I’m unique, just like you

*Human side-channels and their implications for security and privacy*

Matt Wixey
October 2019
Matt Wixey

- Research Lead for the PwC UK Cyber Security practice
- PhD student at UCL
- Previously worked in LEA doing technical R&D
- Black Hat USA, DEF CON, ISF Congress, BruCon, 44Con, BSides, etc
8 “What about computers not connected to the internet?” This was Matt Wixey, a security researcher at PwC UK. His talk was called ‘See no evil, hear no evil’. It was my personal favourite. https://
I'm unique, just like you: Human side-channels and their implications for security and privacy

Disclaimer (Miller, 2018)

tackers. But hackers don’t need the internet. The man had another idea:

In a dour monotone,

How to send ambient light sensors of the computer, the things that adjust the screen to board by a few wires, stood in front of a normal laptop, not connected
Aims

• Be aware of 3 human side-channels and how they work
• Practical takeaways for each side-channel, including tools
• Examine implications for security and privacy
• Know about possible countermeasures
• Explore future research ideas
The John Christie case

Background
How can we use identifiers to find an offender?

• Various things we can look at in real-world crimes
  – Fingerprints, DNA, gait, irises, voice, etc
• What about digital offences?
  – IP and MAC addresses, domains, subscriber info, emails, usernames etc
  – New problem: easily obfuscated, spoofed, anonymised
  – Other methods take us further away from the individual
    • Activity correlated to timezones (Rid & Buchanan 2014)
    • TTPs (Symantec 2011)
A possible solution

• Computers have “side-channels”
  – Unintentional leakage in primitive outputs, as a result of operations

• Is there a real-world equivalent?
  – Humans as bio-computers (Lilly, 1968) with outputs (writing, speech, etc)
  – Unintentional leakage (behavioural theory)
  – Distinctive and consistent (Shoda et al, 1994; Zayas et al, 2002)
    • Based on education, experience, training, environment, goals, etc
    • “Human side-channels”
Me: Professor, I’d like to do my essay on the etymology of the word “f***”. I just wanted to check you’d be OK with that, or would it be inappropriate?

Professor: I don’t give a s***.
Theory of forensic linguistics

- Covers other aspects, but we’re looking at one in particular:
  - Authorship attribution via stylometry
    - Spelling and orthography
    - Grammar
    - Lexicon
    - Idiom
    - Identical expressions
Real-world use cases

- Law enforcement investigations – ransom notes, texts, etc
- Plagiarism investigations
- Literature:
  - Shakespeare, The Federalist Papers, Primary Colors, JK Rowling
- Uncovering miscarriages of justice
  - e.g. police officers collaborating on statements
What forensic linguistics isn’t

- Detection of deception (cp. Van Der Zee et al, 2018; Wixey, 2018)
- Detection of intention
- Creating/comparing ‘textual fingerprints’
- Handwriting analysis
- Assessing context or content
Stylometry techniques

Complex

• Create corpus, extract features of interest
  – Parts of speech; word length; sentence length; pronouns; function words; hapax legomenon; dis legomenon; etc

• Statistical comparison of features
  – Support Vector Machines; Principal Component Analysis; Delta; etc

Basic

• Observing and noting unusual spellings/punctuation use
• Corpus/Google searching for these
Case studies (Olsson, 2009)

http://news.bbc.co.uk/1/hi/england/south_yorkshire/4407944.stm
Cyber-specific case studies

• Academic research
  – Tweets (Sultana et al, 2017; Silva et al, 2011)
  – Sockpuppet detection (Solorio et al, 2013)
  – Forum posts (Abbasi & Chen, 2005)
  – Emails (Iqbal et al, 2010)
  – Source code (Caliskan-Islam et al, 2015; Frantzeskou et al, 2007)
  – Detecting authorship deception (Pearl & Steyvers, 2012)
Cyber-specific case studies

• Operation Tripoli (Check Point, 2019)
  – Large Facebook social engineering campaign
  – Searching for repeated spelling and grammatical errors
  – Revealed multiple profiles (over 30), appear to be by same actor
• Qualitative study of IRS phone scammers (Tabron, 2016)
  – Polar tag questions, narrative violation
  – “Strengthening the human link”
• Guccifer 2.0 (Argamon, 2016)
Other use cases

- Spearphishing – different pretexts, same author
- Missives and manifestos posted online
- Ransomware instructions/notes
- Posts/Tweets claiming responsibility, coordinating attacks, etc
- Satoshi Nakamoto!
Scenario example

• A new spearphishing email comes into your org
• You notice an unusual turn of phrase
• You Google it (using special operators)
• This leads you to a forum post with a username
• Law enforcement can attribute that username to an IP address, subscriber data, etc
Scenario example

• You crawl forum posts of known threat actors and store them in a database (your corpus)
• Your org is hit by a DDoS attack using reflection/spoofing
• You notice the attack appears to be being coordinated on Twitter
• You search for other Tweets and compare them to your corpus
• You get a high match with posts by a particular user
• That user may be behind this attack
But how?

