SimChar: Building a Dataset of Visually Similar Characters

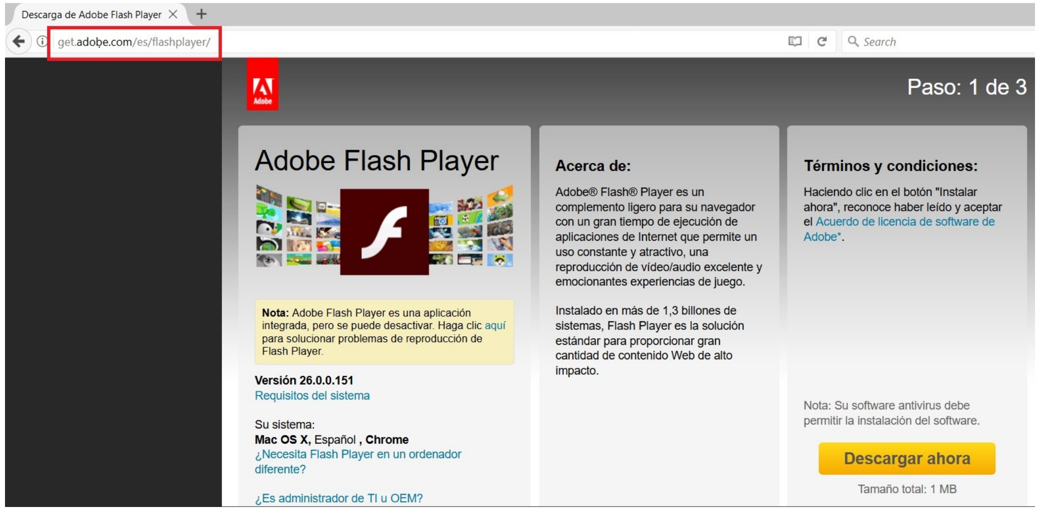
Sep 18 2019 @ W3C TPAC Break Out Session (AHA)

Tatsuya Mori
Waseda University

Background

IDN homograph attacks getting vital these days.

Aug 2017



Descarga de Adobe Flash Player

get.adoḅe.com/es/flashplayer/

Search

Paso: 1 de 3

Adobe Flash Player

Nota: Adobe Flash Player es una aplicación integrada, pero se puede desactivar. Haga clic aquí para solucionar problemas de reproducción de Flash Player.

Versión 26.0.0.151
Requisitos del sistema

Su sistema:
Mac OS X, Español **, Chrome**
¿Necesita Flash Player en un ordenador diferente?

¿Es administrador de TI u OEM?

Acerca de:

Adobe® Flash® Player es un complemento ligero para su navegador con un gran tiempo de ejecución de aplicaciones de Internet que permite un uso constante y atractivo, una reproducción de vídeo/audio excelente y emocionantes experiencias de juego.

Instalado en más de 1,3 billones de sistemas, Flash Player es la solución estándar para proporcionar gran cantidad de contenido Web de alto impacto.

Términos y condiciones:

Haciendo clic en el botón "Instalar ahora", reconoce haber leído y aceptar el Acuerdo de licencia de software de Adobe*.

Nota: Su software antivirus debe permitir la instalación del software.

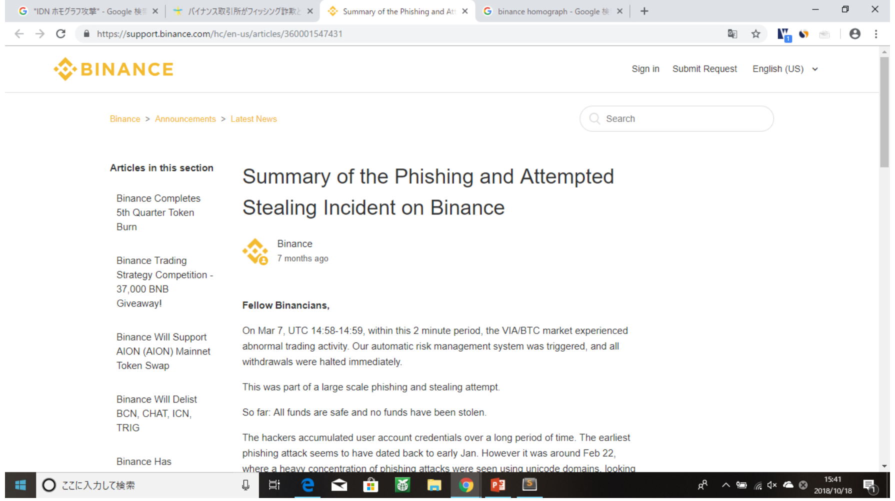
Descargar ahora

Tamaño total: 1 MB

adobe.com targeted (adoḅe.com)

Latin small letter b with dot below (U+1E05)

Mar 2018



"IDN ホモグラフ攻撃" - Google 検索 | バイナンス取引所がフィッシング詐欺と | Summary of the Phishing and Att | binance homograph - Google 検索

https://support.binance.com/hc/en-us/articles/360001547431

BINANCE — Sign in — Submit Request — English (US)

Binance > Announcements > Latest News

Search

Articles in this section

Binance Completes 5th Quarter Token Burn

Binance Trading Strategy Competition - 37,000 BNB Giveaway!

