ShamFinder: An Automated Framework for Detecting IDN Homographs

ACM IMC 2019

H. Suzuki[1], D. Chiba[2], Y. Yoneya[3], T. Mori[1,4], and S. Goto[1]

1 Waseda University, 2 NTT Secure Platform Laboratories
3 JPRS, 4 NICT

NSL Networked Systems Laboratory

Outline

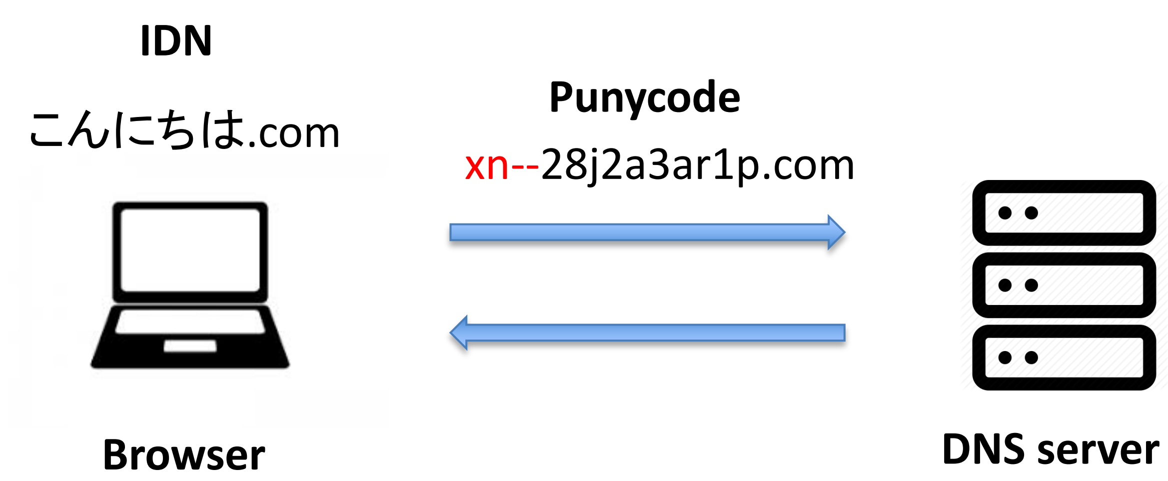
- Background
- ShamFinder framework
- Measurement Study
- Evaluation of human perception
- Discussion / Future Research

Outline

- Background
- ShamFinder framework
- Measurement Study
- Evaluation of human perception
- Discussion / Future Research

Background

- IDN (Internationalized Domain Name)
  - Domain name using Unicode Characters.
  - Punycode is used for resolving IDN



