Carat: Collaborative Energy Diagnosis for Mobile Devices

Adam J. Oliner, Anand P. Iyer, Ion Stoica
AMP Lab, UC Berkeley
{oliner, api, istoica}@eecs.berkeley.edu

Eemil Lagerspetz, Sasu Tarkoma
University of Helsinki
{eemil.lagerspetz, sasu.tarkoma}@cs.helsinki.fi

ABSTRACT

We aim to detect and diagnose *energy anomalies*, abnormally heavy battery use. This paper describes a collaborative black-box method, and an implementation called Carat, for diagnosing anomalies on mobile devices. A client app sends intermittent, coarse-grained measurements to a server, which correlates higher expected energy use with client properties like the running apps, device model, and operating system. The analysis quantifies the error and confidence associated with a diagnosis, suggests actions the user could take to improve battery life, and projects the amount of improvement. During a deployment to a community of more than 500,000 devices, Carat diagnosed thousands of energy anomalies in the wild. Carat detected all synthetically injected anomalies, produced no known instances of false positives, projected the battery impact of anomalies with 95% accuracy, and, on average, increased a user's battery life by 11% after 10 days (compared with 1.9% for the control group).

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques; H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness); D.2.8 [**Metrics**]

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

mobile, battery, energy, diagnosis, analytics, collaborative

1 Introduction

Mobile computing, especially smartphones and tablets, is becoming ubiquitous. Recent work [31] acknowledged the rise of a class of mobile software misbehavior: energy bugs. These bugs add to the list of causes of poor battery life that already includes system configurations, user behavior, and power-hungry apps. Significantly increased battery drain, called an *energy anomaly*, frustrates users, creates poor press for vendors, and can render devices unusable. For such a user, the goal is to understand what is using up the battery, whether or not that is normal, and what can be done.

For some devices, there are third-party apps and OS services for quantifying energy use and in some cases attributing it to specific processes [21]. Unfortunately, a single device has limited diagnostic power because there is no *a priori* specification of normal energy use (c.f. many correctness bugs; crashing is almost always bad). Local instrumentation alone is insufficient to determine whether observed energy use is normal or merely a consequence of local configuration parameters, system or device properties, or user behaviors. Without seeing the app running under different conditions, we cannot say whether changing some aspect of the system would improve battery life or by how much. No amount of local instrumentation can enable these capabilities; the information is simply not present on any single device.

We overcome this limitation by using a *community* of devices; ours is the first collaborative approach to energy diagnosis. Measurements aggregated from multiple *clients* allow us to collect more data more quickly, account (statistically) for individual variation in configurations and usage, say whether energy use is normal, and project the impact of certain actions. Each client occasionally records the battery level and other local data. We aggregate these measurements and compare average discharge rates under different conditions, such as which third-party apps (a common source of battery problems) are running.

If the average discharge rate while running some app A is higher than when A is not running (but any other apps may be), that app is an energy *hog*. A hog may be caused by a coding error (e.g., it prevents the screen from dimming) or because such energy use is intrinsic to the app's function (e.g., it frequently requires the GPS). If an app B is not a hog, it may be an energy *bug* on client X if the average rate on X is higher than the average on all the other clients running B. Energy bugs may be caused by a code error that only triggers under certain conditions (which our analysis tries to discover), configurations, or user behaviors. Distinguishing between hogs and bugs requires a collaborative method.

Our method for diagnosing energy anomalies uses the community to infer a specification (expected energy use), and we call deviation from that inferred specification an anomaly [9]. Unlike previous work, we are looking for regularity and deviation in the use of energy and leveraging this insight to characterize the abnormal use of that resource (the battery). Deviant energy use is an anomaly, regardless of the cause (e.g., coding error or user behavior). Our method further computes diagnosis trees called *MCADs*, which enable us to advise users what actions they can take to improve battery life and to estimate the amount of improvement (accompanied by error and confidence bounds).

SenSys'13, November 11–15, 2013, Roma, Italy.
Copyright 2013 ACM 978-1-4503-2027-6 /13/11 $15.00
http://dx.doi.org/10.1145/2517351.2517354.

Some prior work has aimed to understand energy use by employing a combination of hardware, OS, and app source code or binary instrumentation [11, 23, 32, 44]. In this paper, we present a non-invasive inference method for diagnosing energy anomalies that uses all the information available to a user app on both the Android and iOS platforms. In addition to being a pragmatic point in the design space, our solution naturally possesses several desirable qualities:

- Software-only. Hardware solutions are expensive, require technical skill, and void warranties.
- No kernel modifications. Hacking an OS requires skill; even "jailbreaking" may result in the user bricking their device or introducing bugs or security vulnerabilities.
- Black-box apps. The user does not have access to the source code for most of the apps they run or, usually, the ability to instrument binaries.

Extensions to our method could take advantage of platform-specific information (our implementation does so), but the aim of this paper is to evaluate how far we can take diagnosis without relying on such data. Distribution mechanisms like the app stores make it easy to get instrumentation onto off-the-shelf devices if that instrumentation is a standard app.

We take a black-box approach with process-level granularity; when we observe anomalously high energy use, we implicate one or more processes. Although this restriction may seem severe, for a method that can still be distributed via the App Store, our method is *maximally invasive*. Despite the limitations, these data are sufficient to diagnose anomalies with enough accuracy to provide actionable recommendations that improve battery life in practice.

In this paper, we do the following:

- Present a collaborative inference method for detecting and diagnosing energy anomalies by looking for deviation from typical battery use (see Section 2) and an implementation as an app called Carat for iOS and Android (see Section 3), and
- Evaluate our method with a 500,000-device deployment, showing a 100% detection rate of injected energy anomalies and partial corroboration for the thousands of anomalies we diagnosed in the wild (see Section 5).

The battery life of a device for which Carat generated action recommendations improves by an average of 41% during the first three months (compared with 7.9% for devices without Carat recommendations), 95.2% of the projected battery improvements (e.g., "Killing app A will increase battery life by 45m ± 5m") match the actual improvements within the 95% confidence bounds, and the battery overhead of running Carat is negligible (indistinguishable from running nothing, according to hardware power metering experiments). We conclude with a discussion of the limitations of our approach (see Section 6), an explanation of our place among the related work and how we distinguish ourselves (see Section 7), and a summary of the conclusions (see Section 8).

2 Method

Our method builds and compares conditional probability distributions of rates of energy use to look for *energy anomalies*; e.g., the rates when an app is running on a client with one OS version (the *subject distribution*) may be significantly higher than when running on clients with another OS version (the *reference distribution*). We focus on two kinds of anomalies: hogs and bugs (see Section 2.1). In Sections 2.2–2.4, we compute the magnitude of an anomaly, corresponding to the expected improvement in battery life that an average user experiencing the anomaly would see if they became like the average user not experiencing it. We quantify the error and uncertainty of these projected improvements and decrease that uncertainty by classifying measurements according to various conditions (e.g., rates taken when WiFi was, or was not, available). We generate the classifiers for an anomaly as a diagnosis tree (see Section 2.5–2.6), which we then reduce to a minimal, complete set of actionable recommendations (*MCAD*). An MCAD translates to anomaly diagnoses, such as "With C% confidence, killing app A would increase battery life by $d_1 \pm e_1$ minutes; upgrading to OS version V would increase battery life by $d_2 \pm e_2$ minutes; disabling WiFi..." and so on.

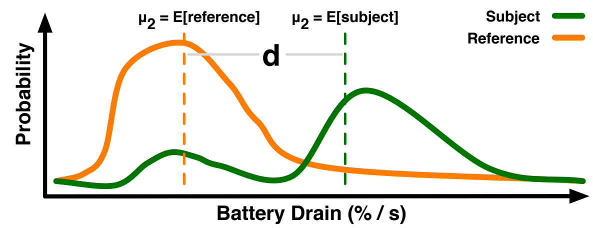

Figure 1: We compare the expected values of conditional distributions of energy drain rates to classify apps as hogs, bugs, or neither. The distance d shown is used to estimate the severity of the anomaly.

2.1 Hogs and Bugs

We define two categories of anomalies, hogs and bugs, by the types of subject and reference distributions we compare. Informally, an app is an energy hog when using that app drains the battery significantly faster, in a statistical sense defined in Section 2.4, than the average app. In contrast, an app has an energy bug when some running instances of the app (the ones in which the bug manifests) drain the battery significantly faster than other instances of the same app (the ones in which the bug does not manifest). Anomalies do not imply incorrect behavior; they may have innocuous causes. Hogs and bugs are computed as follows.

First, we build a (reference) distribution of battery discharge *rates* for devices used normally: playing games, browsing the web, making phone calls, leaving it idle, etc. Introduce an app A into the community, which some subset of clients will install and use, possibly in place of certain other apps. Build another (subject) distribution consisting only of rates observed while A is running. If the expected battery life while A is running is significantly lower than the expected lifetime without A, we call A an energy *hog*.

Intuitively, a hog lowers the community's average battery life. Note that an app may make use of energy-demanding device resources (e.g., WiFi or GPS) without being considered a hog; anomalous apps tend to overuse these resources. An app could be a hog because of a coding error that affects many clients or because an app legitimately needs to use large amounts of energy to serve its function. Regardless, a user seeking to improve their battery life would do well to not have a hog running. Although per-device instrumentation, such as Android provides, can quantify energy use relative to other apps on one device, it cannot say whether that use is abnormal relative to other devices or to apps not running on the device, and so cannot detect or diagnose hogs.

An app B that is not a hog may still use much more energy on some client X. If the expected discharge rate of B running on client X (subject distribution) is significantly higher than that of B running on other clients (reference distribution), we call B an energy *bug* on client X. No amount of instrumentation on a single device can detect or diagnose bugs.

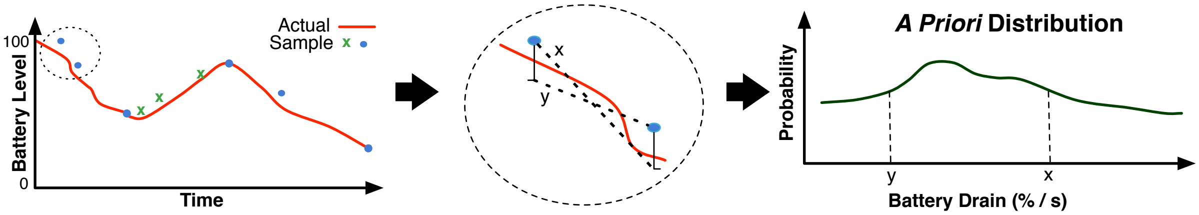

Figure 2: The process of converting battery level samples to rate distributions using the *a priori* distribution. Samples marked with green Xs are discarded because the device was charging. iOS may report a battery level up to 5% above the actual level. The slope bounds (x and y) determine the *a priori* slice.