Binance Will Support AION (AION) Mainnet Token Swap

Binance Will Delist BCN, CHAT, ICN, TRIG

Binance Has

Summary of the Phishing and Attempted Stealing Incident on Binance

Binance
7 months ago

Fellow Binancians,

On Mar 7, UTC 14:58-14:59, within this 2 minute period, the VIA/BTC market experienced abnormal trading activity. Our automatic risk management system was triggered, and all withdrawals were halted immediately.

This was part of a large scale phishing and stealing attempt.

So far: All funds are safe and no funds have been stolen.

The hackers accumulated user account credentials over a long period of time. The earliest phishing attack seems to have dated back to early Jan. However it was around Feb 22, where a heavy concentration of phishing attacks were seen using unicode domains, looking

ここに入力して検索

15:41
2018/10/18

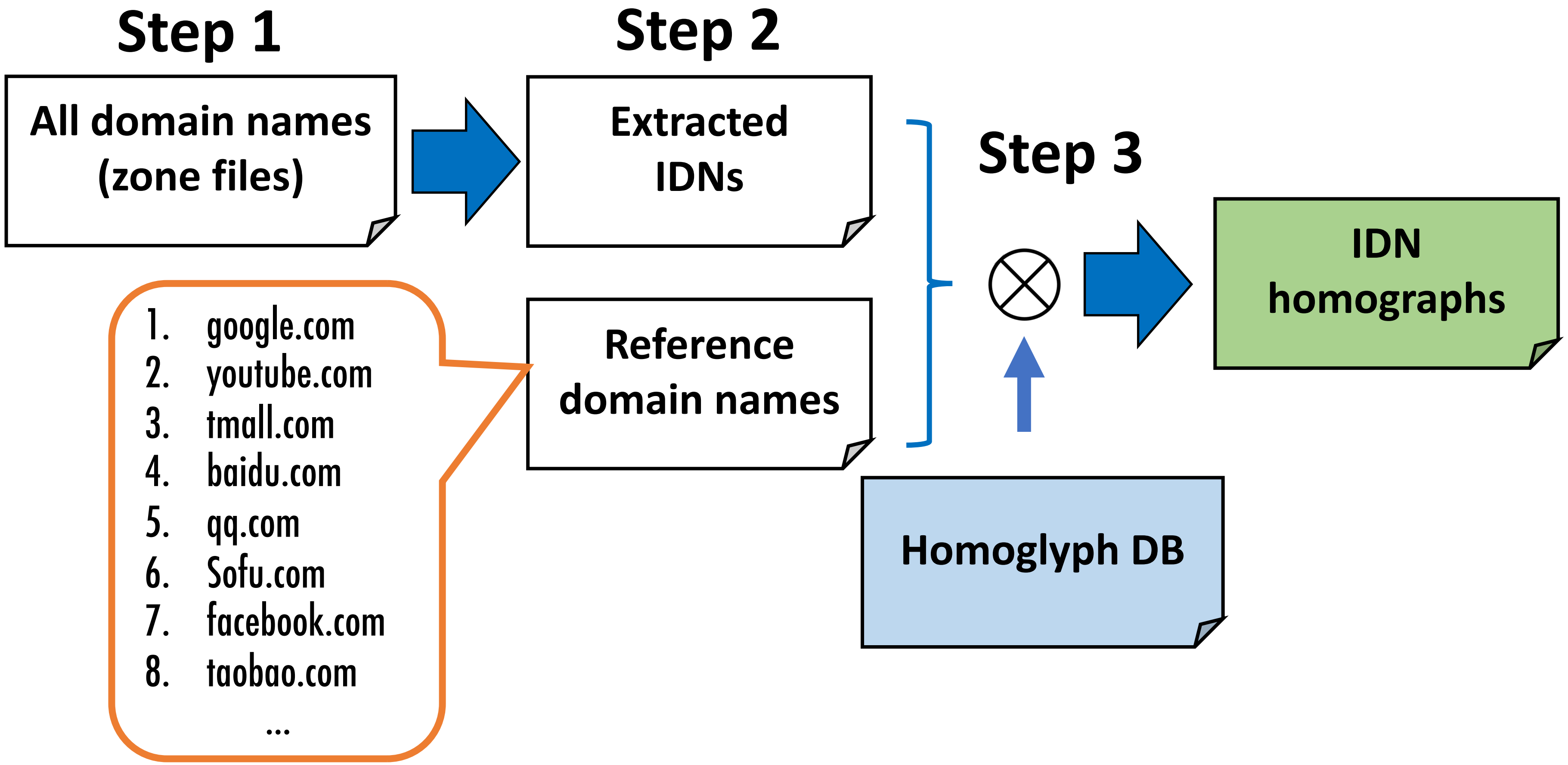
binance.com targeted (bịnạnce.com)