• How do you do all this?
• Isn’t forensic linguistics a really specialist discipline?
• Don’t I need at least an MSc in linguistics to do it?
• And don’t I need machine learning models, expensive statistics software, etc etc?
• Nope!
JGAAP (github.com/evllabs/JGAAP)
## Delta Calculation Worksheet 2019

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>© David L. Hoover</td>
<td>Argamon’s Delta: SUM(ABS((Test-Primary)/S.D.))</td>
<td>Analysis Area</td>
<td>Analysis Parameters</td>
<td>Instructions: View &gt;1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MAX</td>
<td>394.85 D%chg 1-2</td>
<td>Do It All</td>
<td>34 Primary Samples</td>
<td>20 Secondary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MIN</td>
<td>221.60</td>
<td>Y Delete Personal Pronouns? If “Y”, “personal pronouns”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>MEAN</td>
<td>335.25 D%chg 1-2</td>
<td>70.00 Culling %--words for which a single text supplies more</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>STDEV</td>
<td>35.41</td>
<td>2000 Words to Process--the number of MFW on which the ne</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Primary Sample</td>
<td>Stoker</td>
<td>2000.00 MFW</td>
<td>4050 Word Count: the number of words in this sheet availab</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Sample</td>
<td>The Watter’s Mou’ (1)</td>
<td>delta-scon delta-scon The Watter’s Std.Dev.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Jane Eyre (1)</td>
<td>Bronte, C_Jane Eyre (1)</td>
<td>311.96</td>
<td>7.650246</td>
<td>854827636</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Shirley (1)</td>
<td>Bronte, C_Shirley (1)</td>
<td>332.09</td>
<td>3.6533914</td>
<td>609670005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Villette (1)</td>
<td>Bronte, C_Villette (1)</td>
<td>296.80</td>
<td>2.5831853</td>
<td>348677134</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>54HideSeek (1)</td>
<td>Collins,54HideSeek (1)</td>
<td>322.29</td>
<td>3.662363</td>
<td>263877386</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>56After Dark (1)</td>
<td>Collins,56After Dark (1)</td>
<td>300.09</td>
<td>1.8176374</td>
<td>255871729</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>57DeadSecr (1)</td>
<td>Collins,57DeadSecr (1)</td>
<td>356.16</td>
<td>1.8176374</td>
<td>186516269</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>60WomanWh (1)</td>
<td>Collins,60WomanWh (1)</td>
<td>322.95</td>
<td>1.3728509</td>
<td>196612829</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>62NoName (1)</td>
<td>Collins,62NoName (1)</td>
<td>359.02</td>
<td>0.661561</td>
<td>149288976</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>66Armada (1)</td>
<td>Collins,66Armada (1)</td>
<td>349.14</td>
<td>1.578137</td>
<td>183335864</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>68Moonston (1)</td>
<td>Collins,68Moonston (1)</td>
<td>337.16</td>
<td>0.8125909</td>
<td>095727245</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>70ManWife (1)</td>
<td>Collins,70ManWife (1)</td>
<td>354.43</td>
<td>0.541467</td>
<td>117467696</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>72Poor (1)</td>
<td>Collins,72Poor (1)</td>
<td>356.93</td>
<td>1.0053234</td>
<td>144218948</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>73NewMagd (1)</td>
<td>Collins,73NewMagd (1)</td>
<td>382.86</td>
<td>1.3446392</td>
<td>101865682</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>75LawLady (1)</td>
<td>Collins,75LawLady (1)</td>
<td>361.46</td>
<td>0.7401784</td>
<td>067779825</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>76TtwoDest (1)</td>
<td>Collins,76TtwoDest (1)</td>
<td>338.09</td>
<td>0.975109</td>
<td>14678382</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>79FallenL (1)</td>
<td>Collins,79FallenL (1)</td>
<td>375.22</td>
<td>1.1288705</td>
<td>157972639</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>80Jezebel (1)</td>
<td>Collins,80Jezebel (1)</td>
<td>356.84</td>
<td>0.6604045</td>
<td>082783681</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>81BlackR (1)</td>
<td>Collins,81BlackR (1)</td>
<td>369.01</td>
<td>0.5343797</td>
<td>165845115</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>82HeartSci (1)</td>
<td>Collins,82HeartSci (1)</td>
<td>362.57</td>
<td>0.7714756</td>
<td>086844666</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>84IsayNo (1)</td>
<td>Collins,84IsayNo (1)</td>
<td>394.85</td>
<td>1.6831709</td>
<td>180241452</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
I'm unique, just like you: Human side-channels and their implications for security and privacy