An energy bug is therefore a pair: an app and a client it afflicts. An energy bug may be caused by a coding error that affects a small group of clients, a rare configuration that uses more energy ("correct" or otherwise), or unusual user behavior (which requires a community to detect). If the buggy app is getting caught in a bad state, restarting the app may return the app to normal; otherwise, the remedy is the same as for a hog. Other actions may be suggested by our diagnosis trees (Section 2.6), but the current app UI does not reflect this.

We added a caveat that a hog cannot also be a bug to distinguish anomalies that affect all or most clients (hogs) from those that affect only a subset. Hogs are unlikely to be fixed by a restart, so we recommend killing them. This difference in appropriate response motivated the naming, and we found the distinction useful.

The subject and reference distributions are built using battery level samples from the community, as we explain in the following sections. The expected values of these distributions converge rapidly to the true expected value as the number of clients increases (see Section 5.7).

Note that even perfect knowledge of app behavior on a single client could not distinguish hogs from bugs; heavy energy use on one device could be a matter of configuration, user behavior, or some other bug trigger that stays static across runs. In order to say whether an app or app instance is anomalous, a community is required.

2.2 Conditional Distribution Model

As discussed in Section 2.1, to detect energy anomalies we compare two distributions of the battery drain (see Figure 1). This section explains how such a conditional distribution is modeled, and how we quantify the associated uncertainty. The input is a set of n *rates*, tuples consisting of a feature vector c and a rate probability *distribution* u, computed from some pair of samples (see Section 2.3). We model these as being randomly sampled from a true distribution U_c, with mean μ and variance σ^2, composed of measurements satisfying predicate c (e.g., iPhone 4 with WiFi access).

We first take the expected value of each u to yield a *rate* r. Consider the conditional distribution R_c of rates r satisfying c. To compute the error and confidence bounds on the expected value of R_c, we model it as n independent samples from U_c. These rates—means computed from a large number of random i.i.d. variables—are therefore approximately normally distributed as $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, according to the Central Limit Theorem (CLT).

This result can also be obtained by starting with the assumption that R_c is distributed as $\mathcal{N}(\mu, \sigma^2)$. Although we do not know the parameters μ and σ^2, we can estimate them using the rates $(r_1, \ldots, r_n)$. The well-known maximum likelihood estimators for these parameters—obtained by maximizing the log-likelihood function—are as follows:

$$\hat{\mu} = \bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2.$$

By the Lehmann-Scheffé theorem, $\hat{\mu}$ is the uniformly minimum variance unbiased estimator for μ: $\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

This agrees with the CLT method. The estimator $\hat{\sigma}^2$, however, is biased, so we apply Bessel's correction to obtain the uniformly minimum variance unbiased estimator for the sample variance:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2.$$

By our normality assumption, we can construct the t-statistic $t = (\hat{\mu} - \mu)/(s/\sqrt{n})$, which has the Student's t-distribution with $n-1$ degrees of freedom. We can approximate the error bounds on this estimate of μ using a standard formula, where h is chosen according to the desired confidence level:

$$\mu \approx\in \left[\hat{\mu} - \frac{hs}{\sqrt{n}}, \hat{\mu} + \frac{hs}{\sqrt{n}}\right] = \hat{\mu} \pm \epsilon$$

For 95% confidence error bounds, $h = 1.96$; we use this value for all experiments in this paper. Crucially, to estimate the mean μ and to assign error and confidence bounds to that estimate, we require only the rates r, not the original distributions u.

As we gather more data, the uncertainty associated with these expected values decreases. We gauge empirically how convergence occurs in practice in Section 5.7.

2.3 Computing Rate Distributions

To compute rate distributions, our method must first convert a set of *samples* from a single client into a set of *rates*. A sample is a measurement taken at a particular point in time that consists of the battery level (%) and a list of features: device model, OS version, names of running processes, battery state (e.g., unplugged), etc. Let $s_t = (b, p, q, \hat{c})$ denote a sample taken at time t, triggered by reason q (e.g., the device was unplugged), where the battery level was observed to be at fraction $0 \leq b \leq 1$ and the battery state was p (e.g., unplugged). The remaining features are denoted collectively as a set $\hat{c}$ of key-value pairs (e.g., "OSVersion=5.0" or "AppXRunning=YES").

First, we sort the samples by t and filter them using the p values to retain only those adjacent samples that span a period during which the device was not plugged in, restarted, or otherwise increasing in battery level: that is, only periods when the battery was discharging. This reduces the initial set of all samples to a set of *consecutive pairs*. We compute discharge rates from these pairs.

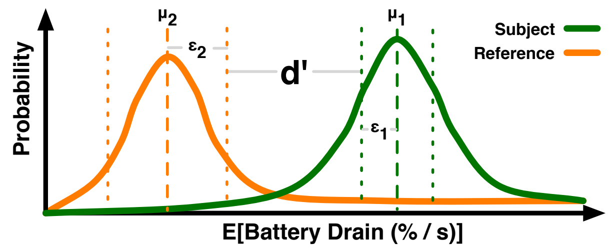

Figure 3: We compare distributions of the expected values of battery drain to identify anomalies ($d' > 0$) and quantify the error and confidence ranges for expected battery drain under different conditions.

Our method allows for imprecision in both the battery level and time measurements by converting a consecutive pair $s_{t_1} = (b_1, p_1, q_1, \hat{c}_1)$ and $s_{t_2} = (b_2, p_2, q_2, \hat{c}_2)$ not to a single rate number but to a rate distribution u. We associate this distribution with a set of features, yielding the pair $R = (u, c)$, computed from the features of the constituent pair of samples, as explained below.

If both endpoints, (b_1, t_1) and (b_2, t_2), are exact, then the rate distribution is $u = \frac{b_1 - b_2}{t_2 - t_1}$ with probability 1. Discharging yields a positive rate.

On iOS, we only get such exact measurements when the `UIDeviceBatteryLevelDidChangeNotification` is triggered. Otherwise, we estimate a probability distribution for the rate. There are a variety of techniques one might employ, depending on the nature of the uncertainty. In this paper, we address the case of iOS measurements, which present unique challenges. Specifically, the API provides battery level measurements at a granularity of 0.05. In other words, if we request the battery level at an arbitrary time during execution and get 0.95, the true level may be in the range $(0.90, 0.95]$.

The true rate, therefore, lies between $\frac{b_1' - b_2}{t_2 - t_1}$ and $\frac{b_1 - b_2'}{t_2 - t_1}$, where $b_1' = b_1 - 0.05$ and $b_2' = b_2 - 0.05$, and subject to the constraint that the rate is nonnegative. Not all values in this range are equally likely, however, so we use this range to take a "slice" of an *a priori* rate probability distribution (see Figure 2), computed using the rates that clients were able to compute exactly, as described above. There was sufficient data in this distribution to bootstrap our method. We convert the slice to a probability distribution by dividing by the slice mass and use it as the rate distribution u.

We compute c from $\hat{c}_1$ and $\hat{c}_2$ by taking the union: $c = \hat{c}_1 \cup \hat{c}_2$. Features like device model do not change between consecutive samples. We conservatively say that an app was running during the period $[t_1, t_2]$ if it was seen in either sample. It would be straightforward to use a different function if the semantics of the features demanded it.

2.4 Comparing Rate Distributions

Let c_1 be the conditions of the subject distribution (e.g., app A is running) and c_2 be the conditions of the reference distribution (e.g., app A is not running). We aim to ascertain whether c_1 corresponds to *significantly greater* energy use than c_2. For this to be answered in the affirmative, we require the following:

$$\hat{\mu_1} - \frac{hs_1}{\sqrt{n_1}} - \hat{\mu_2} - \frac{hs_2}{\sqrt{n_2}} = \hat{\mu_1} - \hat{\mu_2} - (\epsilon_1 + \epsilon_2) \quad > \quad 0.$$

Otherwise, the data does not support the assertion with the desired confidence. Graphically, this corresponds to a positive value of d' in Figure 3.

Carat suggests actions that would improve battery life along with the expected value of that improvement for an average client (starting from full charge and fully draining the battery). The improvement if the client were to change from c_1 (experiencing the anomaly) to c_2 (not experiencing it) follows directly from the distance metric $d = \hat{\mu_1} - \hat{\mu_2}$. Within our confidence bounds, however, the value of d could be as much as

$$e = h\left(\frac{s_1}{\sqrt{n_1}} + \frac{s_2}{\sqrt{n_2}}\right).$$

This is symmetric about the expectation. The estimated improvement is therefore $d \pm e$.

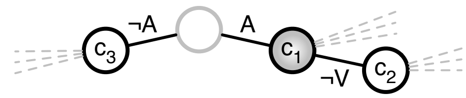

Figure 4: The minimal complete actionable diagnosis (MCAD) for the example anomaly c_1 described in Section 2.6, consisting of c_2 and c_3. The dashed lines indicate nodes and subtrees that, while produced via splits when the tree was constructed, did not meet the criteria for an MCAD.

2.5 Splitting Distributions

In order to more confidently diagnose anomalies, we build a tree that separates conditional distributions by features that significantly affect energy use. Let each conditional distribution be a node in this tree, uniquely identified by its condition c. Starting with some distribution c (e.g., app A is running), iterate through each feature $f \notin c$ and attempt a split by creating new child nodes $c \wedge f$ and $c \wedge \neg f$. For instance, if f is whether the client is running a Galaxy S II, then one child would get the rates from node c taken from Galaxy S IIs and the other would get all other rates satisfying c.

Splitting has two competing effects on the error bounds. First, it reduces n, thereby increasing the error (increasing uncertainty). Second, if feature f divides rates from distributions having significantly different means, then it will likely reduce the sample variance of at least one child and thereby decrease the error (decreasing uncertainty).