Latin small letter i with dot below (U+1ECB)

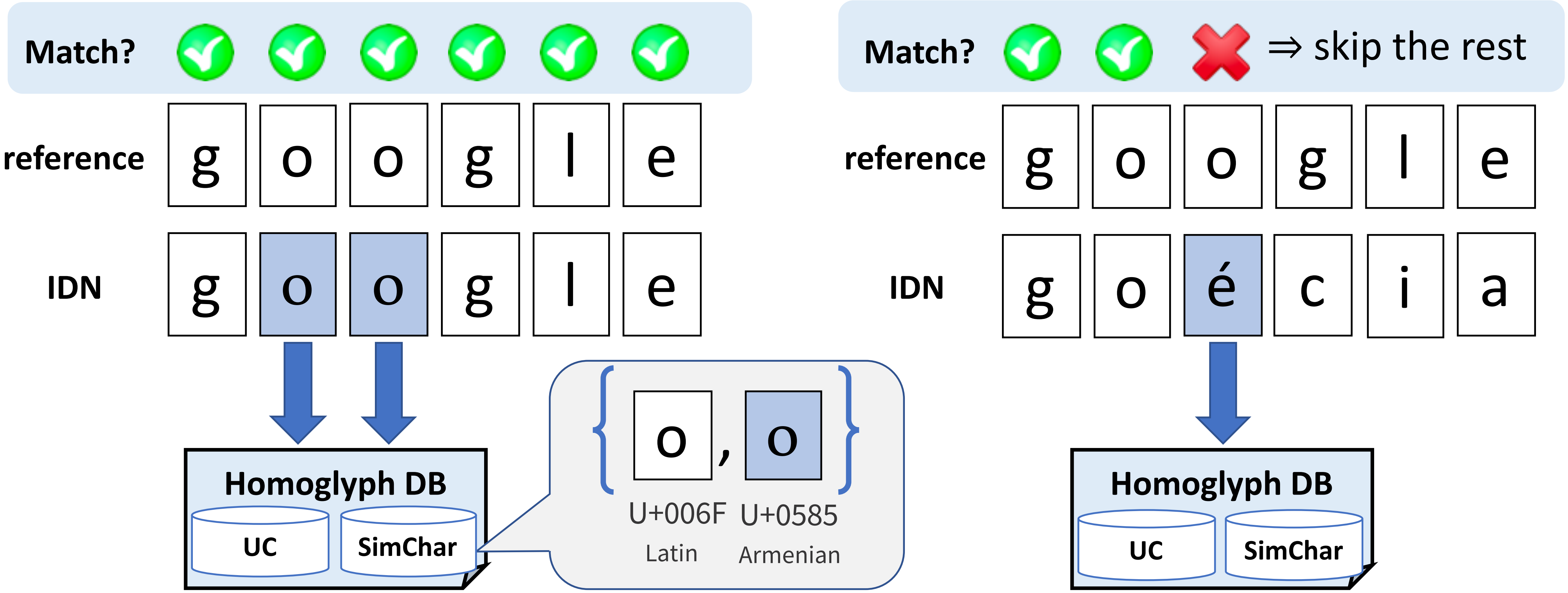
Latin small letter a with dot below (U+1EA1)

ShamFinder

- A framework we built to detect potential IDN homographs automatically.