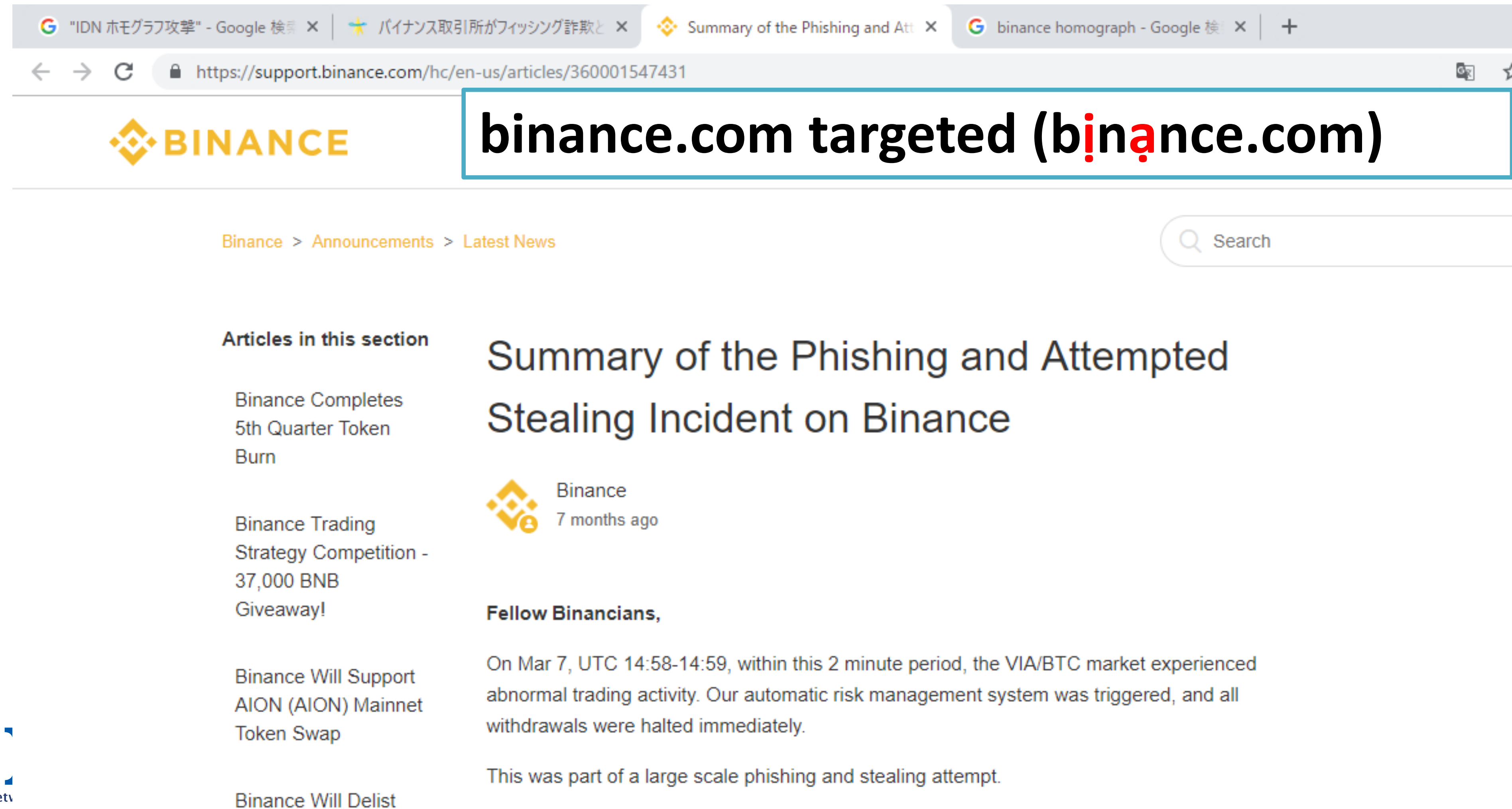
Background

- IDN Homograph attack
  - Exploits the similarity of characters to deceive users.
  - Ex) apple.com vs. applẹ.com

  - Visually similar characters is called **homoglyph**.

Background

- IDN homograph attack has become viable



Outline

- Background
- ShamFinder framework
- Measurement Study
- Evaluation of human perception
- Discussion / Future Research