PwC

October 2019
stylo (R library) - github.com/computationalstylistics/stylo
Shylo (stylo wrapper) - github.com/severinsimmler/shylo

I'm unique, just like you: Human side-channels and their implications for security and privacy

Forensic linguistics

October 2019
• Shylo (stylo wrapper)

I'm unique, just like you: Human side-channels and their implications for security and privacy

Forensic linguistics
## Summary of tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Free?</th>
<th>Ease of use</th>
<th>Method(s)</th>
<th>Outputs</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>JGAAP</td>
<td>Yes</td>
<td>Hard</td>
<td>Multiple</td>
<td>Numeric</td>
<td>Possible</td>
</tr>
<tr>
<td>Delta sheets</td>
<td>Yes</td>
<td>Moderate</td>
<td>Delta</td>
<td>Numeric</td>
<td>Difficult</td>
</tr>
<tr>
<td>Stylometry</td>
<td>Yes</td>
<td>Easy</td>
<td>PCA</td>
<td>Graphs</td>
<td>Possible</td>
</tr>
<tr>
<td>Stylo (R)</td>
<td>Yes</td>
<td>Easy</td>
<td>Multiple</td>
<td>Graphs</td>
<td>Possible</td>
</tr>
<tr>
<td>Shylo</td>
<td>Yes</td>
<td>Easy</td>
<td>Multiple</td>
<td>Multiple</td>
<td>Possible</td>
</tr>
</tbody>
</table>

I'm unique, just like you: Human side-channels and their implications for security and privacy

PwC
Caveats

- Register makes a big difference
- Need a baseline of text – sizeable samples
- Ground truth may also be required (depending on objective)
- Strategy will be decided by circumstances
- Time lapse may affect results
- Not fingerprints, no 100% accuracy – not a silver bullet
I'm unique, just like you: Human side-channels and their implications for security and privacy

Forensic linguistics
Privacy implications

- Attribution of texts written under a separate identity
- Diminish anonymity
Countermeasures

- Linguistic style is often unconscious
- Awareness of it can facilitate disguising it
- Imitating another’s style, either during or after writing
  - Writing in another ‘voice’ (cp. 1984)
- Google Translate
- Combining with other authors
- Running forensic linguistic tools – Anonymouth (Brennan et al, 2012; McDonald et al, 2012)
What can I do now?

- Test tools out
  - Text from previous attacks & open source data
  - Start building corpus
  - Have a play, let me know what you think!
- Explore how useful/applicable it would be for your use cases
- Think about other scenario/contexts it could be used in
Behavourial signatures

“I got an AUC of 0.99 but that’s basically 1” – Jay-Z (a ROC fella)
Active area of research in attribution: who hacks, and why
- Motivation, skills, attack behaviours (Landreth, 1985; Salles-Loustau et al, 2011)
- Attitudes and culture (Chiesa et al, 2008; Watters et al, 2012)
- Psychological elements (Shaw et al, 1998)
- Specific actions undertaken (Ramsbrock et al, 2007)
Background

• What hasn’t been done: comparing profiles of attackers
• Case Linkage Analysis (CLA)
  – Linking crimes together based on common features
  – Note: this is not offender profiling!
  – Offender profiling: After analysing this crime, I think the offender is a charismatic security researcher with a fast-disappearing hairline
  – CLA: After analysing crimes A and B, they have features XYZ in common. I know charismatic balding researcher Matt Wixey committed crime A, so he may have also committed crime B
• Statistical comparison of crime scene behaviours (Woodhams & Grant, 2006)
• Some success in academic literature, with real-world crimes
  – Homicide, burglary, robbery, sexual assault, arson, etc
  – But not cyber attacks (until this research!)
• Grubin et al, 1997; Mokros & Alison, 2002; Tonkin et al, 2008
• Based on same principles of distinctiveness and consistency
Why? What’s the point?