ShamFinder



Match?	✓	✓	✓	✓	✓	✓
reference	g	o	o	g	l	e
IDN	g	o	o	g	l	e

Homoglyph DB

UC SimChar

{ o , o }

U+006F Latin

U+0585 Armenian

Match?	✓	✓	✗ ⇒ skip the rest			
reference	g	o	o	g	l	e
IDN	g	o	é	c	i	a

Homoglyph DB

UC SimChar

Question

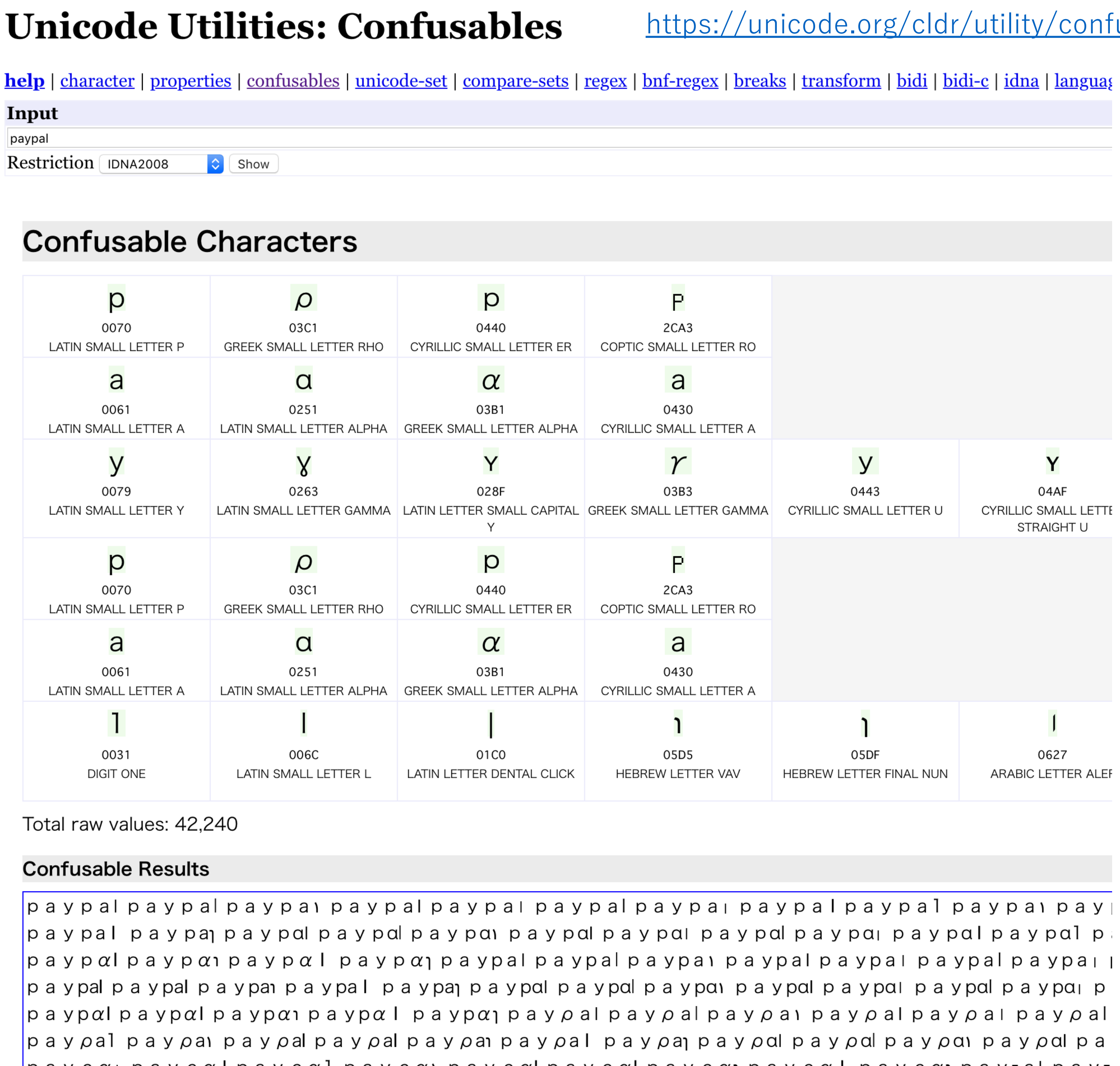
- Is there a homoglyph DB out there?

A solution

- Yes, we can make use of confusables.txt

http://unicode.org/reports/tr39/#Data_Collection

Visually Confusable Characters:
Provides a mapping for visual confusables for use in detecting possible security problems. The usage of the file is described in *Section 4, Confusable Detection*.