A split is performed if the child nodes c_1 and c_2 yield a positive gap, $d' > 0$, as in Figure 3. Splitting generates two leaves, children of c, with edges f and $\neg f$. Otherwise, we make no changes to the tree and proceed to test the next feature. When no more features remain, we can recursively repeat the process on any new leaves.

2.6 Diagnosis

This section describes how to generate a diagnosis for an anomaly, which involves building a tree structure similar to a classification or decision tree [24, 39], and conclude with an example. Consider a node c_1 corresponding to a subject distribution for an anomaly (see Section 2.1). A *diagnosis* is a set of nodes with significantly lower energy use than c_1. Intuitively, a node in this diagnosis is some condition under which the anomaly does not occur. The diagnosis is *complete* if it includes all such nodes.

Let node c_2 be said to be *reachable* from node c_1 if, in the problem domain, it is possible to initially be in a state satisfying c_1 and, by performing some actions, then satisfy c_2. We define an *actionable* diagnosis to be one consisting only of reachable nodes.

A diagnosis is *minimal* if every subtree entirely contained in a complete diagnosis is replaced by its root. The minimal complete actionable diagnosis (MCAD) is unique, but note that it may include paths from c_1 to multiple different states.

For example, consider the node for running app A, $c_1 = A$, with significantly more energy use compared with $\neg A$; it is a hog. Say, for simplicity, that there are only two other features of the

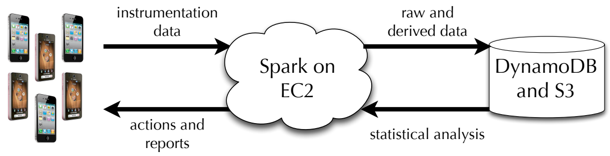

Figure 5: The Carat architecture, consisting of the crowd-based front end, the central server with the analysis running in the cloud, and the stored samples and results.

device—model M and OS version V—and only one other possible OS version. Every node in the subtree rooted at $\neg A$ has significantly lower energy use than c_1, as does every node with $\neg M$ or with $\neg V$. In our domain, a user cannot change their device model, so all nodes with $\neg M$ are excluded from the actionable diagnosis despite showing less energy use. To make the diagnosis minimal, replace with their respective roots the nodes in the subtrees rooted at $A \wedge \neg V$ and $\neg A$. Thus, the MCAD (illustrated in Figure 4) is exactly these two nodes (c_2 and c_3); the interpretation is that the client can improve their battery life either by changing OS versions or killing the hog.

These trees helped diagnose problems in the wild, such as the Kindle bug in Section 5.4.3 where WhisperSync was using far more energy when syncing over GSM. Our analysis discovered the bug was correlated with the iPhone 4 and only occurred on iPads when they did not have WiFi. There are dozens of such diagnoses that we have investigated, and in some cases reported to the developers, and thousands more produced by Carat.

Although the client UI only displays recommendations to kill or restart an app or to upgrade the operating system, our analysis computes diagnoses—and can make recommendations based on—features like internet connectivity status (radio or WiFi), mobility, device model, app versions, GPS activity, the user ID (usually indicating a bad battery or strange user behavior), and so on. Thus, our MCADs can recommend actions like turning on/off the WiFi/GPS/radio, upgrading the app/OS to a newer version, or avoiding an app under certain conditions (e.g., while moving around or when not connected to the internet).

3 Implementation

The Carat architecture consists of a mobile app for device users (see Section 3.1), a central server that collects the data (see Section 3.2), and an analysis running in the cloud (see Section 3.3). Figure 5 shows an overview.

3.1 Carat App

We implemented Carat as an app on both the iOS and Android platforms. It is available as a free download on Apple's App Store, Google's Play Store, and as source code on GitHub, all of which are linked from the project homepage[1]. The clients are lightweight; e.g., the iOS app is ~6000 lines of Objective-C, excluding third-party libraries like Flurry (for collecting usage statistics), ShareKit (for enabling sharing over social networks), Thrift (for handling messaging protocols), CorePlot (for plotting), and several others. This number also excludes auto-generated code related to the UI.

Carat runs as a user-level app on stock devices. This places platform-specific restrictions on what information is accessible and when our app is allowed CPU time to measure it. Our implementation records the following information using the public APIs:

- battery level fraction,
- battery state (e.g., plugged in or unplugged),
- names of running processes (each non-OS process roughly equates to a single user app),
- state of memory (e.g., number of active pages),
- OS and version,
- device model, and
- a unique, anonymous, Carat-specific client ID.

This information resides in persistent storage until the app is brought to the foreground, at which point it communicates with the Carat server over TCP. Our communication model is client-initiated (since they are situated behind NATs) and utilizes Apache Thrift to define the service interface.

The app intermittently transfers stored samples to the server over 3G or WiFi. Since we optimized Carat with respect to energy use, the client invokes a data transmission to the server only when it is running in the foreground and when the user is interacting with the UI. At this time, the app also requests results from the server to update the UI.

To comply with legal restrictions and to alleviate user concerns, our implementation neither records nor transmits personally-identifying information. What it does record is visible within the app (see Section 3.1.1), so the user knows exactly what Carat is measuring. Furthermore, our EULA (required by the App Store and also available on the project webpage[1]) includes an additional clause making it clear exactly what our app will do. Finally, the app is open source under a BSD license and is available on GitHub[1].

Although jailbroken iOS devices allow us to collect more data (e.g., app versions), requiring jailbreaking also would have restricted the size of our userbase, biased our data toward a certain class of users, and prevented us from distributing Carat on the App Store. We opted for less data from more users, and our results demonstrate that energy anomalies diagnosing does not require intrusive instrumentation.

On Android, Carat samples when the `ACTION_BATTERY_CHANGED` Intent fires, at 1% battery level granularity. As we discuss for the remainder of this section, not only is Carat more restricted on iOS than Android with respect to what it can measure, but also when. Carat does not fall into the class of apps that are allowed to run as proper background tasks, which are given intermittent CPU time to perform tasks such as buffering audio, maintaining VoIP server connections, or continuously tracking the GPS coordinates of the device using location services. This means that, in order to take samples while Carat is suspended, our app subscribes to several notifications. When one of these notifications is triggered, iOS allows Carat a small amount of time to take measurements and save these to persistent storage; there is not enough time to communicate with the server.

Carat subscribes to battery-related events (`UIDeviceBatteryLevelDidChangeNotification` and `UIDeviceBatteryStateDidChangeNotification`) and significant location changes (`startMonitoringSignificantLocationChanges`). The location change feature is especially valuable for us. It not only uses far less energy than using the full-fledged location service, but it means that the OS will automatically relaunch Carat if it is terminated while the service is active. (In our deployment, while Carat was in the background, roughly half of samples were triggered by location services and a third were triggered by the battery level event.)

3.1.1 User Interface

When the Carat app is launched, it sends locally stored samples to the server. When Carat is in the foreground, the temporal resolution of sampling increases several-fold. These observations—that increased user engagement leads directly to data being recorded

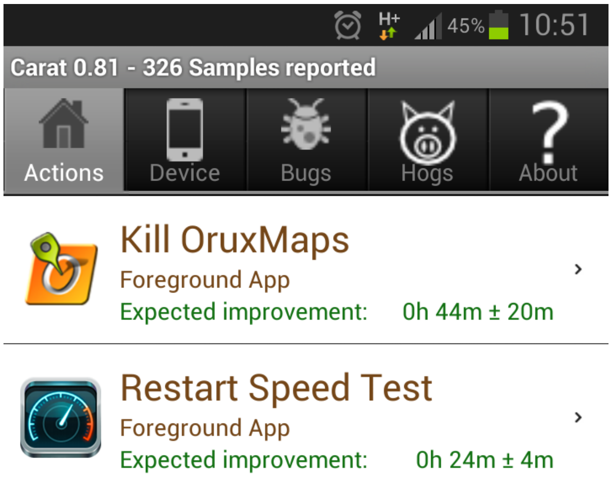

Figure 6: The top of the main screen of Carat on Android, showing recommended actions and projected battery life improvements.

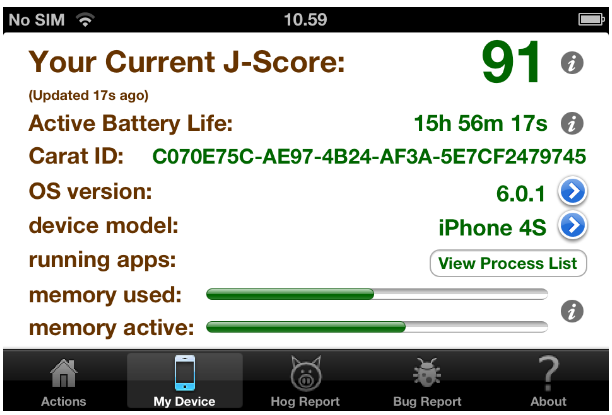

Figure 7: The Device tab on the iOS client. The J-Score indicates the percent of the community with worse battery life than this device.

more often and reported sooner—motivated us to spend time honing the user interface, which we now present.

The main screen of Carat is the Actions list, shown in Figure 6, which presents actions the user can take to improve battery life, based on what Carat has learned about their device (e.g., what apps they run), sorted by the expected improvement if that action is taken. For example, the figure shows an action "Kill OruxMaps" that would result in an expected increase of 44m. This means our analysis observed that a typical device running this game will run a full battery down to zero almost 44 minutes sooner than a typical device running typical apps but not OruxMaps. Carat will suggest restarting bugs, admitting the possibility that the instance is caught in a bad state; if restarting does not help, it may be a configuration problem or specific to user behavior. Finally, our current implementation suggests upgrading the operating system if it observes that a newer version is correlated, across the community, with better battery life. The current UI does not reflect all information present in the diagnosis trees; that is planned for a future release.

The Device tab displays information about the client's device, including most of the information that is being recorded and transmitted to our server: the process list, the device model and OS, the state of memory, etc. This tab also prominently displays a number called a J-Score, which is the percentile into which the client's battery life falls within the community; a J-Score of 65 means a better active battery life than 65% of similar devices. Active battery life is computed based on Carat sampling and omits idle periods. This client's average battery drain when using the device would fully deplete the battery in about 16 hours.