- If we can conclude that two crimes are linked, we can:
  - Save time and resources by investigating them together
  - Build up a body of evidence against an offender
  - Potentially identify weaknesses/flaws in offender’s strategies
  - Attribute multiple crimes to one offender if/when they’re identified
  - Decision-making aid
Example – Crime A

Example – Crime B

Linking crimes

- We know these crimes are probably linked
- But how do we prove it?
- What features of the crime might we look at?
Step 1: Identify behaviours

- Create behavioural domains – broad categories of the crime, e.g. “equipment used”, “property targeted”, etc
- For each domain, look at very granular behaviours and turn them into yes/no questions
- E.g. for “equipment used”: did attacker use stencil? Did they use colour? Did they sign the image? Did they use X paint? Or Y paint?
- Repeat this for all behavioural domains – the more granular, the better!
Step 2: Similarity coefficient

\[ J = \frac{x}{(x + y + z)} \]

- Jaccard’s Coefficient (Tonkin et al, 2008)
- 1 per domain
- \( X = \) count of behaviours present in both attacks
- \( Y = \) count of behaviours present in Crime A, but not B
- \( Z = \) present in Crime B, but not A
- \( 1 = \) perfect similarity, \( 0 = \) perfect dissimilarity
Step 3: Logistic regression

- Can we predict whether the crimes are paired (e.g. committed by the same person)?
- Logistic regression lets us test this out
- Statistical way of finding out which domain contributes more
  - e.g. is “equipment used” more effective than “property targeted”?
- And, combined, how well they can be used to predict linkage?
  - SPSS, R, etc – loads of tutorials online
Step 3: Logistic regression

• Run for each behavioural domain to get:
  – Positive or negative correlation
  – A p-value (statistical significance)
  – Amount of variance that a variable explains

• Repeat with forward stepwise logistic regression
  – Will automatically start with one domain, and add one at each step
  – If it contributes to predictive power, keep it, else discard from the model
  – Determines optimal combination of domains
Step 4: ROC Curves

- Put regression results into ROC curves
- Graphical representation of performance
- Commonly used to look at predictive accuracy of machine learning
  - Plots x (prob of false positive) against y (prob of true positive)
  - More reliable measure of predictive accuracy (Tonkin et al, 2008; Swets, 1988)
  - You’ll get ‘area under the curve’ (AUC) values
Step 4: ROC Curves

- Diagonal: no better than chance
- The higher the AUC value, the greater the predictive accuracy
  - 0.5 – 0.7 = low
  - 0.7 – 0.9 = good
  - 0.9 – 1.0 = high
- Swets, 1988

https://www.statisticshowto.datasciencecentral.com/receiver-operating-characteristic-roc-curve/
Why apply it to cyber attacks?

- In principle, same concepts will apply
- Never been done before
- OSCP, 2014 – idea
- New contribution to CLA body of literature
Cyber attacks - scenario

- In 2017, Business Corp is attacked
- The attacker infects the network with a malicious macro doc
- And then pokes around the filesystem
- Sets up a permanent backdoor
- And starts exfiltrating data
- In 2019, Business Corp is attacked again
- The methodology looks similar – but how do we know it’s the same threat actor?
Experiment – cyber attacks

- Modified open source Python SSH keylogger (strace)
  - [https://github.com/NetSPI/skl](https://github.com/NetSPI/skl)
- Two VMs, exposed on internet over SSH (like honeypots)
- One account per user per box
- Deliberate privesc vulnerabilities, plus fake data to exfiltrate
- 10x pentesters/students asked to SSH in (2 attacks each)
  - And get root, steal data, cover tracks, poke around
Classification

- Define behavioural domains e.g. ‘navigation’, ‘enumeration’, etc
- Classify keystrokes as commands (‘behaviours’)
  - Turn into ‘yes/no’ questions
  - “Did attacker try to wget malware from a remote site after compromise?”
  - Assign 1 if yes, 0 if no
  - End up with binary string for each offence in each domain
Experiment

- Keystrokes collated per user, split into behavioural domains
  - Navigation, enumeration, exploitation
  - 40 individual behaviours per domain

```plaintext
sudo    chmod 755
su      chmod 777
sudo [command] chmod +x
sudo [username] chmod +x [dir]
sudo -n vi
su root nano
su - [username] cat /etc/sudoers
sudo -s sudo -s
sudo su sudo -I
gcc file.c -o file bash
CVE exploits looks for ssh authorized keys
wget mount
```
Experiment

