ON AUTOMATICALLY DETECTING SIMILAR ANDROID APPS

Mario Linares-Vásquez
Andrew Holtzhauer
Denys Poshyvanyk
You Own a Start-Up

In two weeks, you are meeting with this venture capitalist.

He might invest >$1 million in you.
What Should You Do?

Your Product

A new web-enabled device is on the market.

You can build the first and only web browser for it.

What do we do?
What should you do?

Requirements

Design

Implementation

Verification

Maintenance
What should you do?

Agile

Strategize

Design

Build

Test

Software
Study Existing Solutions

Feature 1
“browser tabs”

Feature 2
“search bar”

Feature 3
“image downloading”
Importance of Detecting Similar Apps

Plagiarism/clones  Salient features in successful apps  API usages

*Images from http://howtowebdesign.org/excellent-android-logo-showcase/*
Detecting Similar Apps is Challenging

*Image from http://trendblog.net/the-story-of-the-android-logo/*
1.9M apps - Google Play

50B downloads - Google Play

23 releases (Android) since 2008
Similar to Angry Birds

- Angry Birds Rio
  Rovio Entertainment Ltd
  ★★★★★ FREE

- Subway Surfers
  Kiloo
  ★★★★★ FREE

- Fruit Ninja Free
  Halfbrick Studios
  ★★★★★ FREE

- Angry Birds Star Wars
  Rovio Entertainment Ltd
  ★★★★★ FREE

- Zombie Tsunami
  Mobigame S.A.R.L.
  ★★★★★ FREE

- Angry Birds Space
  Rovio Entertainment Ltd
  ★★★★★ FREE

- Transformers
  Rovio Entertainment Ltd
  ★★★★★ FREE

- Lep's World 2 🍎медведь
  nanobyte GmbH
  ★★★★★ FREE

- Piano Tiles 2 (Don't Tap the White Tile)
  Clean Master Games
  ★★★★★ FREE

- Jetpack Joyride
  Halfbrick Studios
  ★★★★★ FREE

- Hungry Shark Evolution
  Future Games of London
  ★★★★★ FREE

- Sonic Dash
  SEGA
  ★★★★★ FREE

*Screenshots from http://play.google.com
JOIN THE ANGRY BIRDS IN THEIR BIGGEST ADVENTURE YET!
A long time ago in a galaxy far, far away... a group of desperate rebel birds faced off against a galactic menace: the Empire's evil Pigtroopers!

Rebel birds, striking from a hidden base, have won their first victory against the evil Imperial Pigs. During the battle, Rebel spies managed to steal secret plans to the Empire's ultimate weapon, the PIG STAR, and are racing to deliver the plans to the Rebel Birds. Now they need your help!

Join an epic adventure with the Angry Birds in the legendary Star Wars™ universe! Use the Force, wield your lightsaber, and blast away Pigtroopers on an intergalactic journey from the deserts of Tatooine to the depths of the Pig Star -- where you'll face off against the terrifying Darth Vader, Dark Lord of the Pigs! Can you become a Jedi Master and restore freedom to the galaxy?

Time to grab your lightsaber and join the adventure! May the birds be with you!

HOURS AND HOURS OF ENGAGING GAMEPLAY - Explore more than 200 levels in iconic locations like Tatooine and the Pig Star. Can you dodge Imperial pigs, laser turrets, Tusken Raider pigs, and the dark side of the Force to get all three stars?

NEW GAMEPLAY MECHANICS - Use lightsabers, Blasters and Jedi powers to wreak havoc on the Imperial Pigs!

LEVEL UP YOUR BIRDS - Keep playing and level up your birds to improve their skills!

SECRETS AND HIDDEN GOODIES - Can you unlock all the R2-D2 and C-3PO bonus levels?

THE MIGHTY FALCON
Stuck on a tricky level? Earn stars and call the Mighty Falcon to rain down the destruction. New goals, achievements and gameplay!

*Screenshots from http://play.google.com*
CLANDroid: Key Ideas

Similar apps *share* some *semantic anchors* (e.g., API calls)

Requirements or features are implemented by a *combination of different semantic anchors*

Some semantic anchors are *more expressive* in terms of describing *silent features*
Semantic Anchors

<table>
<thead>
<tr>
<th>APIs</th>
<th>JAVA / ANDROID APPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td></td>
</tr>
<tr>
<td>Permissions</td>
<td>ANDROID SPECIFIC</td>
</tr>
<tr>
<td>Intents</td>
<td></td>
</tr>
<tr>
<td>Sensors</td>
<td></td>
</tr>
</tbody>
</table>
Permissions

Access to system resources is controlled with install time permissions
Permissions

Required for:

