Automated Semi-supervised Authorship Attribution of Android Binaries

Hugo Gonzalez, Natalia Stakhanova, Ali A. Ghorbani
Faculty of Computer Science, University of New Brunswick

The Problem
- Malware on the Android platform has increased exponentially. Growing more than 600% in the last two years.
- New threats appear each day in the form of botnet, ransomware, adware or even worse: all together
- Typical methods to perform malware analysis focus on binaries it self, creating signatures or behavioral profiles.

Proposed solution
- It is our believe that only a reduced set of skilled authors produce primary strands of the malware, meanwhile the available samples are repackaged, reused, recompiled or evolved versions from the main strand.
- Our proposed approach looks for similarities in Android binary code from authors perspective.
- It involves 3 step process:
  - Creation and evaluation of benchmark profiles
  - Creation and evaluation of incremental profiles
  - Emergent behavior layer, where a large number of apps are analyzed.

Contributions
- Android Authors Dataset
- Method to classify and cluster on top of Random Forest algorithm
- Approach to create label profiles
- Large scale evaluation

Future work
- Evaluate complementary features
- Improve random forest performance using distributed task
- Use small number of files to create possible profiles.
- Web service to analyze apps.