We created the J-Score (see Figure 7) to increase user interest and sharing, hoping that it would introduce an element of social competitiveness to energy efficiency. It appears, anecdotally, to have worked. For instance, upon observing that her score had dropped precipitously due to an influx of new users, one user remarked (tongue-in-cheek) that she was "no longer confident in our analysis results." She continues to check her score regularly, incidentally sending us samples each time.

The Actions list only suggests killing or restarting an app that is currently active (i.e., in the process list). The Hogs tab shows the top hogs ever reported to have run on the device. The same is true for bugs under the Bugs tab. Clicking on one of the hogs or bugs brings up a detail page where the user can explore the data further.

3.2 Carat Server

The Carat server collects samples from instances of the Carat app running on clients' mobile devices and stores them for use by the backend analysis (see Section 3.3), and it serves actions and other analysis results to clients.

The server is a $<$1300-line Java application (excluding code autogenerated by Thrift) that listens on TCP port 8080 for incoming client connections. We host with Amazon EC2 because it provides a mechanism to scale the server by spawning new instances and to run a load-balancer to distribute incoming connections.

Received samples undergo lightweight processing to remove junk or malformed data and are then sent to persistent storage. This preprocessing removes OS daemons from the list of processes. We manually maintain a blacklist of such daemons, as it does not appear that the iOS API provides enough information to determine this automatically.

3.3 Backend Analysis

The Carat analysis consists of approximately 5000 lines of Scala, written in the Spark framework [45]. Spark is a cluster computing framework designed for iterative and interactive jobs, distinguished by its use of Resilient Distributed Datasets (RDDs). RDDs are read-only collections of objects partitioned across a set of machines that can be rebuilt if a partition is lost. Parallelism in Spark is provided through operations on the RDDs (e.g., `map`, `reduce`, and `filter`).

Existing data-flow based frameworks such as Hadoop or Dryad depend on intermediate data being written and read from disk, incurring a huge performance hit for iterative jobs. In contrast, Spark provides an efficient environment for multi-stage jobs by reusing the same worker nodes across iterations. In addition, it provides a robust programming model for interactive queries where it is desirable to load data into memory and query it repeatedly (with different filters). These features, along with fault tolerance and its memory management model, made Spark a good fit for implementing Carat's analysis.

The production version of Carat runs in a 20-node cluster composed of high-memory Amazon EC2 instances. This section provides an overview of Spark, the challenges related to parallelizing our analysis, and our solutions.

After converting samples to rates, the computation proceeds in two main stages: identifying hogs and bugs and then generating MCAD trees (see Section 2). The first stage is summarized in Algorithm 3.1

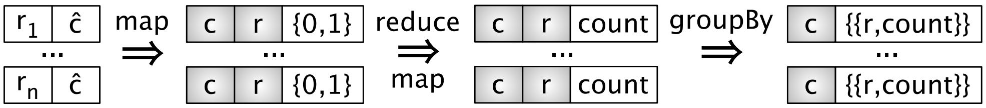

Figure 8: The parallelization process starts with rates as an RDD. Each rate r has features $\hat{c} = (c_1 \dots c_n)$. To compute rate distributions on feature c (e.g., each app), we map the RDD to a structure with (c, r) as the key (shaded) and, as the value, 1 if the feature occurs and 0 otherwise. A reduce operation yields the rate frequencies for features. We map again, now with c as the key, and (r, count) as the value. Grouping by key then gives the frequency of every R for every F. With slight modifications to the mapping and grouping fields, we use this parallelization strategy for hogs, bugs, J-Scores, etc.

Algorithm 3.1: ANALYZERATES($allRates, aDist$)

---

comment: Hog detection

for each $app \in allApps$

do {
- $filt \leftarrow$ ALLRATES.FILTER(app in _.allApps)
- $filtNeq \leftarrow$ ALLRATES.FILTER(app not in _.allApps)
- $d' \leftarrow$ COMPAREDISTRIBUTIONS($filt, filtNeq, aDist$)
- **if** $d' > 0$
  - **then** {**comment:** store hog and distributions

comment: Bug detection

for each $id \in allIds$

do {
- $fid \leftarrow$ ALLRATES.FILTER($\_.id = id$)
- $notFid \leftarrow$ ALLRATES.FILTER($\_.id{!=}id$)
- **comment:** Consider apps reported by id, omit hogs
- $fidNonHogs \leftarrow$ FID.MAP($\_.allApps$) $\setminus$ $Hogs$
- **for each** $app \in fidNonHogs$
- **do** {
  - $appFid \leftarrow$ FID.FILTER(app in _.allApps)
  - $appNotFid \leftarrow$ NOTFID.FILTER(app in _.allApps)
  - $d' \leftarrow$ COMPAREDISTRIBUTIONS($filt, filtNeq, aDist$)
  - **if** $d' > 0$
    - **then** {**comment:** store bug and distributions
- $scoreDist \leftarrow$ GETDIST($fid, notFid, aDist$)
- **comment:** Save scoreDist for J-Score calculation

comment: Write J-Scores based on the processed distributions

In Section 2.3, we discussed how Carat converts consecutive samples into rates. This computation involves a dependency between samples that complicates the parallelization process.

To remove this inter-sample dependency, we create RDDs of consecutive sample pairs. This new RDD is free of dependencies, so the Spark runtime can independently assign data and conversion tasks to workers. This is done by applying a `map` operation to every item in the RDD. The result of this operation is another RDD consisting of rates. We add metadata for backtracking.

3.3.1 Parallelizing Distribution Building

The bulk of Carat's analysis is the process of building and comparing rate distributions. We load the rates into an RDD, which Spark automatically distributes to all compute nodes. The parallelization strategy must compute distributions on features in parallel. That is, when building distributions on feature c, the technique must compute distributions for all values of feature c. We devise such a strategy using Spark's RDD operations as follows.

We begin with items in the rate RDD, composed of rates r and their associated features $(c_1, ..., c_n)$, split among worker nodes. We compute distributions of rates conditioned on c and compare them with distributions satisfying $\neg c$. (We compute the distribution for $\neg c$ by subtracting the distribution for c from the full distribution.)

The first step maps items to the format $((c, r), \{0, 1\})$, keyed on c and r and with a value of 0 or 1, indicating the presence of the rate, computed from the apriori (see Section 2.2). A `reduce` operation computes the frequency of each (c, r) pair. We remap the reduced RDD and make c the key and (r, count) the value. When we apply a `groupBy` on the key, we obtain the frequency of every rate for every value of c, or a sequence of $(c, (r, \text{count}))$ (see Figure 8).

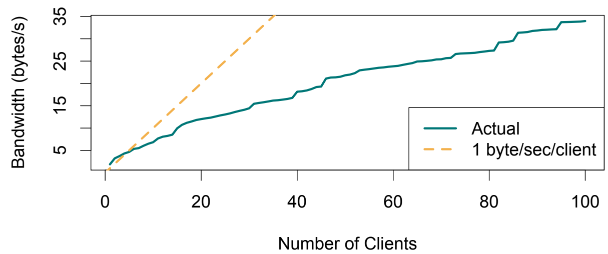

Figure 9: The Carat server sees minimal traffic from individual clients, and the growth of this traffic is linear in the number of users.

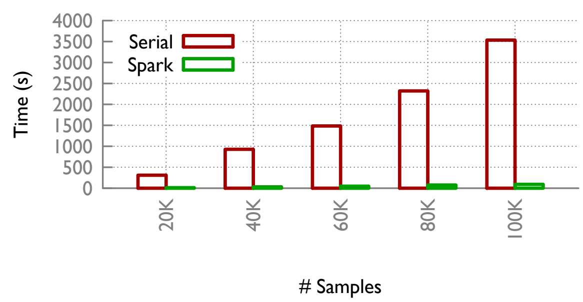

Figure 10: The Carat analysis scales almost linearly when parallelized, while a serial implementation shows exponential complexity.

We now have two RDDs, one with the frequency of rates satisfying c and its complement. The RDDs are joined using a `groupWith` operation. A final `map` operation passes them through our distribution building and comparison module in a parallel fashion, thus obtaining the expected improvements and the correlations. The same parallelization strategy is applied to compute hogs (features are apps), bugs (features are (UserID, App) pairs), J-scores (features are UserIDs). We observe that most other feature-grouping required in Carat's analysis can be reduced to this parallel model.

3.4 Performance and Scaling

The success of our approach depends on an active community and generates better results as that community grows, so the implementation must be scalable.

Our frontend experienced linear traffic scaling with the size of our deployment, at a rate far below 1 byte per second per client (see Figure 9). Sample reporting is presumed to be unreliable; a client with no disk space or network access is allowed to throw away samples and an overloaded server may drop packets. Five

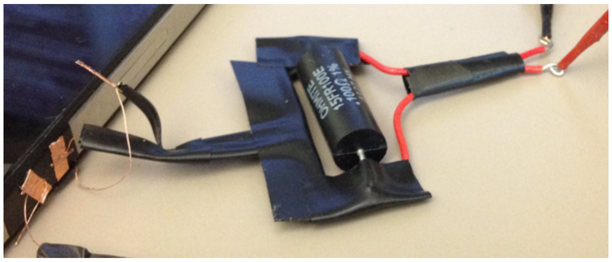

Figure 11: Close-up of the wiring rig that connects our iPhone 4S test phone with the Monsoon Power Monitor.

medium Amazon EC2 instances behind an Elastic Load Balancer (ELB) has been handling our userbase of half a million devices.

Our current implementation of the analysis backend (see Section 3.3) uses the Spark cluster computing framework. The computation is massively parallel, as every distribution and comparison can be computed independently. Figure 10 compares the runtime for an optimized serial implementation of the analysis algorithm compared to a parallel implementation in Spark for increasing number of samples. The results underline the need for parallelization. As our userbase grew, we made numerous optimizations. The analysis program now computes all reports for all our users (24 million samples) from scratch in approximately 45 minutes.

4 Ground Truth and Overhead

For Carat to accurately account for when energy is being used, it must convert intermittent (low precision) battery level samples into energy drain rates in a way that is faithful to the ground truth. Furthermore, the practicality of our method relies on sampling that is sufficiently low-overhead that it does not have a significant impact on the energy use, itself. In this section, we attach mobile devices to power metering hardware: an iPhone 4S to a Monsoon Power Monitor[2] (see Figure 11) and a Galaxy Tab 2 10.1 to Leyden Energy's[3] battery-testing equipment. Our results confirm that Carat generates accurate energy distributions while consuming few resources (i.e., almost no battery).

