STAR!

Your Usage ACTIVE

SENSOR-SPECIFIC MEASUREMENT

We noticed that you are in a new place that is not in our service.

Want to share some data so we can help put it on the map?

GLADLY  NOT RIGHT NOW

TIME BOUND MEASUREMENT

How long should we measure for?

AS LONG AS YOU NEED

SHORT AS POSSIBLE

JUST FOR AN HOUR

TODAY ONLY
Onion Routing & Certificate Pinning

```java
if (CleanInsights.getInstance(mPiwik.getContext()).isTorEnabled())
{
    int proxyPort = CleanInsights.getInstance(mPiwik.getContext()).getTorHttpPort();
    mProxy = new Proxy(Proxy.Type.HTTP, new InetSocketAddress("127.0.0.1", proxyPort));
}

StrongOkHttpClientBuilder builder = StrongOkHttpClientBuilder.forMaxSecurity(mPiwik.getContext());
OkHttpClient client = null;

if (mCertPin != null) {
        .add(packet.getTargetURL().getHost(), mCertPin)
        .build();

    client = new OkHttpClient.Builder()
        .proxy(mProxy)
        .certificatePinner(certificatePinner)
        .build();
}```
public boolean allowMeasurement() {
    if (checkLocationPermission()) {
        Location locationNow = locationManager.getLastKnownLocation(LocationManager.PASSIVE_PROVIDER);
        float distanceNow = locationNow.distanceTo(locationNear);
        if (distanceNow <= distanceLimit) {
            return true;
        }
    }
    return false;
}

try {
    Date startDate = SimpleDateFormat.getDateTimeInstance().parse("4/20/2017");
    Date endDate = SimpleDateFormat.getDateTimeInstance().parse("4/21/2017");

    getMeasurer().addThreshold(new SessionLengthThreshold(true, 60));
    getMeasurer().addThreshold(new DateThreshold(false, startDate, endDate));
}
catch (ParseException pe) { }
@Override
protected void onPause() {
    super.onPause();

    // when the app pauses do a private, randomized-response based tracking of the number of likes
    MeasureHelper.track().privateEvent("Vote", "Like per Session", Integer.valueOf(mLikeCount).floatValue(), getMeasurer())
        .with(getMeasurer());

    // dispatch the current set of events to the server
    ((CleanInsightsApplication)application()).getMeasurer().dispatch();
}
private Encoder createRandomizingEncoder() {
    // TODO: Choose appropriate parameters
    return new Encoder(mMeasurer.getUserSecret(),
        ENCODER_ID,
        4096,
        13.0 / 128.0,
        0.25,
        0.75,
        1,
        2);
}

public synchronized MeasureMe set(@NonNull QueryParams key, int value) {
    final String stringValue;
    if (key == QueryParams.EVENT_VALUE) {
        key = QueryParams.EVENT_NAME;
        stringValue = BaseEncoding.base64().encode(createRandomizingEncoder().encodeOrdinal(value));
    } else {
        stringValue = Integer.toString(value);
    }
    set(key, stringValue);
    return this;
}
public Encoder(byte[] userSecret, String encoderId, int numBits,
        double probabilityF, double probabilityP, double probabilityQ,
        int numCohorts, int numBloomHashes) {

    this(
        null, // random
        null, // md5,
        null, // sha256,
        userSecret,
        encoderId,
        numBits,
        probabilityF,
        probabilityP,
        probabilityQ,
        numCohorts,
        numBloomHashes);

/**
 * Constructs a new RAPPOR message encoder.
 *
 * @param userSecret Stable secret randomly selected for this user. UserSecret must be at least
 *        MIN_USER_SECRET_BYTES bytes of high-quality entropy. Changing the user secret clears the
 *        memoized cohort assignments and permanent randomized responses. Be aware that resetting
 *        these memoizations has significant privacy risks -- consult documentation at go/rappor for
 *        more details.
 * @param encoderId Uniquely identifies this encoder. Used to differentiate momoized
 *        cohort assignments and permanent randomized responses.
 * @param numBits The number of bits in the RAPPOR-encoded report.
 * @param probabilityF The RAPPOR "f" probability, on the range [0.0, 1.0]. This will be
 *        quantized to the nearest 1/128.
 * @param probabilityP The RAPPOR "p" probability, on the range [0.0, 1.0].
 * @param probabilityQ The RAPPOR "i" probability, on the range [0.0, 1.0].
 * @param numCohorts Number of cohorts into which the user pool is randomly segmented.
 * @param numBloomHashes The number of hash functions used forming the Bloom filter encoding of a
 *        string.
 */
Free and open source

Learn more at
https://cleaninsights.org