Unicode Utilities: Confusables

https://unicode.org/cldr/utility/confusables.jsp?a=paypal&r=IDNA2008

help | character | properties | confusables | unicode-set | compare-sets | regex | bnf-regex | breaks | transform | bidi | bidi-c | idna | languag

Input

paypal

Restriction IDNA2008 Show

Confusable Characters

p 0070 LATIN SMALL LETTER P	ρ 03C1 GREEK SMALL LETTER RHO	р 0440 CYRILLIC SMALL LETTER ER	ⲣ 2CA3 COPTIC SMALL LETTER RO		
a 0061 LATIN SMALL LETTER A	ɑ 0251 LATIN SMALL LETTER ALPHA	α 03B1 GREEK SMALL LETTER ALPHA	а 0430 CYRILLIC SMALL LETTER A		
y 0079 LATIN SMALL LETTER Y	ɣ 0263 LATIN SMALL LETTER GAMMA	ʏ 028F LATIN LETTER SMALL CAPITAL Y	γ 03B3 GREEK SMALL LETTER GAMMA	у 0443 CYRILLIC SMALL LETTER U	ү 04AF CYRILLIC SMALL LETTE STRAIGHT U
p 0070 LATIN SMALL LETTER P	ρ 03C1 GREEK SMALL LETTER RHO	р 0440 CYRILLIC SMALL LETTER ER	ⲣ 2CA3 COPTIC SMALL LETTER RO		
a 0061 LATIN SMALL LETTER A	ɑ 0251 LATIN SMALL LETTER ALPHA	α 03B1 GREEK SMALL LETTER ALPHA	а 0430 CYRILLIC SMALL LETTER A		
1 0031 DIGIT ONE	l 006C LATIN SMALL LETTER L	ǀ 01C0 LATIN LETTER DENTAL CLICK	ו 05D5 HEBREW LETTER VAV	ן 05DF HEBREW LETTER FINAL NUN	ا 0627 ARABIC LETTER ALEF

Total raw values: 42,240

Confusable Results

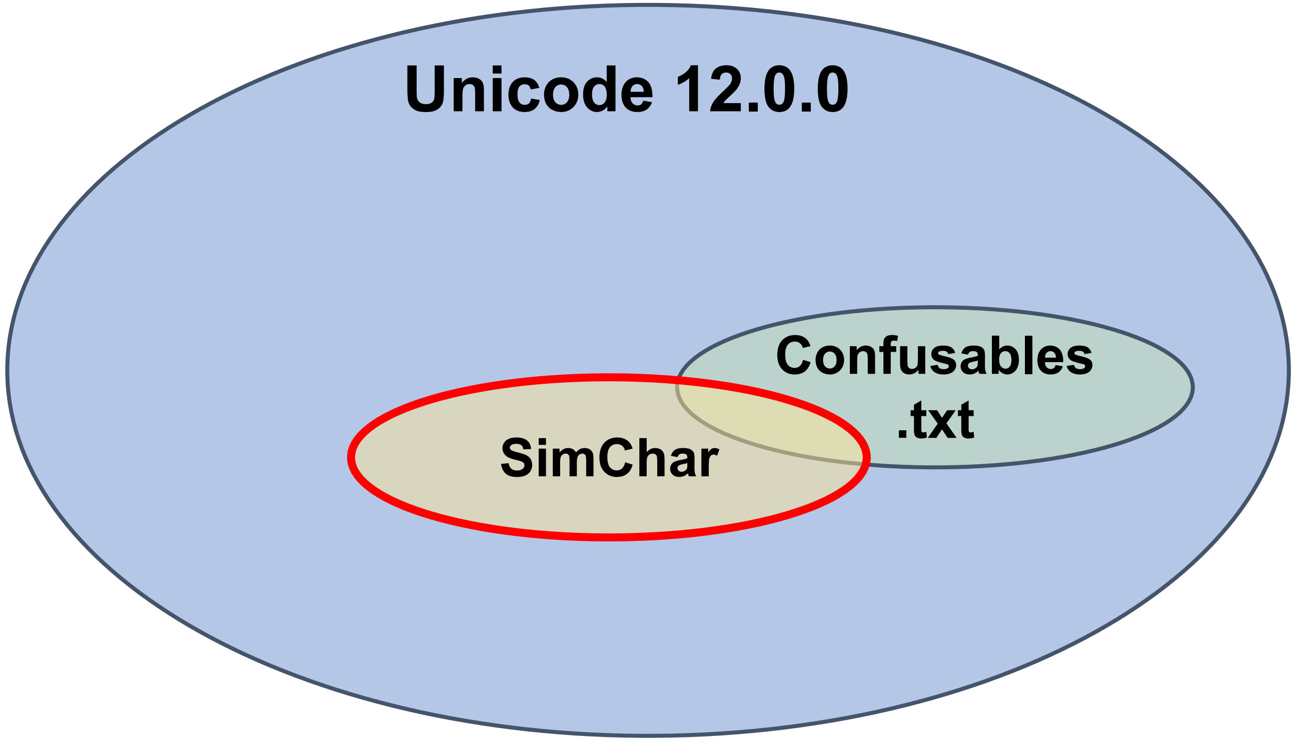
```
paypal paypal paypaו paypal paypaا paypal paypaו paypal paypaǀ paypa1 paypaו pay
paypal paypaן paypɑl paypɑl paypɑו paypɑl paypɑl paypɑا paypɑl paypɑا paypɑl paypɑ1 p
paypαl paypα1 paypαǀ paypαן paypal paypaו paypal paypaا paypal paypaا p
paypal paypal paypaו paypaǀ paypaן paypɑl paypɑl paypɑו paypɑl paypɑا paypɑl paypɑا p
paypαl paypαl paypαו paypαǀ paypαן payρal payρal payρaו payρal payρaا payρal
payρa1 payρaו payρal payρal payρaו payρaǀ payρaן payρɑl payρɑl payρɑו payρɑl pa
```

Our Question

- Are there homoglyphs that are NOT listed in the confusables.txt?