To test the fidelity and cost of our sampling, we ran the devices through a script of varied activities. The script is not intended to be a representative workload, but to repeatably exercise the device features and drain the battery at different rates. It includes such behaviors as downloading and running an app, browsing the web, playing a game, and idle periods. The WiFi was turned on for some periods and off for others.

On each device, we ran through the script under three different arrangements: (1) hooked up to the power meter with and (2) without Carat running and (3) not hooked up to the power meter with Carat running. We compare the data from (1) and (2) to quantify the overhead of running Carat; we compare the data from (1) and (3) to ensure the meter was not influencing Carat's measurements and to assess the fidelity of our sampling and rate estimation. For the runs performed without Carat, where our app appears in the script, we substituted the standard Weather app.

The battery levels reported by the OS, both through the API (Carat samples) and the on-screen indicator, track the actual use of power by the device. Figure 12 shows the iOS data. Between 00:30 and 1:30, Carat took no samples and conflated a higher-rate period with a lower-rate period. Higher frequency sampling would have avoided this error.

The expected energy discharge rates computed from the Carat samples approximate the values computed using power metering hardware. During the 9-hour iOS experiment, Carat took 9 samples at 5% granularity; the power meter took 13,549 samples at effectively 0.0001% resolution. Carat overestimates the average discharge rate by only 0.00088%/sec (see Figure 13). On the Galaxy Tab, where Carat took twice as many samples as on iOS (19), the error is an order of magnitude less (0.00015%/sec). This accuracy is possible thanks to the *a priori* distribution, which uses knowledge of community behavior to refine noisy and incomplete measurements; imprecision in per-client measurements is further mitigated by the statistical backend analysis.

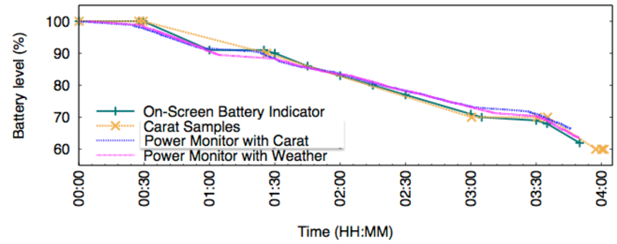

Figure 12: The battery levels during our iOS power metering experiments, either taken directly from the on-screen battery indicator, the Carat samples, or computed from the meter's readings.

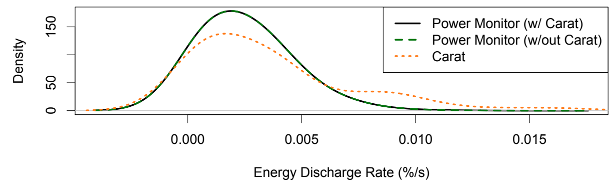

Figure 13: The energy rate distributions from our iOS power metering experiments, smoothed with a Gaussian kernel estimator for visibility. Using the *a priori*, Carat is able to faithfully estimate the distribution with sparse sampling, overestimating the mean energy drain rate by only 0.00088% from 9 samples.

Carat imposes negligible energy overhead. Our power metering hardware indicates that running through our iOS script with Carat running used *less* energy (53.691 mAh or ∼3.5% of the battery less) than executing that same script with the Weather app running in its place (i.e., 54 minutes less battery life running Weather instead of Carat). We also ran the script without substituting another app but found battery life with Carat running was slightly higher than without; Carat's energy use is less than the experimental precision. Similar results held on Android. We can afford to perform sparse, low-overhead sampling on individual clients because we aggregate such data from many clients.

5 Deployment Evaluation

Carat became available as a free download on Apple's App Store and on Google's Play Store in mid-June of 2012. Days later, it was featured on the popular TechCrunch blog[4]; the story was soon picked up by dozens of other news sources. Within 24 hours of the article's publication, we went from a few hundred users to more than 100,000. This doubled in the subsequent 24 hours. Carat has been installed more than 560,000 times; of those, 475,041 clients reported data (some never ran the app or never when connected to the internet); 409,867 reported enough data to yield diagnoses.

Our salient results (see Sections 5.4–5.7) are that we found no instances of false positives among the reported anomalies; after two weeks, users who received Carat recommendations improved battery life by 13% (c.f. 3% for those who did not); and 95.2% of the predicted battery life improvements fell within the predicted 95% confidence bounds.

Device Model	Number	% Total	% Platform
iOS			
iPhone 4S	85,267	20.8	37.6
iPhone 4	54,853	13.4	24.2
iPhone 5,2	12,590	3.07	5.56
iPhone 3GS	12,364	3.02	5.46
iPhone 5,1	12,239	2.99	5.40
Other	49,258	12.0	21.7
Android			
unknown	22,057	5.38	12.0
GT-I9100	15,770	3.85	8.60
Galaxy Nexus	10,333	2.52	5.64
GT-I9300	7238	1.77	3.95
GT-N7000	5009	1.22	2.73
Other	122,889	30.0	67.0

Table 1: The most common device models in our deployment, showing the percent of users from whom we had sufficient data to generate diagnoses.

OS Version	Number	% Total	% Platform
iOS			
5.1.1	136,485	33.3	60.2
6.0	35,708	8.71	15.8
6.0.1	21,068	5.14	9.30
6.1	10,009	2.44	4.42
Other	23,301	5.69	10.3
Android			
4.0.4	40,512	9.88	22.1
4.0.3	24,439	5.96	13.3
2.3.6	19,782	4.83	10.8
unknown	18,075	4.41	9.86
Other	80,488	19.6	43.9

Table 2: The most common operating system versions in our deployment, showing the percent of users from whom we had sufficient data to generate diagnoses.

5.1 Data

Our users ran iOS (55%) and Android (45%). Tables 1 and 2 show breakdowns of the most common device models and operating systems. In aggregate, the devices recorded 16.5 million rates, launching our app 7.4 million times (a median of 1.9 sessions per day).

The community ran 102,421 different apps, with a disproportionate number (56%) coming from Android users. Of these apps, 10,110 (9.9%) were classified as hogs, of which 83% were Android apps. Carat detected energy bugs in thousands of apps; of the 21,529,249 total possible bugs (user-app instance pairs), 1.1% were classified as such.

Clients reported samples at a wide variety of rates, clustering into casual users recording a few samples daily and heavier users sampling sometimes a hundred times as often. The average number of samples per day was nearly the same on both platforms (36.8 samples per user per day on iOS and 37.7 on Android), but the variance of this rate on Android was 32% higher than on iOS. This is, in part, because some Motorola devices only triggered the battery level intent at 10% levels while most other Android devices triggered every 1%; iOS devices triggered consistently at 5% increments.

5.2 User Behavior

The frequency and duration of user engagement matters. The more often users launch Carat, the fresher our data will be (that is when it is sent to our server). On both iOS and Android, the longer users keep Carat in the foreground, the more samples it can record. The session length data (see Table 3) and click-path data show that many stay in the app to explore the reports or check their J-Score.

Session Length	Sessions	% of Sessions
0–3 secs	257,632	4.15
3–10 secs	893,793	14.4
10–30 secs	2,100,538	33.9
30–60 secs	1,397,873	22.5
1–3 mins	1,109,035	17.9
3–10 mins	163,478	2.63
10+ mins	282,645	4.56

Table 3: The length of Carat sessions. The app only reports data when it is opened and can sample more aggressively in the foreground. So, incentivizing the user to open the app and explore results from within the UI helps us collect more data.

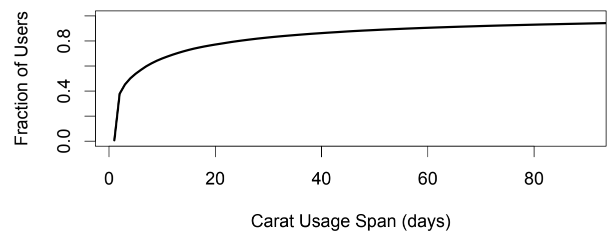

Figure 14: The number of days over which Carat users open the app. Some users check Carat only over a period of several days (to see their initial reports) and then never again; the majority, however, check back with the app occasionally over the following weeks or months.

Almost half of the sessions last more than 30 seconds.

Figure 14 shows the period of time over which users open Carat. After a month, we retain roughly 25% of our users; only 6% use the app for more than 90 days. The median user opens Carat 1.9 times per day and 3.0 times per week.

5.3 Injected Anomalies

We added energy anomalies to an existing app—initially with no apparent misbehavior—to confirm that Carat is able to detect the new bugs. We chose the Wikipedia Mobile app made by Wikimedia Foundation for iOS because it is an open-source app used by many of our clients but was not reported as an anomaly. We added several behaviors to the Wikipedia app that consume large amounts of energy when activated, with each one repeatedly using a different resource: radio, CPU, and GPS.

We installed the buggy Wikipedia instance on one of our test devices, an iPhone 3GS. Wikipedia Mobile was already in use by several clients at this point, so a baseline distribution had been established and Carat did not consider the app to be anomalous. We ran the app for one day for each injected bug (i.e., radio, CPU, and GPS), activating the app a handful of times during the day but only leaving it open for a couple of minutes (casual use). At the end of the third day, we ran the analysis with the real, non-buggy data as the reference distribution and once each with the data from exactly one of the buggy days as the subject distribution. Thus, we could declare success if the analysis reported three bugs, one for each injected behavior.

Indeed, after performing the injection, Carat correctly detected each of the three bugs (no false negatives). Figure 15 shows the reference distribution and each of the three subject distributions for the iPhone 3GS running our buggy Wikipedia build. The expected improvement reported for fixing each bug (i.e., returning the app to typical Wikipedia Mobile behavior) was 27m 26s for the CPU bug, 9m 22s for the GPS bug, and 55m 28s for the Radio bug, which agreed with what the experimenter observed on the device.