Networked Systems Laboratory

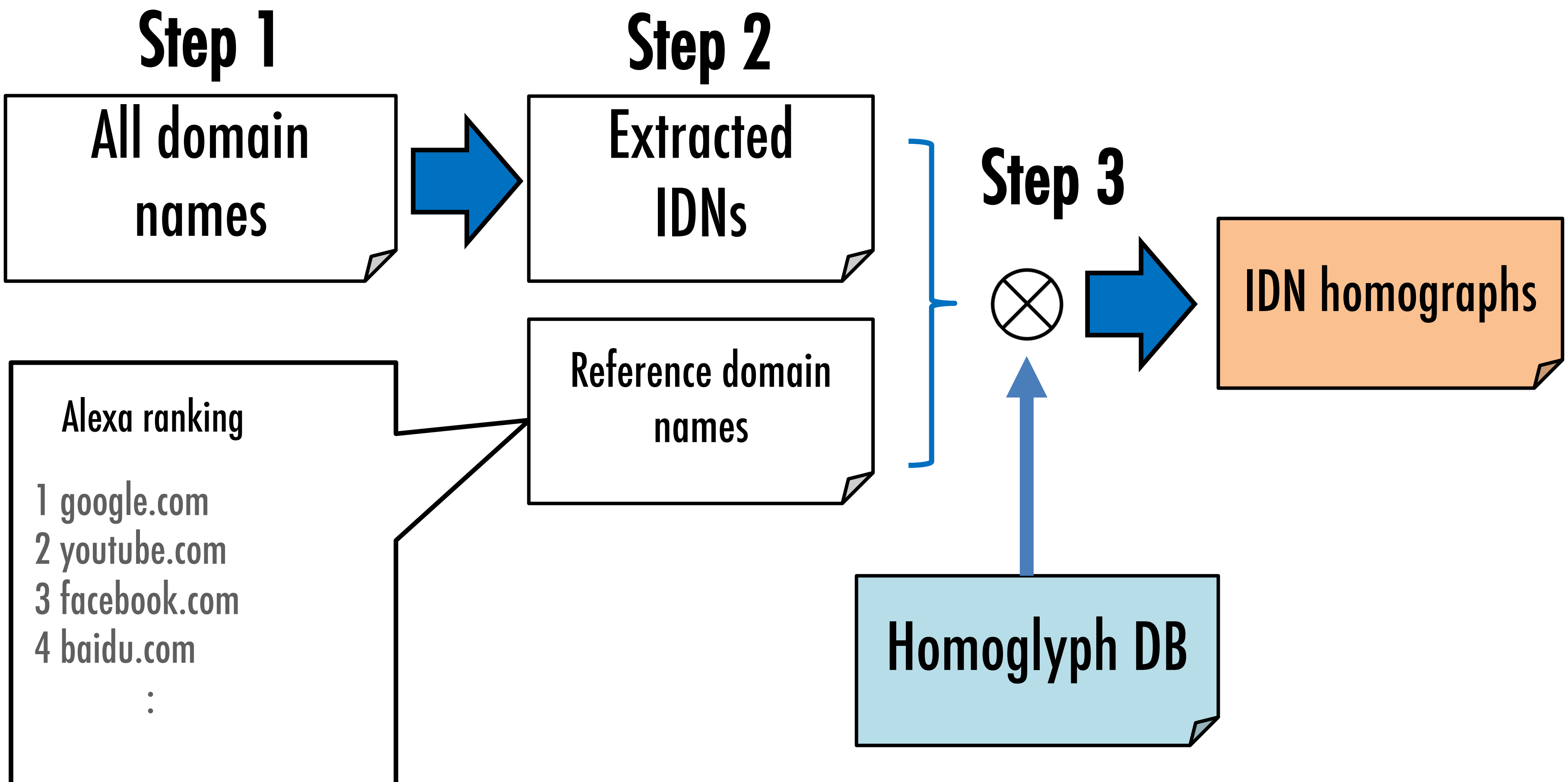
Possible Solution and Challenges

- Straightforward countermeasure
  → identify possible IDN homographs

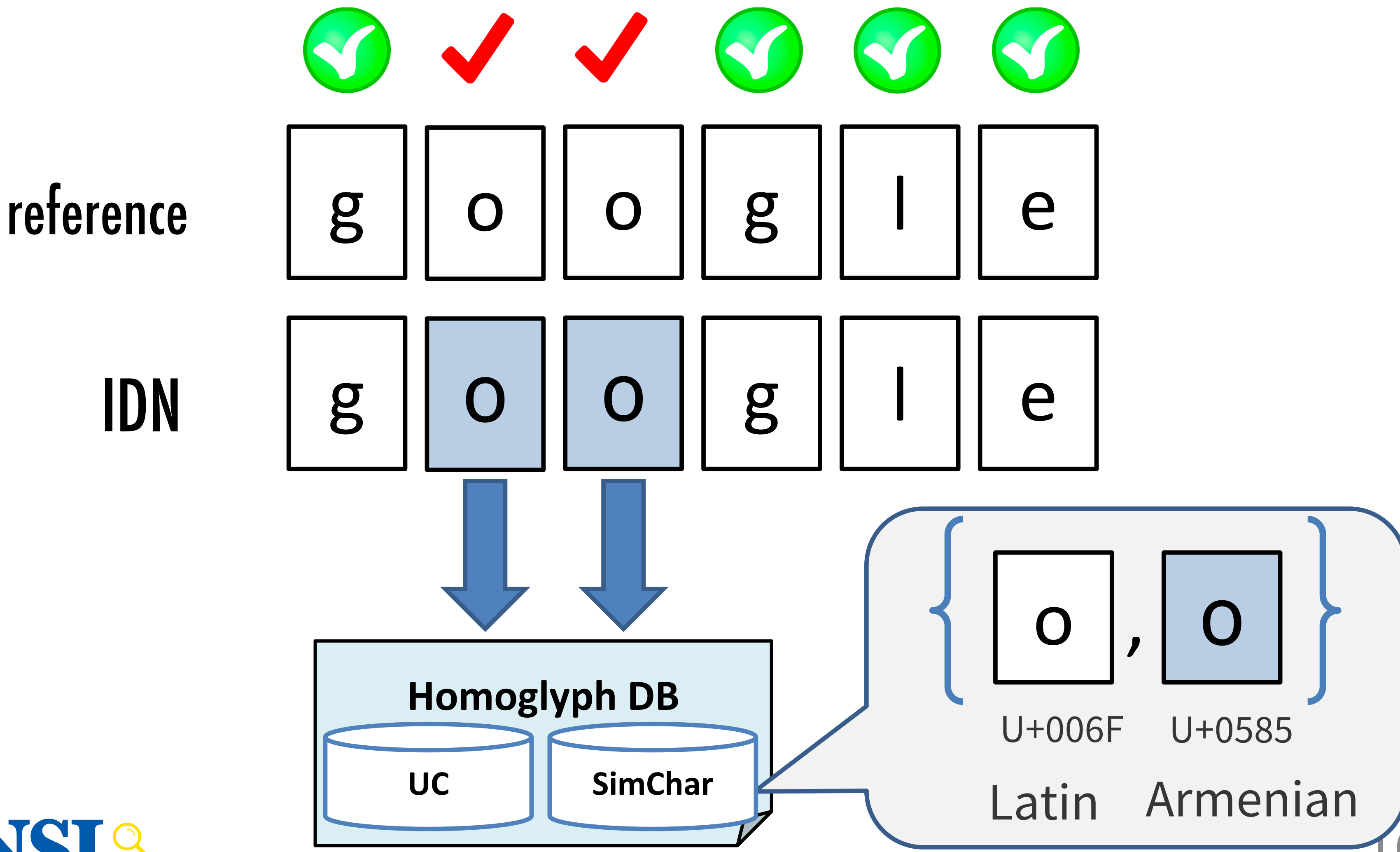
- Challenges
  – There are so many Unicode characters
  – There are so many IDN registered

→ How can we detect IDN homograph efficiently?