Answer

- Yes, we found several homoglyphs not listed in the confusables.txt



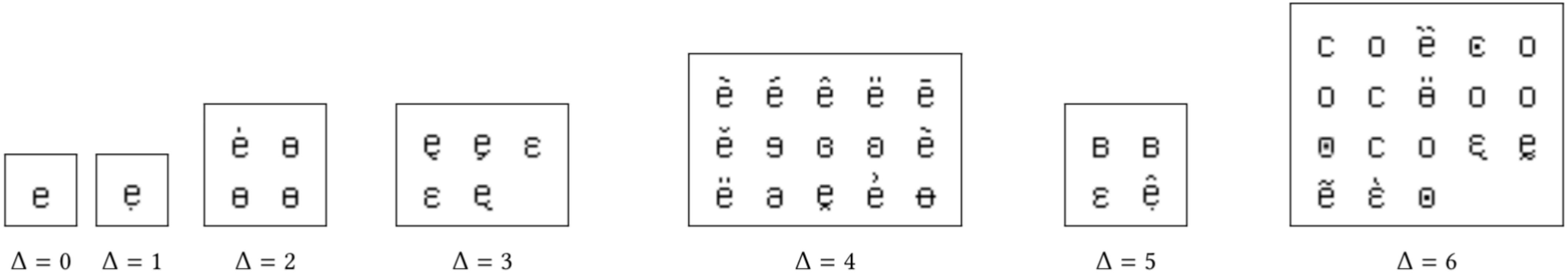
The process of building a homoglyph DB (SimChar)

1. Get the visual images of characters (Unicode BMP) by using GNU Unifont.

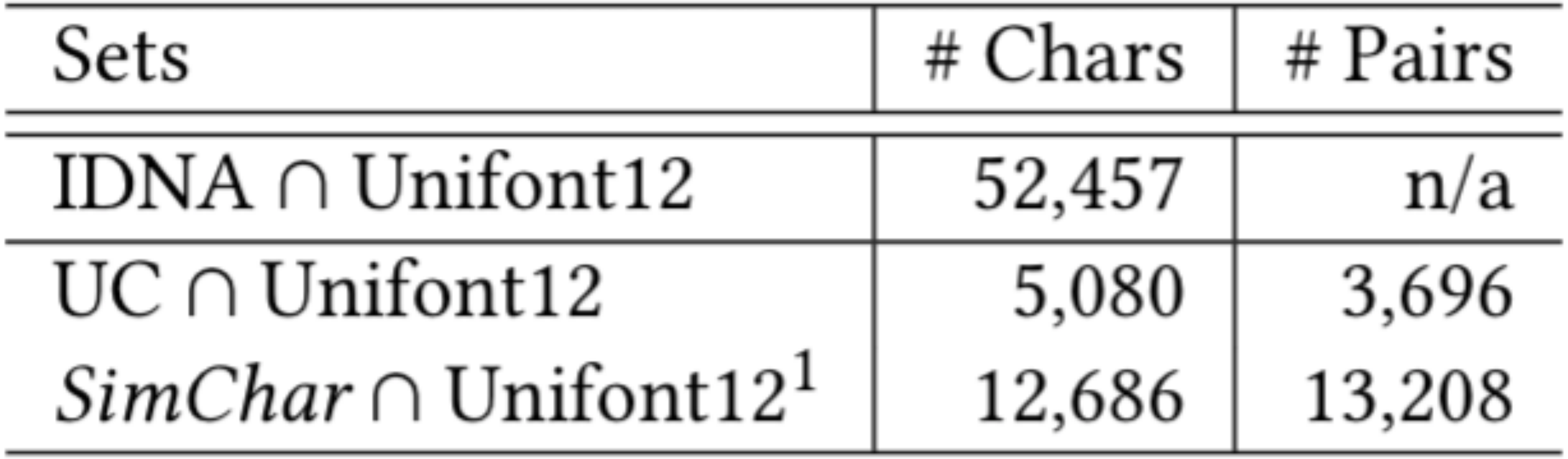
2. Compute the distance of two characters (images) with the number of different pixels.

