Beiwe Background Information

We developed the Beiwe (bee-we) research platform to collect smartphone-based high-throughput digital phenotyping data. The fully configurable open-source platform supports collection of a range of social, behavioral, and cognitive data, including spatial trajectories (via GPS), physical activity patterns (via accelerometer and gyroscope), social networks and communication dynamics (via call and text logs), and voice samples (via microphone).

The platform consists of a front-end smartphone application for iOS (by Apple) and Android (by Google) devices and a back-end system, which supports a web-based study management portal for data processing and storage, based on Amazon Web Services (AWS) cloud computing infrastructure.

Data analysis is increasingly identified as the main bottleneck; our data analysis platform, Forest, makes sense of the data collected by Beiwe.

General Study Background

The data generated by increasingly sophisticated smartphone sensors and phone use patterns appear ideal for capturing various social and behavioral disease phenotypes. Given that the majority of the adult population in developed nations now owns and operates a smartphone, the act of measurement no longer needs to be confined to research laboratories but instead can be carried out in naturalistic settings in situ leveraging the actual real-world experiences of patients.

While smartphones can be harnessed to offer medicine a wealth of data on disease phenotypes, the majority of existing smartphone apps are not intended for biomedical research use and, as such, do not generate research-quality data. While several commercial platforms collect similar data streams as the Beiwe research platform, they rarely if ever allow investigators to access the collected raw data. Most offer only proprietary summaries of the data. This approach is problematic not only from the data analysis perspective, but it also makes it harder to replicate research. In a typical biomedical research setting, one first formulates the scientific question of interest, then determines what data are needed to address that question, and finally decides on a statistical approach needed to connect the collected data with the research question. This approach is incompatible with platforms that do not allow access to raw data.

Finally, while many apps are able to collect some behavioral data, without a research platform to support these data, results are difficult to analyze and reproduce. Because the Beiwe platform includes a flexible study portal, customizable app, scalable database, as well as an evolving suite of modeling and data analysis tools, researchers can use it for a diverse set of studies. Equally important, results can be re-analyzed and studies recreated and validated using the same data collection settings and the same data analysis tools as those in the original study, thus significantly enhancing the level of reproducibility and transparency in research.

Beiwe Configurability / Registration / Participation

Every aspect of data collection is fully customizable, including which sensors to sample, sampling frequency, addition of Gaussian noise to GPS location, use of Wi-Fi or cellular data for uploads, data upload frequency, and specification of surveys and their response options.

Study participants simply download the Beiwe application from the app store and enter three pieces of information: a system-generated 8-character user ID, a system-generated temporary password, and an IP address of the back-end server. If no active data is being collected in the study (i.e., no surveys), this is the only time the participant will interact with the application. However, most studies make use of occasional self-reports or EMA, and some use the audio diary feature to collect rich data on lived experiences.

Data Protection

All Beiwe data is encrypted while stored on the phone awaiting upload and while in transit, and are re-encrypted for storage on the study server.

During study registration, Beiwe provides the smartphone app with the public half of a 2048-bit RSA encryption key. With this key, the device can encrypt data, but only the server, which has the private key, can decrypt it. Thus, the Beiwe application cannot read its own temporarily stored data, and the study participant (or somebody else) cannot export the data. The RSA key is used to encrypt a symmetric Advanced Encryption Standard (AES) key for bulk encryption. These keys are generated as needed by the app and must be decrypted by the study server before data recovery. Data received by the cloud server is re-encrypted with the study master key and then stored.

Some of the data collected by Beiwe contain identifiers, such as phone numbers. The Beiwe app generates a unique cryptographic code, called a salt, during the Beiwe registration process, and then uses the salt to encrypt phone numbers and other similar identifiers. The salt never gets uploaded to the server and is known only to the phone for this purpose. Using the industry-standard SHA-256 (Secure Hash Algorithm) and PBKDF2 (Password-Based Key Derivation Function 2) algorithms, an identifier is transformed into an 88-character anonymized string that can then be used in data analysis.

How Beiwe Minimizes Risk

Beiwe was designed to collect large quantities of social and behavioral data, and was designed on the premise that identifying data, such as phone numbers, will be protected by permanently anonymizing them when collected and that, in addition, all data should be encrypted at all times. Beiwe uses a store-and-forward architecture for managing data, meaning that data are buffered on the device and as soon as a Wi-Fi connection is available, all data are securely transmitted in a HIPAA-compliant manner (details below). During study registration, the platform provides the smartphone with the public half of a 2048-bit RSA (Rivest-Shamir-Adleman) encryption key. With this key the device can encrypt data, but only the server, which has the private key, can decrypt it. Thus, the Beiwe app cannot read its own data, so even if a phone is lost or stolen, no information is compromised. The RSA key is then used to encrypt a symmetric Advanced Encryption Standard (AES) key for bulk encryption. These keys are generated as needed by the app; therefore they are not stored anywhere and must be decrypted by the study server before any data can be recovered. Our current set-up does not use local servers but instead relies on cloud computing from Amazon Web Services (AWS), where at present we use an Amazon EC2 instance as the study server and an Amazon S3 instance for data storage. In our current configuration, data received by the EC2 server are re-encrypted with a master key provided for the given study and then stored on an Amazon S3 instance, an industry-standard secure storage platform housed in guarded data centers.

The key security aspects of the Beiwe Research Platform include:

- Participants are identified with their unique 8-character Beiwe Participant IDs.
- Participants will login to the Beiwe smartphone application with their ID and password.
- All data collection is tied to the 8-character Beiwe Participant ID. No identifiers, like participant name or contact information, is either collected or stored on the phone or the server. Only research collaborators have access to the study specific master key, which is stored securely.
- All data is encrypted while in transit and while at rest. The application does not store unencrypted data on the participants' smartphones
- Audio recordings are encrypted once recording is complete.
- Indirect identifiers (such as phone numbers and MAC addresses) are permanently anonymized using industry recognized encryption techniques , which renders data unidentifiable.
- No identifiable data is stored on the mobile device. All identifiers are rendered innocuous using an encryption scheme whereby, every phone generates their own unique cryptographic code during the Beiwe registration process, and it then uses that code to encrypt the phone numbers and MAC addresses collected by Beiwe. The only data stream that may contain

identifiable information are the audio recordings which cannot be
• anonymized; however all data, including the audio recordings, are encrypted.

GPS Data

The Beiwe app can record the phone's GPS location in latitude and longitude, as well as the precision of that measure. The GPS is often accurate to within about 10-20 meters. It can be used to construct a map of where a participant traveled and when a participant was at different places, although it cannot identify the mode of travel. The rate of GPS sampling is customizable to each study, or GPS sampling can be disabled for a particular study if GPS data is not part of the research questions/goals.

Beiwe can be configured to anonymize GPS information by adding randomly generated noise to the GPS coordinates on the device.