Our Approach – ShamFinder framework



Step 3: Detecting homograph domain

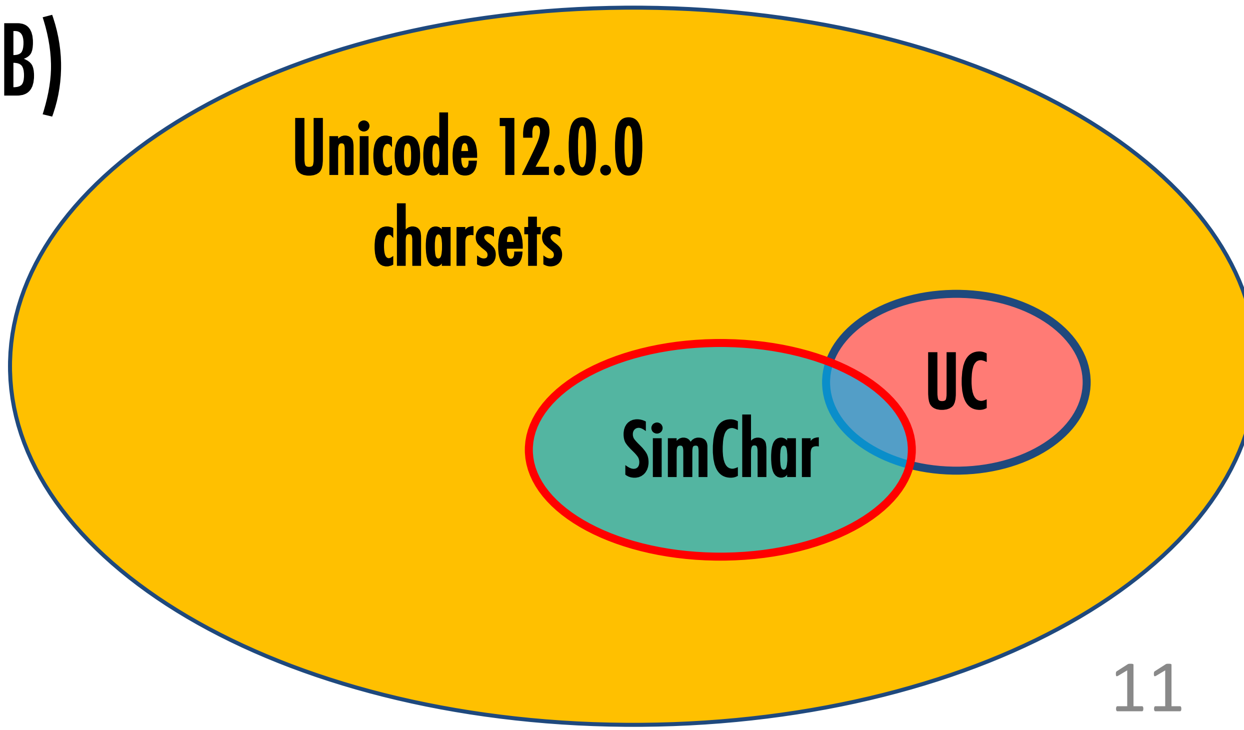


Homoglyph DB

- UC (existing DB)

  Unicode Confusables.txt (Unicode Technical Standard #39)

- SimChar (our DB)



SimChar building setups

- Target Character set
  - IDNA2008 and Unicode 12.0.0 draft
  - draft-faltstrom-unicode12-00
- Visual Appearance
  - GNU Unifont 12.0
- Computing Similarity
  - PSNR (Peak signal-to-noise ratio)
  - The number of different pixels between two images.

Computing the similarity

- PSNR: A metric that quantifies image quality degradation

$$PSNR = 10\log_{10}\left(\frac{1}{MSE}\right) = 20\log_{10} N - 10\log_{10}\Delta$$

$$\Delta = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} |I_1(i,j) - I_2(i,j)|$$ # of different pixels

Networked Systems Laboratory

Procedure of Building SimChar

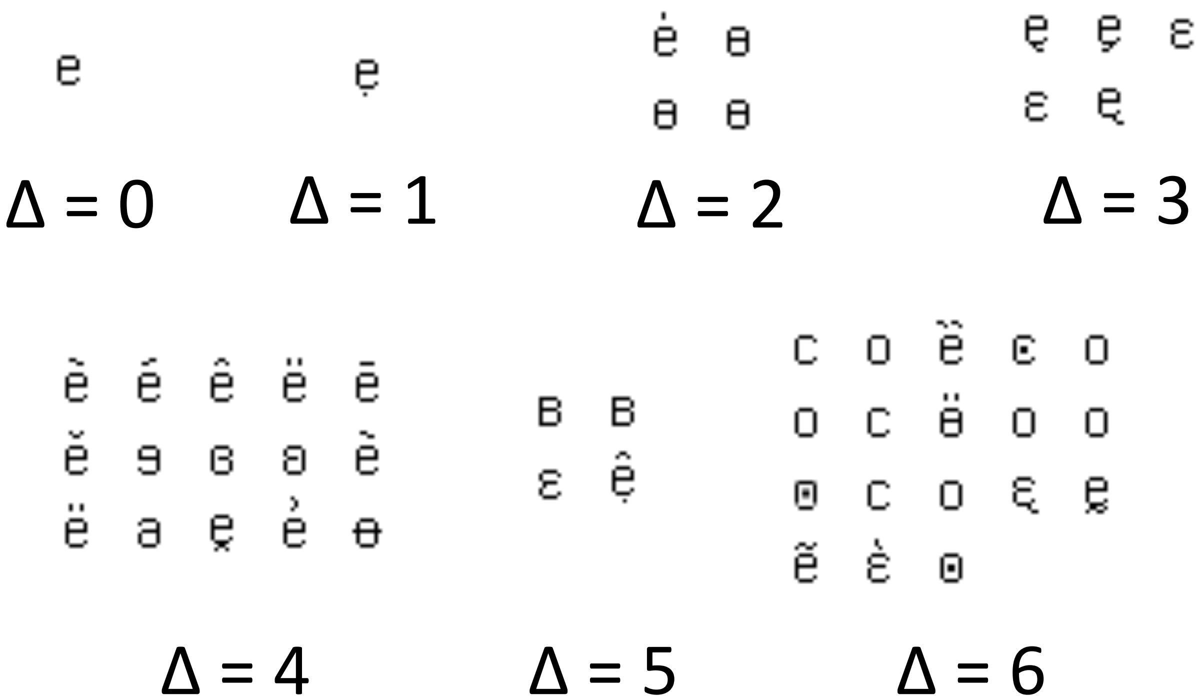
1. Create visual images.

