Android Application Similarity in Five Minutes

Felix Matenaar
felix.matenaar@rwth-aachen.de
@pleed_

27.09.2012

DEXlabs.org

Introduction

Definition

Semantic similarity or semantic **relatedness** is a concept whereby a set of documents or terms within term lists are **assigned a metric** based on the likeness of their meaning / semantic content. - Wikipedia

Motivation

1. IP theft
2. Malware detection
3. Fun

Motivation

1. IP theft
2. Malware detection
3. Fun

→ We want to calculate the similarity between Android applications

Objects to deal with

Dex Object Format

- Strings
- Classes/Types
- Fields
- **Methods**
- Some other data

What is a Method?

0x00somerandombinarystring0x00

What is a Method?

0x00somerandombinarystring0x00
$\rightarrow$ We just don't care for now

Distance Metric

Edit/Levenshtein Distance

String 1: abcdefg
String 2: axbgg
Edits: Insertion, Deletion, Substitution

Edit/Levenshtein Distance

String 1: abcdefg
String 2: axbgg
Edits: Insertion, Deletion, Substitution

1. abgg $\rightarrow$ Deletion

Edit/Levenshtein Distance

String 1: abcdefg
String 2: axbgg
Edits: Insertion, Deletion, Substitution

1. abgg $\rightarrow$ Deletion
2. ab**c**gg $\rightarrow$ Insertion
3. abc**d**gg $\rightarrow$ Insertion
4. abcd**e**gg $\rightarrow$ Insertion

Edit/Levenshtein Distance

String 1: abcdefg
String 2: axbgg
Edits: Insertion, Deletion, Substitution

1. abgg $\rightarrow$ Deletion
2. ab**c**gg $\rightarrow$ Insertion
3. abc**d**gg $\rightarrow$ Insertion
4. abcd**e**gg $\rightarrow$ Insertion
5. abcde**f**g $\rightarrow$ Substitution

Edit/Levenshtein Distance

String 1: abcdefg
String 2: axbgg
Edits: Insertion, Deletion, Substitution

1. abgg $\rightarrow$ Deletion
2. ab**c**gg $\rightarrow$ Insertion
3. abc**d**gg $\rightarrow$ Insertion
4. abcd**e**gg $\rightarrow$ Insertion
5. abcde**f**g $\rightarrow$ Substitution

$\rightarrow$ Distance between String 1 and String 2 is 5 Edits

Distance Metric

Map Distance to similarity value to the interval $[0, 1]$

$$f_{sim}(s_1, s_2) = \frac{edits(s_1, s_2)}{max(|s_1|, |s_2|)}$$ f(abcdefg,axbgg) = 0.7,

$\rightarrow$ 0 would mean equality

Find the global Optimum

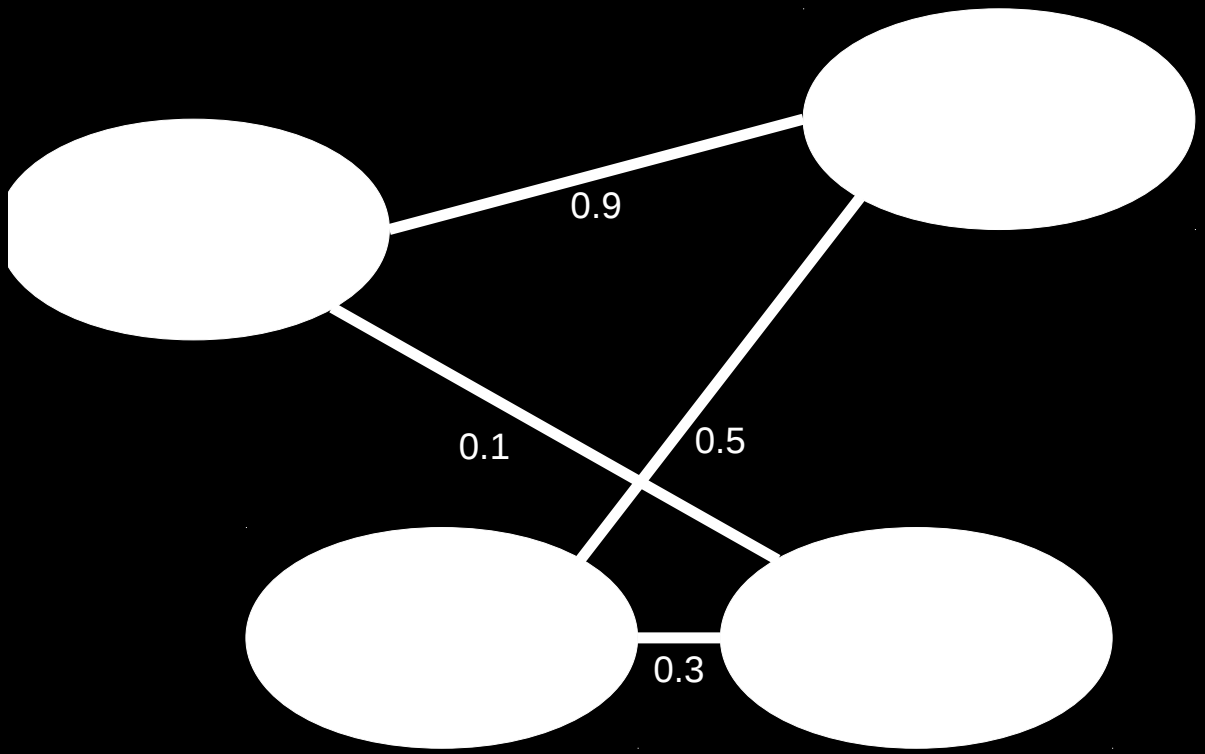
Situation

We got:

1. A set of methods from application 1 (a_1)
2. A set of methods from application 2 (a_2)
3. $f_{sim}(s1, s2) | s1 \in a_1, s2 \in a2$

Naive Approach

Decide locally after minimum cost candidate



Correct Solution^{TM}

- Use munkres algorithm to find global minimum in $O(n^3)$

Conclusion

Conclusion

We've learned

- *one* way to calculate the similarity of binary strings
- *one* way to find the global optimum between sets of methods

Contact

Felix Matenaar
felix.matenaar@rwth-aachen.de
@pleed_

www.dexlabs.org