SAND2014-16825D

*Exceptional service in the national interest*

Sandia National Laboratories

# iOS Clone Detection

Michael Bierma
UC Davis
mhbierma@ucdavis.edu

Nicholas Ward
UC Berkeley
kingsyphax@berkeley.edu

Kevin Wu
UIUC
kcwu2@illinois.edu

Yung Ryn Choe
Sandia National Labs
yrchoe@sandia.gov

## Objective

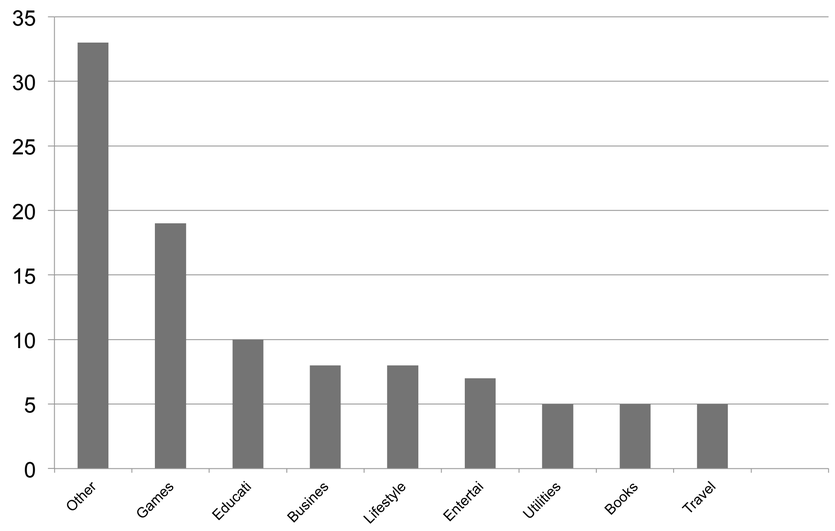
Perform large-scale static analysis of iOS applications to detect the presence of application clones using analysis techniques including normalized compression distance and opcode-level n-gram similarity analysis.

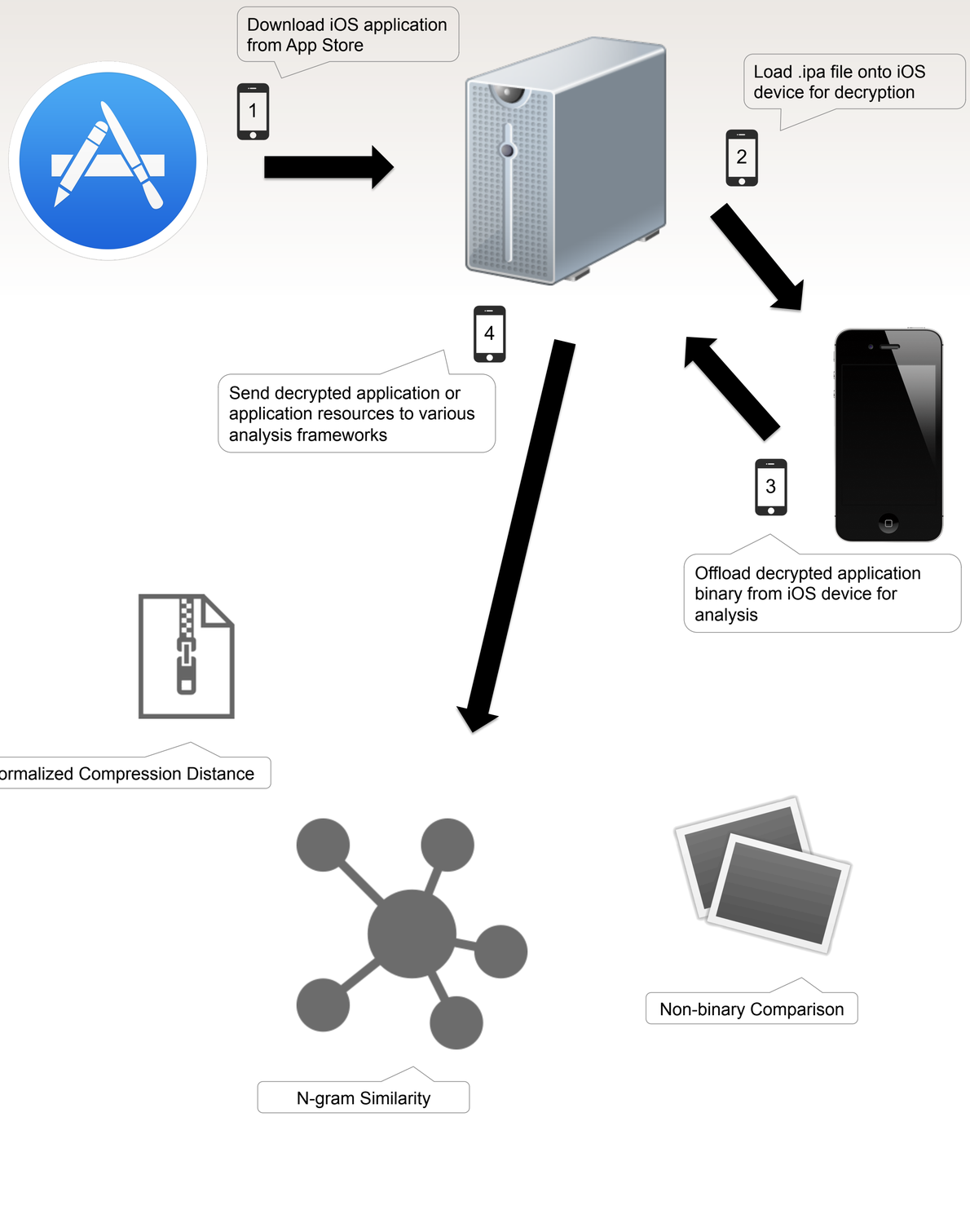
## Background

Apple's iOS platform is second only to Google's Android OS in mobile operating system market share, with over 400 million iOS devices sold. A signature feature of the iOS system is that it is closed, in terms of both source code and application distribution. This improves Apple's control over iOS apps, but it also hinders academic research related to the iOS platform. Compared to Android applications, iOS apps have been the subject of relatively little academic security research.

On the iOS platform, there have been numerous reports of malicious developers cloning popular applications in order to profit from the success of others. While clone detection has been investigated on the Android platform, this problem has not been rigorously analyzed with regard to iOS apps.

### Application Categories




### Normalized Compression Distance

If the compression of two application binaries together is similar to the size of the compression of the individual binary, the the apps are similar, suggesting cloning.

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

### N-gram Similarity

The binaries' assembly-level instructions are first split up into N-grams (groups of N consecutive instructions, where N is an integer). Fuzzy hashes are computed over combinations of N-grams, which are then used for comparison. Similarity of these fuzzy hashes indicates the apps have similar N-grams, and thus their code is similar.

### Non-binary Comparison

Applications include resources other than the application binary. Comparing some of the included resources, such as images, gives us another parameter for helping us to uncover cloned applications. We plan to utilize a variety of methods for image similarity analysis, including decision trees, perceptual hashes and keypoint matching.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2014-15758D