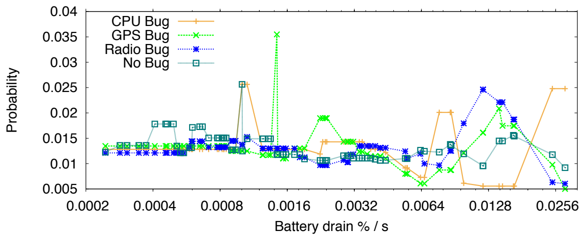

Figure 15: The reference (anomaly-free) and anomalous rate distributions for the modified Wikipedia Mobile, using only the a priori from the private deployment. Carat successfully detects all of the injected bugs.

5.4 Wild Anomalies

Carat detected 10,110 hogs and 233,258 buggy app instances among the 102,421 apps run by the 409,867 users for whom we had sufficient data to generate reports. We ranked the hogs and bugs by a function of severity (predicted battery impact) and popularity (number of users than ran the hog or had a buggy instance), resulting in one list for each kind of anomaly. Although our manual validation process prevented us from checking the entire list, we did check the first two dozen from each list using a combination of user complaints, news coverage, analysis tools (see Section 5.4.1), or experimental results in the literature (e.g., [32, 33]). Among these anomalies, there were no false positives. Later in this section, we describe a subset of these manually-checked anomalies that we feel highlight interesting circumstances or salient aspects of our analysis (see Sections 5.4.2–5.4.3). Note that the number of apps for which we performed manual validation ($\sim$50) already makes this paper a high-water mark for evaluating energy diagnosis on mobile devices, even without considering the other 100,00+ apps that Carat analyzed or the many thousands of diagnoses it generated.

Our attempts to acquire the tools used in prior work to validate our results were unsuccessful; the authors either did not respond, told us the tools were not in a state to be used by people other than themselves and they didn't have time to help us, or they simply refused to furnish the tool. Regardless, no existing tool that we know of would have allowed us to validate all tens of thousands of anomalous apps and app instances that Carat discovered.

5.4.1 External Validation with ARO

AT&T provides a tool called the Application Resource Optimizer (ARO) that uses network traces to identify communication-related misbehavior. We selected the four most severe hogs (GO SMS Pro, Advanced Task Killer, Line: free calls and messages, and Chant for Twitter) and four non-anomalies (Lookout Antivirus, Facebook, Gachinko Tennis, and Dropbox) on Android that showed a strong correlation between increased energy use and network connectivity.

The tool indicated that all four hogs had bursts of network communication that could be more tightly grouped. Three were missing cache headers that might have reduced retransmission; the fourth, Advanced Task Killer, was implicated for wasting energy by not closing network connections. Although half the non-anomalies also lacked cache headers, they did not perform redundant downloads like some of the hogs. ARO corroborated these hogs, but also gave some indications of misbehavior by the non-anomalies; only the accompanying energy measurements separated the misbehavior that hurts battery life from that which doesn't. Furthermore, without a collaborative method like Carat that collects data from multiple devices, it is hard to say whether any of this behavior is intrinsic to the app or a function of device- or user-specific factors.

5.4.2 Hogs

Of the 102,421 apps seen during our deployment, 10,110 (9.9%) were categorized as hogs. (Before checking for statistical significance, there were 15,038 (14.7%).) Recall that an app is a hog if the community-wide average discharge rate while running the app is significantly greater than the average rate while not running it (see Section 2.1) and that we can compute the expected improvement in battery life by killing a hog (see Section 2.4). Hogs may be caused by an oft-triggered code bug or may be simply intrinsic to the app. Users concerned about battery life are advised by the Action list to kill hogs; the user is not concerned about the intention, or lack thereof, behind the energy use.

While some hogs were unsurprising to us (e.g., Pandora and Skype), others were (e.g., some Android themes and wallpapers). For instance, while most apps for searching airline fares and booking flights are not among the hogs—they use the network but not heavily and do not use many other resources—there were a handful of such apps that appeared among the top hogs. We discovered that all those airline apps were written by the same developer and were suffering from a systematic programming inefficiency.

The top ten hogs (by severity) on iOS all fall into the category of utilities, including iDesp Money (for budget management), Ushahidi (for sharing stories within a community), and the Citi Mobile banking app. There were no games; despite being typically resource-intensive, they did not use energy as anomalously as other kinds of apps. Similarly, the top hogs on Android were primarily utilities, but there were also several wallpaper apps (e.g., Beach at Night and Heart and Love) and one game (which has since been removed from the app store).

We now describe a couple of hogs from among those we manually checked (again, there were no false positives) and cite corroborating evidence that the app does, indeed, consume an anomalously large amount of energy.

Pandora Radio: Carat classifies Pandora Radio, which 7116 iOS users ran, as a hog and says killing it will increase an client's average battery life by 50m 43s. This is corroborated by user reports, one of which claimed Pandora drained the battery to 30% in a few hours even with the screen off[5]. To improve battery life while using Pandora, the MCAD suggests using WiFi for connectivity (an additional 25–35m). Pandora is an example of an intuitive hog, as it uses several energy-hungry resources, but Carat quantifies the cost.

Skype: 27,741 iOS clients were running the Skype VoIP app, which was also reported as a hog. This is also confirmed by the forums; one user even used the term "power hog" to describe Skype[6]. Skype's energy use is driven by network connectivity; when no network connection is available, expected battery life is about 6.5h above average.

Go launcher exe new theme. . . : (*sic*) Is an unlikely hog on the Android platform that costs most users between 2h 1m and 2h 53m of battery life. Experiences with Go Launcher and its variants, which change the UI of the device, vary among users[7], but generally "fancier" themes and widgets cause higher battery drain[8].

Live wallpapers: Carat identifies several Android Live Wallpapers as energy hogs. Two that rank among the top 10 most severe hogs on the Android platform are Beach at Night[9] and Heart and Love[10]. They cost most users 2h 33m–2h 49m and 2h 37m–2h 51m battery life, respectively. Both are ad-supported; the detrimental effects of adware are known [32]. Both live[11] wallpapers[12] and adware[13] have been blamed for abnormally fast battery drain.

5.4.3 Bugs

Recall that a bug is an app that is not a hog (it usually consumes below-average energy) but consumes far more energy on some clients

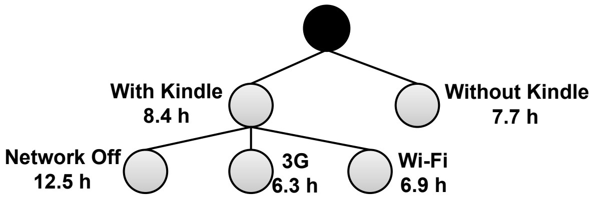

Figure 16: MCAD for the Kindle app on iOS, showing the expected battery life when using exclusively this app under various conditions. The diagnosis points to network connectivity as the primary determinant of energy use. Note that, as with all bugs, Kindle uses less energy than a typical app ("Without Kindle") when the bug is not triggered.

than others (see Section 2.1). Although the current Carat client-side UI only suggests restarting a bug (in case it is simply caught in a bad state), the MCAD diagnosis computed on the backend enables more specific recommendations, such as disabling WiFi or turning on GPS; we plan to add this in later versions of the app. Note that, without a community of clients, distinguishing bugs from hogs would be impossible and identifying the triggers would be difficult.

The maximum number of bugs that Carat could report is the sum over clients of the number of non-hog apps they ran, which was 9.1 million in our dataset. Our method reported 233,258 buggy app instances (1.1%); we describe some examples below.

Many popular apps, including Facebook and Youtube (on iOS) and Twitter and Chrome (on Android), exhibit anomalously high energy use among small subsets of users. This suggests that those apps have configurations or usage modes that consume significantly more energy. By severity, however, most of the bugs are again less popular utilities: e.g., Koder and Raved on iOS and Police Scanner and Are You Watching This?! on Android. There were two games among the top ten most severe bugs: Tower of Fortune (iOS) and Papaya Diamond (Android). Unlike the Android hogs, no wallpapers were among the top bugs.

Kindle: This electronic book app was reported as a bug for 254 out of 2617 iOS clients (9.7%). Figure 16 shows a diagnosis tree for Kindle, in which 3G connectivity appears especially detrimental. The support forums blame the problem on WhisperSync[14], which synchronizes notes, bookmarks, previous location, and Popular Highlights. When syncing over GSM, in particular, the device uses much more energy than syncing over WiFi. Our data support this hypothesis, which had previously been only anecdotal.

Facebook Messenger: Was anomalous on 792 of 7350 Android clients (10.8%). The MCAD indicates that upgrading the OS improves battery life (71–83m), and that WiFi is more energy efficient than other connectivity options. Using the app while stationary gives a 63–97m boost to battery life. (Note that Carat does not advise users to stand still.)

YouTube: Was a bug on 3118 of 37475 iOS clients (8.3%). The MCAD shows that while moving, users of mobile Internet have a battery life advantage over WiFi users (25–34m). When compared to immobile WiFi users, mobile network users still have a 20–28m advantage. This is contrary to many apps, where WiFi is less energy-consuming.

Twitter: Was reported as a bug on 2744 of 18651 Android clients (14.9%). The MCAD for Twitter indicates that the most critical cause of battery drain is an old OS version. Users of Ice Cream Sandwich (4.0.4) got 94m to 100m more battery life than other Android Twitter users. Use of WiFi with 4.0.4 yielded another 85m to 105m; this was not observed on other OS versions.

SwiftKey: A popular keyboard application for Android, SwiftKey is one of the top 15 bugs by severity, affecting 2402 users. The developer website indicates that the latest release of the app exhibits high energy drain, especially in newer versions of Android OS[15].

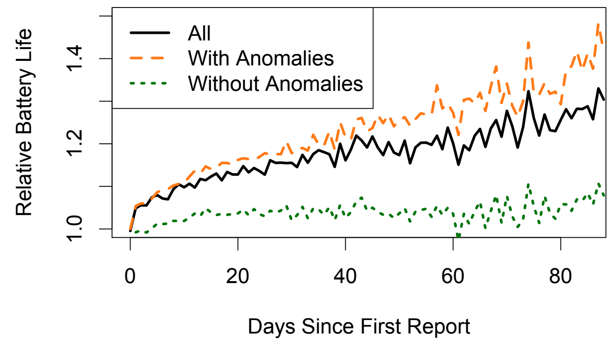

Figure 17: Average relative battery life of Carat users following the generation of their first report (hog and bug lists), using the battery life of the first day as the baseline. A typical user (black line) sees an 7.0% increase after a week, surpassing 23% after two months.