2. Compute Δ for all pairs of characters and extract the ones with delta is 4 or less.

3. Eliminate sparse characters that contain small number of black pixels (fewer than 10)

Detected Characters

- Extracted characters that had small distance from “e”



Statistics of SimChar

Set	#characters	#Homoglph pairs
UC	9,605	6,296
SimChar	12,686	13,208
SimChar ∩ UC	233	127

SimChar						UC ∩ *IDNA*					
	#		#		#		#		#		#
‘o’	40	‘s’	14	‘f’	8	‘o’	34	‘c’	4	‘p’	3
‘e’	26	‘r’	14	‘m’	8	‘l’	12	‘d’	4	‘x’	3
‘n’	24	‘a’	14	‘g’	7	‘y’	10	‘g’	4	‘j’	2
‘w’	20	‘k’	13	‘j’	7	‘i’	9	‘f’	4	‘n’	2
‘c’	19	‘t’	13	‘p’	7	‘u’	9	‘a’	3	‘z’	2
‘l’	18	‘z’	12	‘x’	6	‘w’	8	‘b’	3		
‘u’	18	‘d’	10	‘q’	2	‘v’	6	‘e’	3		
‘h’	17	‘y’	9	‘v’	1	‘s’	5	‘h’	3		
‘i’	16	‘b’	8			‘r’	5	‘q’	3		
Total			351			Total			141		

Outline

- Background
- ShamFinder framework
- Measurement Study
- Evaluation of human perception
- Discussion / Future Research

NSL
Networked Systems Laboratory

Measurement Study

- IDN
  - 955,512 IDNs are collected for .com TLD.
- Reference domain names
  - Alexa top 10k (.com)

Networked Systems Laboratory

Number of detected IDN homographs

Homoglyph DB	Number of IDN homographs
UC	436
SimChar	3,110
UC ∪ SimChar	3,280

Networked Systems Laboratory

Number of blacklisted IDN homographs

Homoglyph DB	hpHosts	GSB	Symantec
UC	28	2	1
SimChar	222	12	7
UC ∪ SimChar	242	13	8

NSL
Networked Systems Laboratory

Analysis of active IDN homographs

Domain name	Category	#resolutions	MX	Web link	SNS
gmaıl[.]com	Phishing	615,447	◐		
döviz[.]com	Portal	127,417		✓	✓
ǵmail[.]com	Parked	74,699	◐		
gmàil[.]com	Parked	63,233	●		
expansión[.]com	Parked	56,918	◐	✓	✓
gmaiĺ[.]com	Parked	49,248	●		
yàhoo[.]com	Parked	44,368	◐		
shädbase[.]com	Parked	38,556		✓	✓
youtubê[.]com	Sale	37,713	●		
perú[.]com	Parked	36,405	●		✓

Networked Systems Laboratory

Analysis of active IDN homographs

Category	Numbers
Domain parking	348
For sale	345
Redirect	338
Normal	281
Empty	222
Error	113
Total	1,647

For business (42%)

21%

Category	Numbers
Brand protection	178
Legitimate website	125
Malicious website	35

Networked Systems Laboratory

Changing speaker

Outline

- Background
- ShamFinder framework
- Measurement Study
- Evaluation of human perception
- Discussion / Future Research

Networked Systems Laboratory

Research Question:

Are the newly detected homoglyphs (SimChar) confusable?