- APIs/Features (e.g., internet connection)
- Content providers (e.g., READ_CONTENTS)
- Databases
- Message-passing system
<uses-permission android:name="com.android.vending.CHECK_LICENSE" /> 
<uses-permission android:name="android.permissionINTERNET" />
<uses-permission android:name="android.permission.WAKE_LOCK" />
<uses-permission android:name="android.permission.ACCESS_NETWORK_STATE" />
<uses-permission android:name="android.permission.ACCESS_WIFI_STATE" />
<uses-permission android:name="android.permission.WRITE_EXTERNAL_STORAGE" />
**Intents**: asynchronous messages that describe operations to perform

Intents

MainActivity ➔ Launch ➔ OtherActivity

Explicit intent

Intent intent = new Intent(MainActivity.this, OtherActivity.class);
intent.putExtra("vo", vo);
intent.putExtra("services_quantity", quantity);
....
startActivity(intent);  Asynchronous call
Sensors

https://commons.wikimedia.org/wiki/File:Mobile_Device_Sensors.png
## Sensors

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Android 4.0 (API Level 14)</th>
<th>Android 2.3 (API Level 9)</th>
<th>Android 2.2 (API Level 8)</th>
<th>Android 1.5 (API Level 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYPE_ACCELEROMETER</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TYPE_AMBIENT_TEMPERATURE</td>
<td>Yes</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>TYPE_GRAVITY</td>
<td>Yes</td>
<td>Yes</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>TYPE_GYROSCOPE</td>
<td>Yes</td>
<td>Yes</td>
<td>n/a¹</td>
<td>n/a¹</td>
</tr>
<tr>
<td>TYPE_LIGHT</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TYPE_LINEAR_ACCELERATION</td>
<td>Yes</td>
<td>Yes</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>TYPE_MAGNETIC_FIELD</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TYPE_ORIENTATION</td>
<td>Yes²</td>
<td>Yes²</td>
<td>Yes²</td>
<td>Yes</td>
</tr>
<tr>
<td>TYPE_PRESSURE</td>
<td>Yes</td>
<td>Yes</td>
<td>n/a¹</td>
<td>n/a¹</td>
</tr>
<tr>
<td>TYPE_PROXIMITY</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TYPE_RELATIVE_HUMIDITY</td>
<td>Yes</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>TYPE_ROTATION_VECTOR</td>
<td>Yes</td>
<td>Yes</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>TYPE_TEMPERATURE</td>
<td>Yes²</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

[1] n/a: Not available
[2] The type orientation sensor is available in Android 4.0 (API Level 14) and above.
Challenges

Different programmers can use the same words to describe different requirements (i.e., the synonymy problem)

Different programmers can use different words to describe the same requirements (i.e., the polysemy problem)
LSI reveals latent concepts
The JDK contains close to 115,000 API calls that are exported by a little more than 13,000 classes and interfaces that are contained in 721 packages

Similarities among terms and words in LSI
Singular Value Decomposition (SVD)
Latent Semantic Indexing

document term
\[ m \times n \]

term dims
\[ m \times r \]

dims r x r

document dims
\[ r \times n \]
LSI and Semantic Anchors

TDM (APIs) → LSI (APIs)

TDM (Source code) → LSI (Identifiers)

TDM (Permissions) → LSI (Permissions)

TDM (Intents) → LSI (Intents)

TDM (Sensors) → LSI (Sensors)

LATENT CONCEPTS
Data Extraction

1. Google Play → APK files
2. APK files → APKs
3. APKs → apktool
4. apktool → dex2jar
5. dex2jar → JARs
6. Permissions extractor → manifest files
7. Permissions extractor → JAR files
8. JAD → .class files
9. JAD → 7zip
10. 7zip → .class files
11. sensors extractor → sensors
12. sensors extractor → JClassInfo
13. source code → No
14. source code → Error?
15. API calls → JClassInfo
EVALUATION
**RQ1:** What semantic anchors used in CLANDroid produce better results when compared to the others?

**RQ2:** How orthogonal are the apps detected by CLANDroid as compared to Google Play?

**RQ3:** Do third-party libraries and obfuscated apps impact the accuracy of CLANDroid?
RQ1: Survey

6 versions of the survey (one for each semantic anchor + combined)

12 Questions, 4 apps
RQ1: Survey

27 Participants

48 Apps

Intro.
Dear participant,

Thanks for your help and your time. In the following survey you will evaluate whether a set of mobile apps are similar to 12 target Android applications. In order to evaluate similarity of the applications you will use the following Likert scale:

1) Completely dissimilar – The participant is highly confident that the app is dissimilar to the target app.

2) Mostly dissimilar – It is unclear if the app is similar to the target app.

3) Mostly similar – There are some similarities between the app and the target app.

4) Highly similar – The participant is highly confident that the app is similar to the target app.

We encourage the user to read/check the information of the apps available at Google Play (i.e., description, permissions, screenshots, what’s new summary, etc). Please do not base your judgement using only the name of the app or the icon. We want you to judge similarities based on the features you think each app provides to the users.
RQ1: Survey

Intro.
Dear participant,

Thanks for your help and your time. In the following survey you will evaluate whether a set of mobile apps are similar to 12 target Android applications. In order to evaluate the similarity of the applications you will use the following Likert scale:

1) Completely dissimilar – The participant is highly confident that the app is dissimilar to the target app.