5.5 Diagnosis on Other Features

Carat analyzes the battery life implications of many other combinations of features on the backend as part of the MCAD generation, including the OS version, device model, internet connectivity, and so on. For various reasons, the Carat UI does not recommend that a user take actions like purchasing a different device model or downgrading to an earlier operating system version (those features are not actionable, as discussed in Section 2.6). Other than killing or restarting apps, the only action our current Carat implementation might suggest to users is to upgrade the operating system.

iOS 5.0.1: Shortly after Apple released iOS 5.0, many users complained of issues with poor battery life. The subsequent point release—iOS 5.0.1—was touted, in part, as a fix for these problems. The public reaction was mixed[16]. One user said, "After updating I am seeing my power drain at a much quicker rate"; another claimed his phone was "Still draining at the exact same rate"; and a third, meanwhile, reported that his battery life was "doing much better." In summary, users had a wide variety of anecdotes but no data.

Using the data from our deployment, Carat discovered that, in fact, the average discharge rate for devices running 5.0 was higher than for devices running 5.0.1. Clients running 5.0.1 should expect to see, on average, a 1h 11m 30s increase in battery life, supporting Apple's claims that the update addressed some of the battery problems in the initial release. Users running iOS 5.0 at the time 5.0.1 was released (and this diagnosis was computed) were advised by our app to upgrade.

5.6 Battery Life Improvement

One key metric metric is whether battery life tends to increase over time for our users, a coarse measure of whether using Carat reduces energy use. The metric is coarse because it includes several confounding factors: some of these users may not have followed Carat's recommendations, the population is biased toward users who originally had battery problems (and thus installed Carat), and users may have also employed alternative means to decrease energy use. Some users did not run any apps that Carat considers anomalies and therefore did not receive any reports; that is our control group. Figure 17 shows average relative battery life over time for Carat users who did ("With Anomalies") and did not ("Without Anomalies") receive reports. (The increased variance at higher "Days Since First Report" is due to user attrition; see Figure 14.)

After 2 weeks, the average user sees an 11.7% improvement in battery life, however, users who received reports saw a 13% increase while those who did not gained only 3%. This is more pronounced for long-term users (90+ days); when Carat recommended battery-saving actions, users improved battery life by 41%, compared with 7.9% when Carat did not.

Although users who received recommendations from Carat had a marked improvement in battery life, we considered the possibility that the improvement may have arisen through actions other than those specifically suggested by our app. For example, upon being told to kill App X, the user might instead simply restart their phone, kill all the running apps, or coincidentally stop using App X as part of normal app turnover. This may be partly true, but the data also clearly show that users are performing the actions Carat presents to them; after receiving their first report, anomalous app usage (hogs and bugs) decreased by 60%. This is almost double the decrease for non-anomalous apps (33%). (A number which is probably higher than turnover in the general population due to the more prevalent device restarting and app killing among our users.)

These data suggest that not only do users who receive reports manage to significantly improve their battery life, but that they are following the recommendations contained in those reports. Performing the Carat actions yields increased battery life.

5.7 Improvement Prediction Accuracy

A second key metric is how closely the Carat Actions—and the projected benefits—match the observed benefits. Specifically, when Carat predicts that killing/restarting an app a will improve battery life by $b \pm e$ seconds with 95% confidence, how often is it correct? We found that Carat tended to underestimate the improvement that clients would experience, but 95.2% of these predictions fell within our 95% confidence bounds.

We reached this number using the following analysis. Let $x_{u,a}$ be the fraction of the time that user u reports running app a, within some window of time. The estimated battery life improvement b (in seconds) that Carat quotes assumes a transition from $x_{u,a} = 1 \rightsquigarrow 0$. We assume that the achieved benefit is linear in Δx, so moving from $x_{u,a} = 1 \rightsquigarrow 0.5$ (using the app half half the time instead of all the time) yields an improvement of $0.5b$ seconds; transitioning from $x_{u,a} = 0.5 \rightsquigarrow 0.3$ yields an improvement of $0.2b$ seconds. (Other actions that Carat suggests, such as upgrading the operating system, cannot be done fractionally.) The predicted benefit b is therefore a slope; we compare the predicted improvement curve $y = bx$ (and error margins) with the empirical curve—a least-squares best-fit line through the actual battery life and usage numbers collected by the app—with slope b'.

As stated above, the data show that if Carat advises killing an app and that doing so will increase battery life by $b \pm e$, then across all recommendations made by Carat there is a greater than 95% chance that decreasing the frequency of app use will result in the projected improvements (subject to the scaling described above).

As the number of clients and samples increases, so does the accuracy of our predictions. In particular, Carat's estimate of the expected value—the crucial number used to identify anomalies and compute expected benefits—tends to converge to the true value. Figure 18 shows the shrinking relative error envelope of this estimate for some of the anomalies Carat detected in the wild.

There is no guarantee of convergence in practice because the true rate distribution may be neither stationary nor identically distributed. Indeed, this paper has discussed at length one situation where a rate distribution may not be identically distributed across clients: the presence of an energy bug. As long as a bug affects a constant fraction of the population, however, this convergence happens almost surely, in the mathematical sense (as the number of samples goes to infinity, the estimated expected value converges to the true value with probability 1).

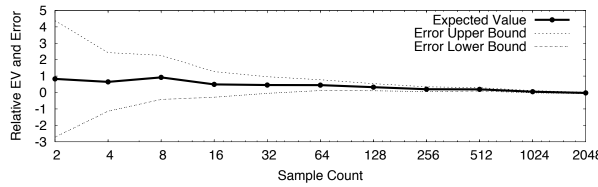

Figure 18: As the number of samples increases, the relative error in our estimate of the expected discharge rate shrinks rapidly. Above is the average expected value for several of the largest anomalies seen in our deployment and the 95% confidence error envelope.

6 Limitations and Future Work

Carat takes a black-box approach to diagnosing anomalies, which carries inherent limitations. Without visibility into the mechanisms (e.g., code, messages, or kernel state) and without the ability to perturb the system (i.e., it is passive and cannot modify other apps), the best possible result is to say what aspects of the system are likely to be involved with the abnormal battery discharge. This is what Carat provides, and it does so by correlating real-valued signals from features without initial assumptions about their relationships. This kind of approach has proven fruitful in prior work [26, 28].

Compared to iOS, Android provides greater visibility into the behavior of apps and the operating system, as would facilitating app instrumentation through a developer API. We opted for feature parity with iOS for this paper in order to evaluate a method that works for both platforms, but plan to leverage such additional data in later versions of the app (and already do so on the backend).

As with any passive approach, which a regulation iOS app must be, our results are limited by the data. If none of the clients ever runs a particular buggy app, Carat will never detect a problem; if two apps are always run together and one is anomalous, they will both be categorized as anomalies and there is nothing that correlation can do to disambiguate. The likelihood of spurious correlations increases with the number of features (apps and configurations). The way to combat this problem is with more data. For example, as we gather more samples involving highly correlated apps that show one but not the other, we can begin to discern which (or possibly both) are responsible for the anomaly. The results show that our data are sufficient for actionable diagnosis.

Carat is targeted at users, but additional in-app instrumentation (such as via a developer API) would enable finer-grained diagnoses for developers, e.g., identifying what user behaviors, app settings, or other environmental conditions trigger abnormal energy use.

7 Related Work

There is a rich body of work in diagnosis for correctness and performance. Recent work identified an emerging class of software misbehavior that afflicts battery life [31] and proposed a method for detecting a specific class of such bugs [33]. We believe our work is the first collaborative method to automatically detect and diagnose abnormal energy use on mobile devices. Unlike previous work, Carat is able to disambiguate between hogs and bugs—anomalies that are intrinsic to an app versus those that may be triggered by device- or user-specific conditions, respectively—a capability that requires measurements from multiple devices. An early prototype and small deployment of the method on a single platform was summarized in our workshop paper [27].

Our approach is a form of statistical debugging, in which (loosely speaking) deviant behavior is called a bug [9]. Such methods have been used to identify code paths correlated with failure [16, 17], concurrency bugs [14], shared influence (surprising behavior that is correlated in time) [26, 28], invariant violation [13], and configuration errors [41]. In the field of security, anomaly-based intrusion detection has a long history [8, 34, 35]. Recently, statistical methods were used to diagnose energy problems by comparing the behavior of an app at different times on a single device [21]; this kind of approach cannot disambiguate hogs from bugs or separate app-intrinsic behavior (many apps consume different amounts of energy depending on what features are being exercised) from device- or user-specific factors.

These statistical methods frequently make use of a large number of instances or users of these programs, which is sometimes called a *community*. A recent paper suggests a collaborative debugging framework called MobiBug for mobile devices [1], but they focus on crashes, not continuous or intermittent measurements. There is prior work for file systems [42] and peer-to-peer networks [22] that generate alerts based on aggregate behavior.

Projects like the Application Communities project [20] use the community to distribute work; instead, we employ uniform, lightweight instrumentation. There are also security applications for the community besides detection, such as diagnosing problems by discovering root causes [41] and preventing known exploits (e.g., sharing antibodies) [7, 25].

Many projects have sought to profile or emulate energy use on mobile devices [10, 11, 23, 29, 30, 32, 44], sometimes for prediction [37, 40], mitigation [3, 18], diagnosis [21], or developer tools [15]. Human interface studies have shown that 80% of mobile users will take steps to improve their battery life [36]; Carat recommends specific, personalized actions for users to take and even estimates the benefit they are likely to see. This is a distinguishing feature of our work.

Energy debugging shares similarities with performance debugging; both areas aim to account for the use or abuse of a shared resource. Some notable performance debugging work includes history-based analysis in datacenters [5], resource accounting [4], and black-box debugging [2]

Pinpoint [6] and Magpie [4] track communication dependencies with the aim of isolating the root cause of misbehavior; they require instrumentation of the application to tag client requests. In order to determine the causal relationships among messages, Project5 [2] and WAP5 [38] use message traces and compute dependency paths. D^3S [19] uses binary instrumentation to perform online predicate checks. Recent work shows how access to source code can facilitate tasks like log analysis [43] and distributed diagnosis [12]. CarrierIQ[17] collects detailed measurements by integrating with the mobile platform, and has drawn criticism for the intrusiveness of their implementation[18]. Unlike the preceding methods, we do not assume such access to code, communications, or binaries, taking instead a black-box approach with broader deployment potential.