3. If the distance is smaller than some threshold, then detect the pair as homoglyph.

The process of building a homoglyph DB (SimChar)



Stats of DBs



Sets	# Chars	# Pairs
IDNA ∩ Unifont12	52,457	n/a
UC ∩ Unifont12	5,080	3,696
SimChar ∩ Unifont12[1]	12,686	13,208

Confusables to Latin letters

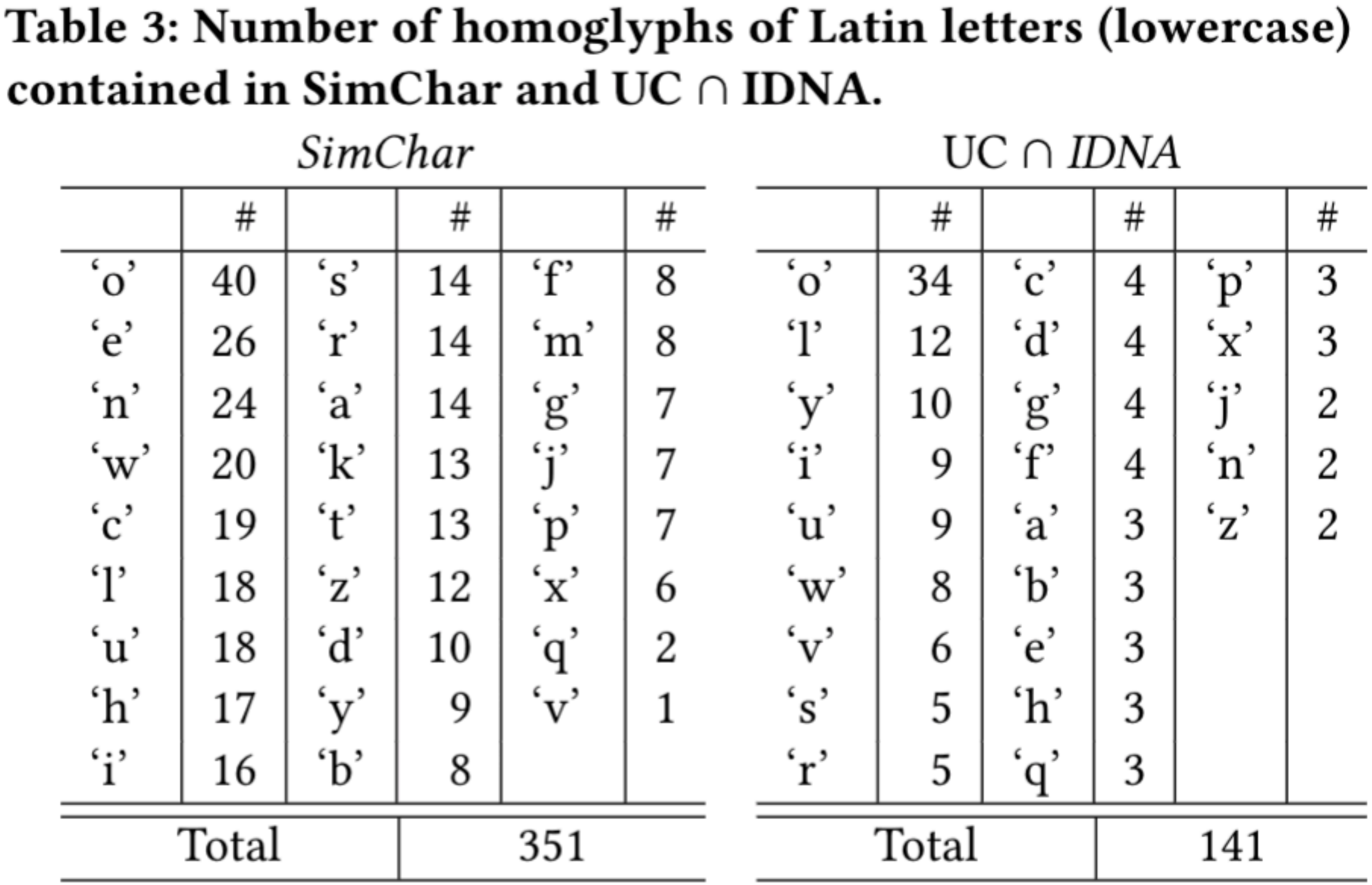


Table 3: Number of homoglyphs of Latin letters (lowercase) contained in SimChar and UC ∩ IDNA.