Studying Human Perception

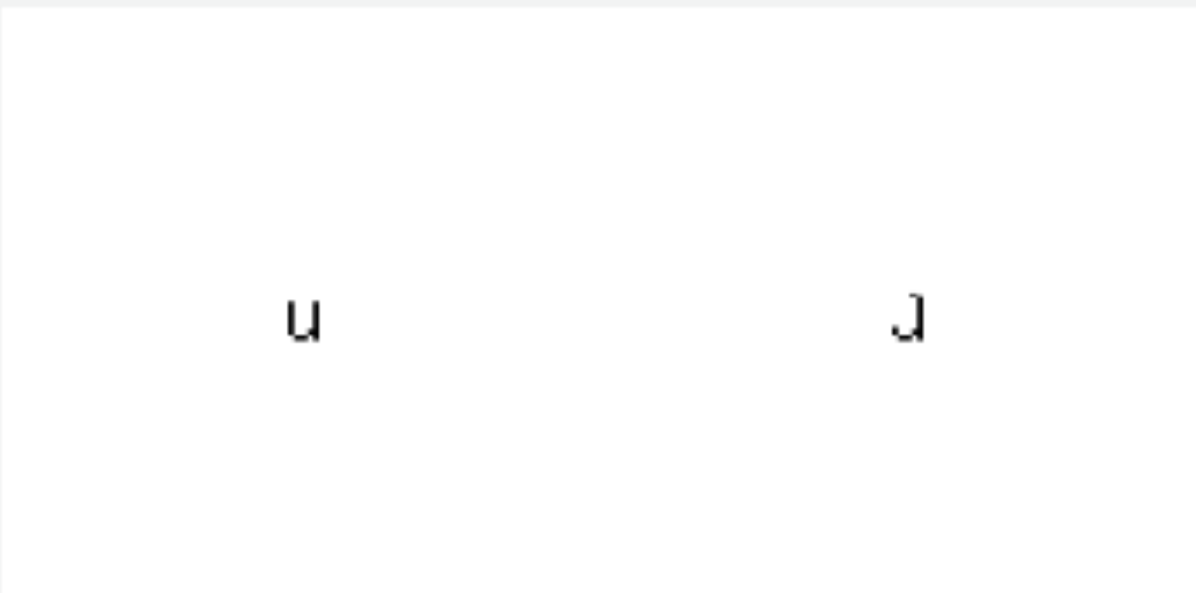
- User study experiments using MTurk
- Test the confusability of homoglyph sets, SimChar and UC.
  - Sampled 20 pairs from each set.
- Random sets for detecting the outliers

Experimental Setup

- **Pilot Study** (w/ small participants)
  – Adjusted the questionnaires
  – Measured time to complete a task
    - Minimum wage in US: 7 – 12 USD / hr
    - 15 seconds / task → 0.05 USD / task

- **Recruiting Participants** (Turkers)
  – #approved tasks > 50
  – approval rate 97%

There are two characters shown in the image. Are they distinct or confusing? Please select your answer.



Select an option

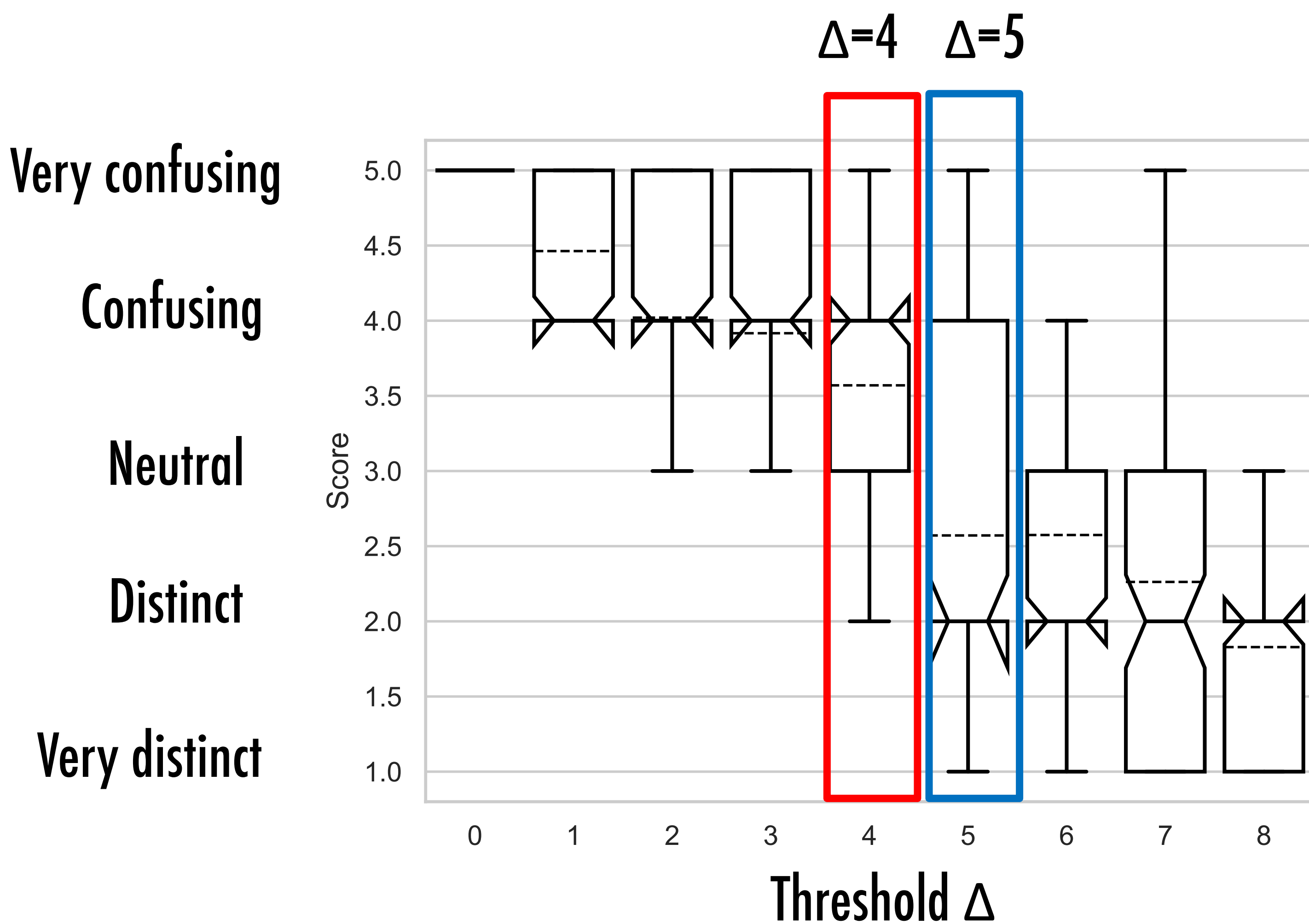
1 (Very distinct)	1
2 (Distinct)	2
3 (Neutral)	3
4 (Confusing)	4
5 (Very Confusing)	5