2) Mostly dissimilar – It is unclear if the app is similar to the target app.

3) Mostly similar – There are some similarities between the app and the target app.

4) Highly similar – The participant is highly confident that the app is similar to the target app.

We encourage the user to read/check the information of the apps available at Google Play (i.e., description, permissions, screenshots, what’s new summary, etc). Please do not base your judgement using only the name of the app or the icon. We want you to judge similarities based on the features you think each app provides to the users.
RQ1: Survey

Q1. Given the following app: air.BasketballDoodFree (description)

Please rate the similarity of the following set of apps to the app described above:

- App 1: air.SoccerDoodFree (description)
- App 2: air.RugbyDoodFree (description)
- App 3: air.FieldHockeyDoodFree (description)

<table>
<thead>
<tr>
<th></th>
<th>Completely dissimilar</th>
<th>Mostly dissimilar</th>
<th>Mostly similar</th>
<th>Highly Similar</th>
</tr>
</thead>
<tbody>
<tr>
<td>App 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>App 2</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>App 3</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
</tbody>
</table>
RQ1: Results

a) Similarity of Top–3 Similar Apps

- Sensors
- Permissions
- Intents
- Combined
- Identifiers

b) Similarity of Top–1 Similar App

- Sensors
- Permissions
- Intents
- Combined
- Identifiers

Table I

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>API</th>
<th>RQ1: SIMILAR APPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLANdroid</td>
<td>14.3</td>
<td>0.30571</td>
</tr>
<tr>
<td>Combined CLANdroid</td>
<td>1.5</td>
<td>0.30251</td>
</tr>
<tr>
<td>CLANdroid</td>
<td>2.5</td>
<td>0.20995</td>
</tr>
<tr>
<td>CLANdroid</td>
<td>3.0</td>
<td>0.20694</td>
</tr>
<tr>
<td>CLANdroid</td>
<td>3.5</td>
<td>0.10333</td>
</tr>
<tr>
<td>API</td>
<td>4.0</td>
<td>0.00076</td>
</tr>
</tbody>
</table>
Target App Example

Race, Stunt, Fight, Lite!
Adrenaline Crew  Racing
Unrated
Offers in-app purchases
This app is compatible with your device.

Add to Wishlist  Install

Similar

SouzaSim - Mop
SouzaSim is a wheelie Simulator, Yeah! thats right!
FREE

Race, Stunt, Fight
Adrenaline Crew
Race stunt and fight your way across the all new motorcycle racing game!
FREE

Traffic Rider
Soner Kara
The Next-Gen of Endless Motorbike Racing
FREE
Top-1 Similar App (CLANDroid)

Crazy Monster Truck - Escape
GAMEANAX  Racing
Everyone
Offers in-app purchases
This app is compatible with your device.

Add to Wishlist  Install

Roar like a daredevil with Natasha’s deadly super bike
Free Bonus Level
Top-2 Similar App (CLANDroid)

F18 Carrier Landing Lite

RORTOS Simulation

Everyone

Offers in-app purchases

This app is compatible with your device.
CLANSDroid is an effective tool for detecting similar apps, in particular when using APIs, identifiers, permissions, and intents as semantic anchors.

Using sensors for detecting similar apps appears to be ineffective.

APIs provided the highest number of apps rated as “highly similar”.
RQ2: Ranking of Goldset Apps

14K+ Target Rankings using 6 CLANDroid instances

Target App  Goldset Rankings of goldset apps
RQ2: Summary

Mann-Whitney tests show statistically significant differences in all the cases except for the comparison between Identifiers and Permissions, and Intents and Combined

Magnitude of the differences (Cliff’s delta) is small or negligible

CLANdroid is able to retrieve similar apps belonging to different categories, while Google Play lists as similar only apps in the same category
RQ3: Results - TPL

The rankings of goldset apps significantly improved

\(~490\) Positions
RQ3: Examples

CLANDroid API with and without TPL: 4,928 vs. 50
CLANDroid Ident with and without TPL: 6,852 vs. 266
RQ3: Summary

CLANDroid is significantly (negatively) impacted by the inclusion of \textit{third-party libraries}.

There are differences in the rankings when \textit{excluding} applications we detected as \textit{obfuscated}, however, the magnitude of these differences is \textit{negligible} in most cases.
CLANDroid (online)

http://www.semeru.info/clandroid/
Summary

Importance of detecting similar apps

Plagiarism/clones  Salient features in successful apps  API usages

CLANDroid

APKs → Data Extractor → TDM Builder

Users/Developers → Search Engine → LSI

RQ1: Survey

6 versions of the survey (one for each semantic anchor + combined)

12 Questions, 4 apps each

A survey → 1 + 3

Target app  Similar apps

RQ1: Results

RQ2: Results

RQ3: Results - TPL

Ranking of goldset apps was improved with significant differences

~490 Positions