- Automated calculation of Jaccard values

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation (linked)</td>
<td>0.756</td>
<td>0.756</td>
<td>0.166</td>
</tr>
<tr>
<td>Navigation (unlinked)</td>
<td>0.163</td>
<td>0.125</td>
<td>0.134</td>
</tr>
<tr>
<td>Enumeration (linked)</td>
<td>0.641</td>
<td>0.708</td>
<td>0.259</td>
</tr>
<tr>
<td>Enumeration (unlinked)</td>
<td>0.108</td>
<td>0.087</td>
<td>0.122</td>
</tr>
<tr>
<td>Exploitation (linked)</td>
<td>0.58</td>
<td>0.555</td>
<td>0.281</td>
</tr>
<tr>
<td>Exploitation (unlinked)</td>
<td>0.091</td>
<td>0.077</td>
<td>0.097</td>
</tr>
</tbody>
</table>
## Experiment

- Imported results into SPSS
- Performed logistic regression (direct and forward stepwise)
- Also used SPSS for ROC curves

<table>
<thead>
<tr>
<th>Variable</th>
<th>AUC</th>
<th>Sig.</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td>0.992</td>
<td>p &lt; 0.001</td>
<td>0.007</td>
<td>0.978 - 1.0</td>
</tr>
<tr>
<td>Enumeration</td>
<td>0.912</td>
<td>p &lt; 0.001</td>
<td>0.081</td>
<td>0.753 - 1.0</td>
</tr>
<tr>
<td>Exploitation</td>
<td>0.964</td>
<td>p &lt; 0.001</td>
<td>0.028</td>
<td>0.91 - 1.0</td>
</tr>
<tr>
<td>Keystroke Interval</td>
<td>0.572</td>
<td>NS</td>
<td>0.102</td>
<td>0.373 - 0.771</td>
</tr>
<tr>
<td>Command Interval</td>
<td>0.58</td>
<td>NS</td>
<td>0.113</td>
<td>0.358 - 0.802</td>
</tr>
<tr>
<td>Backspaces</td>
<td>0.702</td>
<td>p &lt; 0.05</td>
<td>0.094</td>
<td>0.519 - 0.886</td>
</tr>
<tr>
<td>Optimal</td>
<td>1</td>
<td>p &lt; 0.001</td>
<td>0</td>
<td>1.0 - 1.0</td>
</tr>
</tbody>
</table>
Applicability and approaches

- Honeypots
- Build up a corpus of attackers
- Could also identify attackers who’ve trained together
Some offenders show more distinctiveness than others
  - Bouhana et al, 2016

Some behaviours less consistent
  - Bennell & Canter, 2002; Bennell & Jones, 2005

MO is a learned behaviour, and offenders develop
  - Pervin, 2002; Douglas & Munn, 1992

Offenders will change behaviours in response to events
  - Donald & Canter, 2002
Caveats

• This experiment:
  – Small sample, only commands
  – Only one OS/scenario
  – Not ‘real’ attackers – knew they wouldn’t suffer consequences
  – Not all attackers will have the same motivations, could affect results
  – Not 100% accurate
Privacy implications

- People can be linked to separate hosts/identities
- Based on approaches, syntax, and commands
- Regardless of anonymising measures
- Regardless of good OPSEC elsewhere
- Could be linked to historical or future activity
Countermeasures

• Similar to defeating authorship identification
• Make a conscious decision to disguise your style
  – CLA different – e.g. alias command would not work
  – Hard to automate – can’t predict commands in advance
  – Could semi-automate, using scripts
  – Randomising ordering of command switches
  – Switching up tools e.g. wget instead of curl; vi instead of nano, etc
What can I do now?

- Give it a go!
  - Keylogger on CTF machines (make sure participants are aware, take appropriate ethical measures)
  - Classification and calculate Jaccard score – pretty simple
  - Calculate logistic regression scores – again, pretty simple
  - ROC curve analysis (same tools)
  - Have a go at automating! R/Python probably best place to start
  - Other behavioural domains, e.g. evasion techniques
  - Whitepaper available (contact me!) or see DEF CON 2018 talk
Cultural CAPTCHAs

“Of course I remember Crinkley Bottom”
Background

• “Is this account a human or a bot?”
  – Lots of academic and practical research (Filippoupolitis et al, 2014)
  – Botometer, Twitteraudit, Botcheck, Botsentinel
• Certain behaviours/features can be “tells”
• Harder question: “Is this account owner really X nationality?”
• Context: hostile accounts influencing conversations or consensus
  – We think they’re probably human
  – But how do we prove they’re authentic?
Background

• Enter “cultural CAPTCHAs”
• Cultural artefacts which haven’t spread beyond origin
• In many cases this can be popular culture, but also:
  – Language
  – Cultural norms and expectations
  – Food
  – Music
  – Traditions, etc
Let’s try an example - who are these two men?