Experimental Setup Con’t

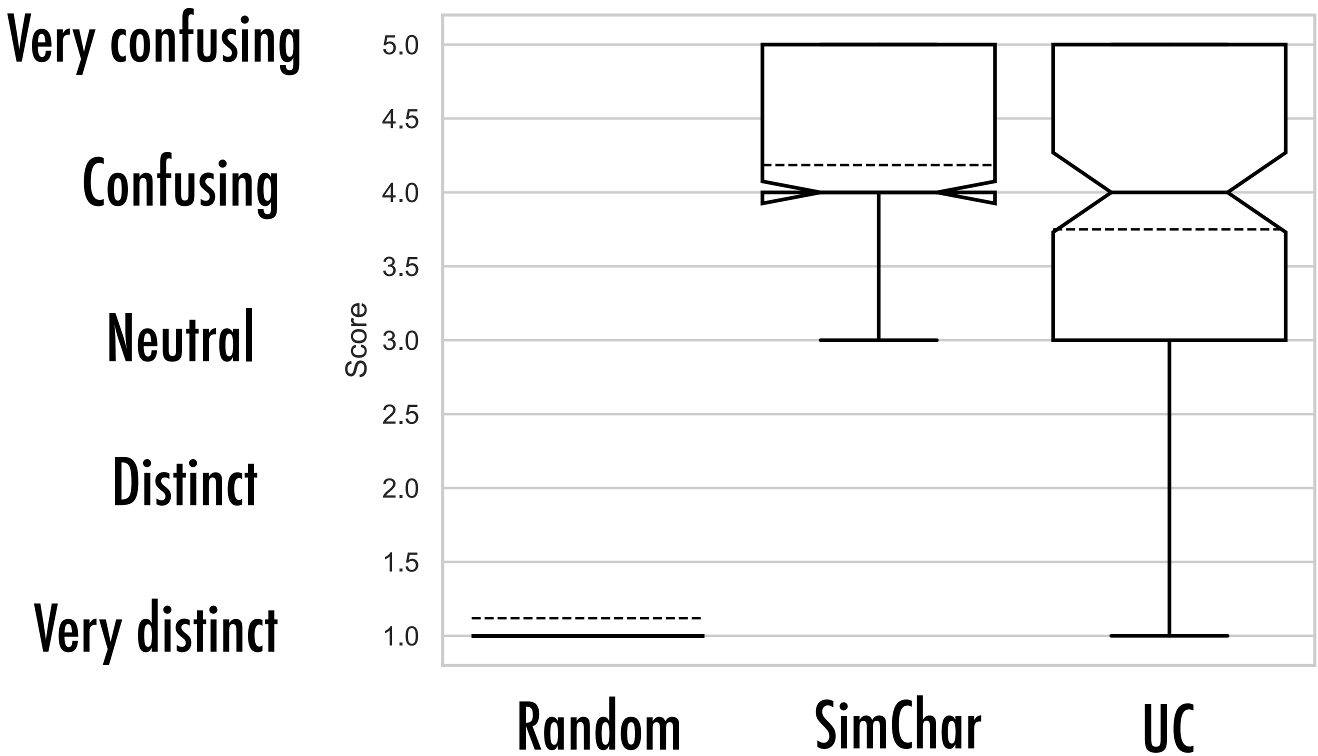
- **Experiment 1:**
  - Studied how the threshold Δ affects human perception

- **Experiment 2:**
  - Compared the confusability of UC and SimChar

Experiment 1



Experiment 2



Homoglyphs contained in SimChar are more confusable than those contained in UC

NSL
Networked Systems Laboratory

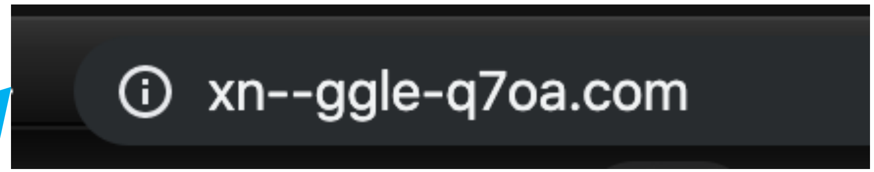
Outline

- Background
- ShamFinder framework
- Measurement Study
- Evaluation of human perception
- Discussion / Future Research