SimChar

	#		#		#
‘o’	40	‘s’	14	‘f’	8
‘e’	26	‘r’	14	‘m’	8
‘n’	24	‘a’	14	‘g’	7
‘w’	20	‘k’	13	‘j’	7
‘c’	19	‘t’	13	‘p’	7
‘l’	18	‘z’	12	‘x’	6
‘u’	18	‘d’	10	‘q’	2
‘h’	17	‘y’	9	‘v’	1
‘i’	16	‘b’	8		
Total			351		

UC ∩ *IDNA*

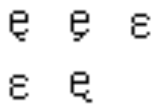
	#		#		#
‘o’	34	‘c’	4	‘p’	3
‘l’	12	‘d’	4	‘x’	3
‘y’	10	‘g’	4	‘j’	2
‘i’	9	‘f’	4	‘n’	2
‘u’	9	‘a’	3	‘z’	2
‘w’	8	‘b’	3		
‘v’	6	‘e’	3		
‘s’	5	‘h’	3		
‘r’	5	‘q’	3		
Total			141		

Examples

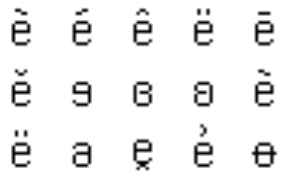
e_0

e_1

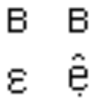
e_2



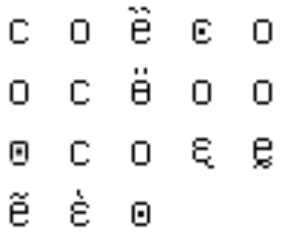
e_3



e_4



e_5



e_6

Our Question

- Are the detected homoglyphs really confusable?

Answer

- Yes, our human study revealed that they are more confusable than those contained in confusables.txt!

Q: Are they distinct or confusing?

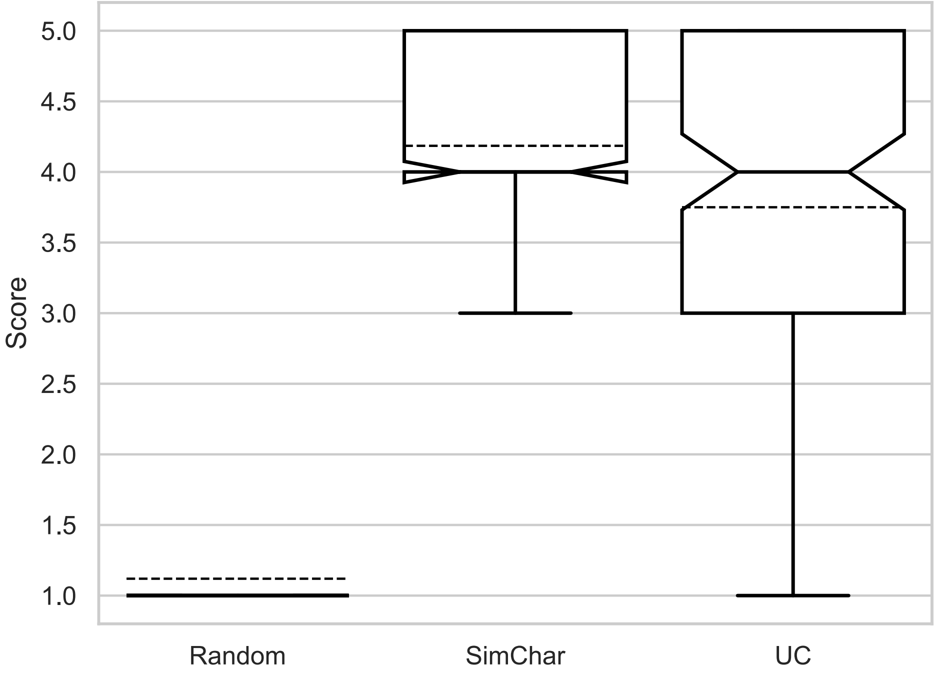
Very confusing

Confusing

Neutral

Distinct

Very distinct



Limitations & Future work

- Evaluation used GNU Unifont only
  → Need to extend the evaluation for other font families
- Participants of Human Study were English speakers
  → Need to consider linguistic/cultural spheres
  - Human perception example:
    ぬ vs. め or わ vs. ね are quite distinguishable for Japanese.

Summary

- ShanFinder is a framework to detect IDN homographs efficiently.
- ShamFinder makes use of SimChar, which is a database of homoglyphs, and confusables.txt
- SimChar contains homoglyphs not listed in the confusables.txt.
- SimChar is available at:
  - https://github.com/shamfinder/shamfinder
  - simchar.json (47MB)
- More technical details are available at arXiv:
  - https://arxiv.org/abs/1909.07539

    The paper will appear at ACM IMC 2019
    https://conferences.sigcomm.org/imc/2019/