RAISE YOUR HAND if you know
Experiment

• Let’s try another

Who's probably on the left?

I'm unique, just like you: Human side-channels and their implications for security and privacy

Vic and Bob - Wikipedia
https://en.wikipedia.org/wiki/Vic_and_Bob

Vic and Bob, also known as Reeves and Mortimer, are a British comedy double act consisting of Vic Reeves and Bob Mortimer (born 23 May 1959). They have ...
I'm unique, just like you: Human side-channels and their implications for security and privacy
Experiment

- One for any Americans 😊
- Who’s this, and where is he from?

Another example

I'm unique, just like you: Human side-channels and their implications for security and privacy

https://knowyourmeme.com/memes/jake-from-state-farm
https://www.reddit.com/r/MovieDetails/comments/7vt5wh/inglourious_basterds_2009_you_can_clearly_see_the/

Three glasses.
Other possible examples

- Food
- Music
- Cultural norms and quirks
- Popular culture
- Education

https://www.youtube.com/watch?v=2cgRd2WJXpo
Case studies

BotSentinel.com

Trollbot Rating: Moderate

This report was created for

Report created: 2019-04-23 09:50:18

Our analysis has concluded that this account exhibits moderate tweet activity and is not a trollbot account.
Case studies

I have administered this test multiple times now, on multiple pro-Brexit accounts with multiple linked patterns of posting. Never gets a reply. They can't answer it.

BOT TEST

Who is the bloke on the left?
Case studies

I'm unique, just like you: Human side-channels and their implications for security and privacy

PwC

Cultural CAPTCHAs

British people challenge liars when they meet them. It's part of our national identity.

Provide irrefutable evidence that I'm not British or that I'm a bot

Answer the question

How about you just fuck off you annoying little worm

Answer it.

You fuck off too
I'm unique, just like you: Human side-channels and their implications for security and privacy

Cultural CAPTCHAs
Applicability and approaches

- ‘CAPTCHA’-style verification system
- For accounts reported as possibly false/hostile?
- Give users option of selecting a different CAPTCHA
  - They genuinely may not know the answer!
I'm unique, just like you: Human side - channels and their implications for security and privacy
Caveats

- Reliant on specific cultural knowledge
  - Some may be age-dependent
  - May become increasingly hard to find examples
  - Users may genuinely not know the answer
  - cp. genuine CAPTCHAs
- Images cannot be searchable online
  - Manipulation/generation to avoid TinEye, reverse image search, etc
What can I do now?

- Come up with your own examples and implementations
- Test on social media
- Research on effectiveness at scale
- How resilient are cultural CAPTCHAs?
  - Not an area I know much about, but with click-farm workers, catfish, etc – how much research do they do into culture and language?
  - Interesting area for future work
Conclusion
Key takeaways

- Human side-channels offer under-explored, unconventional, and often cost-effective, opportunities for attribution and defence

- These are often specialist areas – but barrier to entry isn’t as high as you might think!

- Tools and resources are available now, often open-source, to test these things out
Next steps and future research

- Expanding PoCs, applying techniques to more scenarios
- Other side-channels
- Further research into nature and scope of cultural CAPTCHA
- Further research into applicability and effectiveness of forensic linguistics and behavioural signatures as investigative tools
- Automate some of this stuff, especially FL and CLA
- Get in touch! Let’s discuss 😊
- matt.wixey@pwc.com, @darkartlab
Aims - review

- Be aware of 3 human side-channels and how they work
- Practical takeaways for each side-channel, including tools
- Examine implications for security and privacy
- Know about possible countermeasures
- Explore future research ideas
Thank you!

@dankartlab
matt.wixey@pwc.com
References


github.com/computationalstylistics/stylo

github.com/evllabs/JGAAP

github.com/jpotts18/stylometry

github.com/NetSPI/skl

github.com/psal/anonymouth

github.com/severinsimmler/shylo