Networked Systems Laboratory

Countermeasures using SimChar

Modern browsers



http://g໐໐gle.com

Proposed interface

NSL
Networked Systems Laboratory

Limitations

- Evaluation used GNU Unifont only
  → Need to extend the evaluation for other font families

- Participants of Human Study were English speakers
  → Need to consider the linguistic/cultural background

め vs. ぬ

Networked Systems Laboratory

Contributions to the standardization communities

- W3C AHA (Anti-Homograph-Attack) CG Chartered
  - https://github.com/yoneyajp/AHA/wiki

- ICANN IDN Workshop (next month)

- IETF (TBD)
- Unicode Consortium (TBD)

Summary / Future research direction

- ShamFinder – a framework to detect IDN homographs efficiently
- SimChar – Systematically updatable homoglyph DB
  https://github.com/shamfinder/shamfinder

- Possible security applications
  – Attacks that exploit an intrinsic gap between human perception and machine processing
  – Ex) Viagra → V1@gra

NSL Networked Systems Laboratory

Questions?

Networked Systems Laboratory

Time taken for constructing SimChar.

Process	Time
Generating images	79.2 seconds
Computing Δ for all the pairs	10.9 hours
Eliminating sparse characters	18.0 seconds

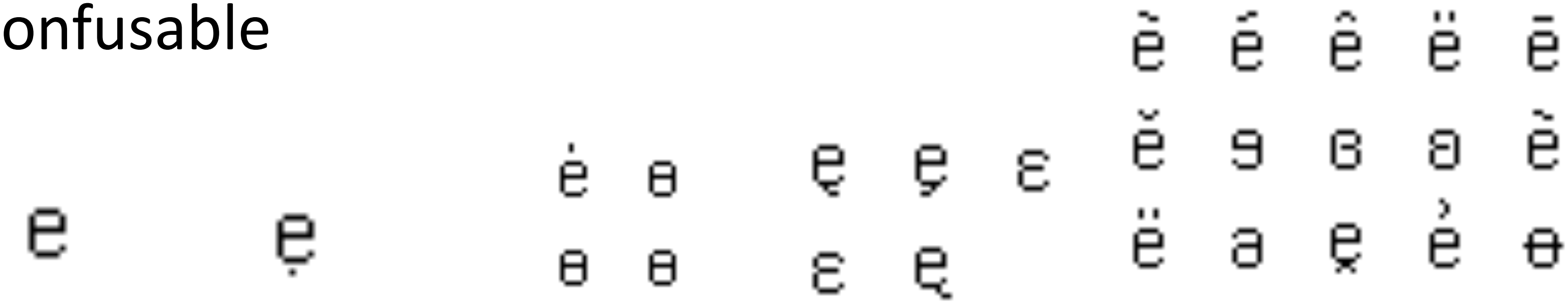
Top languages used for IDNs

Rank	Language	Number	Fraction (%)
1	Chinese	443,865	46.5
2	Korean	101,711	10.6
3	Japanese	88,970	9.3
4	Germany	53,776	5.6
5	Turkish	35,288	3.6

Top-5 ASCII domain names that have the most IDN homographs.

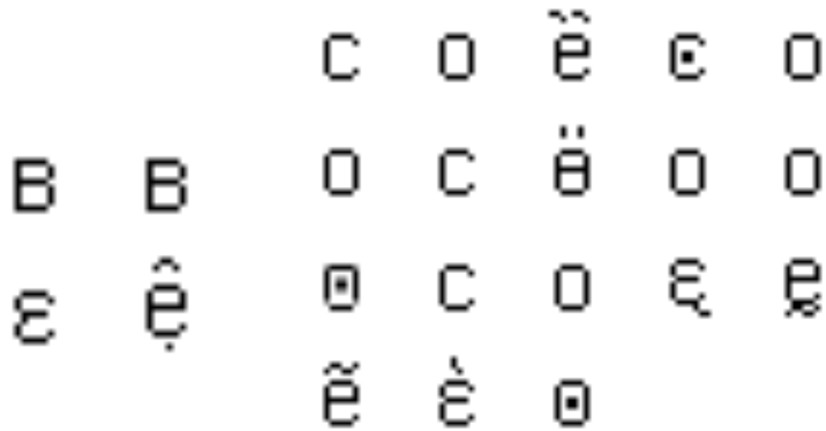
Rank	Domain name	# homographs
1	myetherwallet.com	170
2	google.com	114
3	amazon.com	75
4	facebook.com	72
5	allstate.com	68

confusable



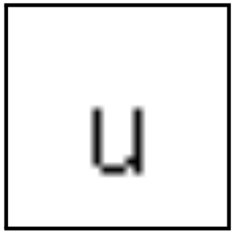
$\Delta = 0$ $\Delta = 1$ $\Delta = 2$ $\Delta = 3$ $\Delta = 4$

distinguishable



$\Delta = 5$ $\Delta = 6$

Not that confusable pairs in UC


U+0075


U+118D8


U+0079


U+028F


U+0079


U+118DC

Networked Systems Laboratory