8 Conclusions

This paper presents a method for diagnosing energy anomalies in the wild given incomplete and noisy instrumentation measurements from a community of clients. We implemented this method as an app for iOS and Android called Carat and deployed it to a community of more than 500,000 devices. Carat diagnosed thousands of anomalies, which involves detecting the anomaly, estimating its severity, quantifying the error and confidence bounds on that estimate, and sometimes identifying the device features that are correlated with the anomaly. We also validated our implementation with hardware measurements and synthetic anomaly injection, showing that Carat can accurately estimate energy use and detect anomalies.

Specifically, Carat imposes negligible overhead on each device, estimates energy use with accuracy comparable to hardware, detected 100% of synthetically injected anomalies in controlled experiments, produced no known false positives (based on corroborating dozens of anomalies using other methods), and predicted the battery impact of anomalies with greater than 95% accuracy. Finally, users receiving reports from Carat improved their battery life by 21% after a month; users who received no reports gained only 5.5% over the same period.

A collaborative approach is required to diagnose energy bugs; even complete knowledge of app behavior on a single client could be specific to a device or user. We believe this is the first collaborative diagnosis of energy anomalies in the wild and represents a crucial extension of previous work in distributed and statistical debugging to include a new class of abnormal behavior related to mobile energy use.

Notes

[1] `http://carat.cs.berkeley.edu`
[2] `http://msoon.com/LabEquipment/PowerMonitor/`
[3] `http://www.leydenenergy.com/`
[4] `http://techcrunch.com/2012/06/14/carat-battery/`
[5] `http://bit.ly/yTIUeU`
[6] `http://bit.ly/wsMraK`
[7] `http://bit.ly/WZ4dQi`
[8] `http://bit.ly/QSiv72`
[9] `com.bobisoft.wallpaper.beachatnight`
[10] `com.custom.lwp.FREE_HeartAndLove`
[11] `http://bit.ly/QSixvT`
[12] `http://bit.ly/TLWRhV`
[13] `http://bit.ly/Scgjs2`
[14] `http://gdg.to/xeK9CZ`
[15] `http://bit.ly/ODNyxQ`
[16] `http://zd.net/y0dyCr`
[17] `http://www.carrieriq.com/`
[18] `http://onforb.es/zd1zmF`

Acknowledgements

We thank the folks at Monsoon for helping us connect their hardware to the iPhone, the beta testers and `/r/compsci` community of reddit for trying early versions of the app, and our colleagues in the AMP Lab for their valuable feedback. We are also grateful to our numerous reviewers for their comments and to our shepherd, Geoffrey Challen, for helping us improve the paper.

This research is supported in part by NSF CISE Expeditions award CCF-1139158, gifts from Amazon Web Services, Google, SAP, Blue Goji, Cisco, Cloudera, Ericsson, General Electric, Hewlett Packard, Huawei, Intel, Microsoft, NetApp, Oracle, Quanta, Splunk, VMware and by DARPA (contract #FA8650-11-C-7136).

9 References

[1] S. Agarwal, R. Mahajan, A. Zheng, and V. Bahl. There's an app for that, but it doesn't work. Diagnosing mobile applications in the wild. In *HotNets*, 2010.

[2] M. K. Aguilera, J. C. Mogul, J. L. Wiener, P. Reynolds, and A. Methitacharoen. Performance debugging for distributed systems of black boxes. In *SOSP*, 2003.

[3] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: A

measurement study and implications for network applications. In *IMC*, 2009.

[4] P. Barham, A. Donnelly, R. Isaacs, and R. Mortier. Using Magpie for request extraction and workload modelling. In *OSDI*, 2004.

[5] P. Bodik, M. Goldszmidt, A. Fox, D. Woodard, and H. Andersen. Fingerprinting the datacenter: Automated classification of performance crises. In *Eurosys*, 2010.

[6] M. Y. Chen, E. Kiciman, E. Fratkin, A. Fox, and E. Brewer. Pinpoint: problem determination in large, dynamic internet services. In *DSN*, June 2002.

[7] M. Costa, J. Crowcroft, M. Castro, A. Rowstron, L. Zhou, L. Zhang, and P. Barham. Vigilante: End-to-end containment of internet worms. In *SOSP*, 2005.

[8] H. Debar, M. Becker, and D. Siboni. A neural network component for an intrusion detection system. In *IEEE Symposium on Security and Privacy*, 1992.

[9] D. Engler, D. Y. Chen, S. Hallem, A. Chou, and B. Chelf. Bugs as deviant behavior: a general approach to inferring errors in systems code. In *SOSP*, 2001.

[10] D. Ferreira, A. K. Dey, and V. Kostakos. Understanding human-smartphone concerns: A study of battery life. In *Pervasive*, 2011.

[11] J. Flinn and M. Satyanarayanan. PowerScope: a tool for profiling the energy usage of mobile applications. In *WMCSA*, 1999.

[12] K. Glerum, K. Kinshumann, S. Greenberg, G. Aul, V. Orgovan, G. Nichols, D. Grant, G. Loihle, and G. Hunt. Debugging in the (very) large: Ten years of implementation and experience. In *SOSP*, 2009.

[13] S. Hangal and M. Lam. Tracking down software bugs using automatic anomaly detection. In *ICSE*, 2002.

[14] G. Jin, A. Thakur, B. Liblit, and S. Lu. Instrumentation and sampling strategies for cooperative concurrency bug isolation. In *OOPSLA*, 2010.

[15] A. Kansal and F. Zhao. Fine-grained energy profiling for power-aware application design. In *HotMetrics*, 2008.

[16] B. Liblit, A. Aiken, A. X. Zheng, and M. I. Jordan. Bug isolation via remote program sampling. In *PLDI*, 2003.

[17] B. Liblit, M. Naik, A. X. Zheng, A. Aiken, and M. I. Jordan. Scalable statistical bug isolation. In *PLDI*, 2005.

[18] K. Lin, A. Kansal, D. Lymberopoulos, and F. Zhao. Energy-accuracy trade-off for continuous mobile device location. In *MobiSys*, 2010.

[19] X. Liu, Z. Guo, X. Wang, F. Chen, X. Lian, J. Tang, M. Wu, M. F. Kaashoek, and Z. Zhang. D3S: debugging deployed distributed systems. In *NSDI*, 2008.

[20] M. E. Locasto, S. Sidiroglou, and A. D. Keromytis. Software self-healing using collaborative application communities. In *NDSS*, 2005.

[21] X. Ma, P. Huang, X. Jin, P. Wang, S. Park, D. Shen, Y. Zhou, L. K. Saul, and G. M. Voelker. eDoctor: Automatically diagnosing abnormal battery drain issues on smartphones. In *NSDI*, 2013.

[22] D. J. Malan and M. D. Smith. Host-based detection of worms through peer-to-peer cooperation. In *ACM Workshop on Rapid Malcode*, 2005.

[23] R. Mittal, A. Kansal, and R. Chandra. Empowering developers to estimate app energy consumption. In *Mobicom*, 2012.

[24] S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 1997.

[25] J. Newsome, D. Brumley, and D. Song. Vulnerability-specific execution filtering for exploit prevention on commodity software. In *NDSS*, 2006.

[26] A. J. Oliner and A. Aiken. Online detection of multi-component interactions in production systems. In *DSN*, 2011.

[27] A. J. Oliner, A. P. Iyer, E. Lagerspetz, S. Tarkoma, and I. Stoica. Collaborative energy debugging for mobile devices. In *HotDep*, 2012.

[28] A. J. Oliner, A. V. Kulkarni, and A. Aiken. Using correlated surprise to infer shared influence. In *DSN*, 2010.

[29] E. Oliver. The challenges in large-scale smartphone user studies. In *HotPlanet*, 2010.

[30] E. A. Oliver and S. Keshav. An empirical approach to smartphone energy level prediction. In *UbiComp*, 2011.

[31] A. Pathak, Y. C. Hu, and M. Zhang. Bootstrapping energy debugging on smartphones: A first look at energy bugs in mobile devices. In *HotNets*, 2011.

[32] A. Pathak, Y. C. Hu, and M. Zhang. Where is the energy spent inside my app? Fine grained energy accounting on smartphones with eprof. In *EuroSys*, 2012.

[33] A. Pathak, A. Jindal, Y. C. Hu, and S. Midkiff. What is keeping my phone awake? Characterizing and detecting no-sleep energy bugs in smartphone apps. In *Mobisys*, 2012.

[34] V. Paxson. Bro: a system for detecting network intruders in real-time. In *Computer Networks*, December 1999.

[35] P. A. Porras and P. G. Neumann. EMERALD: event monitoring enabling responses to anomalous live disturbances. In *NIST/NCSC*, 1997.

[36] A. Rahmati, A. Qian, and L. Zhong. Understanding human-battery interaction on mobile phones. In *MobileHCI*, 2007.

[37] N. Ravi, J. Scott, L. Han, and L. Iftode. Context-aware battery management for mobile phones. In *PerCom*, 2008.

[38] P. Reynolds, J. L. Wiener, J. C. Mogul, M. K. Aguilera, and A. Vahdat. WAP5: black-box performance debugging for wide-area systems. In *WWW*, 2006.

[39] L. Rokach and O. Maimon. Top-down induction of decision trees classifiers—a survey. *Trans. on Sys., Man, and Cybernetics*, 2005.

[40] A. Sinha and A. P. Chandrakasan. JouleTrack: a web based tool for software energy profiling. In *DAC*, 2001.

[41] H. J. Wang, J. C. Platt, Y. Chen, R. Zhang, and Y.-M. Wang. Automatic misconfiguration troubleshooting with PeerPressure. In *OSDI*, 2004.

[42] Y. Xie, H. Kim, D. O'Hallaron, M. Reiter, and H. Zhang. Seurat: a pointillist approach to anomaly detection. In *RAID*, 2004.

[43] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan. Detecting large-scale system problems by mining console logs. In *SOSP*, 2009.

[44] C. Yoon, D. Kim, W. Jung, C. Kang, and H. Cha. Appscope: Application energy metering framework for android smartphones using kernel activity monitoring. In *USENIX Technical*, 2012.

[45] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *HotCloud*, 2010.