Domain and Website Attribution beyond WHOIS

Speaker: Silvia Sebastián
Authors: S. Sebastián, R. Diugan, J. Caballero, I. Sanchez-Rola, L. Bilge
Domain Attribution: WHOIS

>whois imdea.org_

Registrant Identity:  
Who Owns The Domain

Registrar Identity:  
Who Registered the Domain
Domain Attribution: WHOIS

>whois imdea.org_

Registrant Identity: Who Owns The Domain

Registrar Identity: Who Registered the Domain

Limitations

Privacy Protection Services

Privacy Regulations (GDPR)
Law Enforcement

Companies

End Users
Domain Attribution: WHOIS

Law Enforcement  Companies  End Users
Domain Attribution: WHOIS

Law Enforcement

Companies

End Users
Manual Domain Attribution

Who?
Manual Domain Attribution

Who?

WhoseDomain
Manual Domain Attribution

WhoseDomain
> whois cosmopolitan.de

Who?
> whois cosmopolitan.de

The DENIC whois service doesn't disclose any information concerning the domain holder.
> whois cosmopolitan.de

The DENIC whois service doesn't disclose any information concerning the domain holder.
> whois cosmopolitan.de

The DENIC whois service doesn't disclose any information concerning the domain holder.
> whois cosmopolitan.de

The DENIC whois service doesn't disclose any information concerning the domain holder.
cosmopolitan.de

> whois cosmopolitan.de

The DENIC whois service doesn't disclose any information concerning the domain holder.
> whois cosmopolitan.de

The DENIC whois service doesn't disclose any information concerning the domain holder.
<table>
<thead>
<tr>
<th>Domain</th>
<th>Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>365media.com</td>
<td>e.gg</td>
</tr>
<tr>
<td>acompli.net</td>
<td>friendfeed.com</td>
</tr>
<tr>
<td>atlantic.nl</td>
<td>liverail.com</td>
</tr>
<tr>
<td>atlassolutions.com</td>
<td>oversightboard.com</td>
</tr>
<tr>
<td>azure.com</td>
<td>wunderlist.com</td>
</tr>
<tr>
<td>AGGREGATE INTELLIGENCE</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>365media.com</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>META</th>
</tr>
</thead>
<tbody>
<tr>
<td>atlassolutions.com</td>
</tr>
<tr>
<td>friendfeed.com</td>
</tr>
<tr>
<td>commercepartnerhub.com</td>
</tr>
<tr>
<td>e.gg</td>
</tr>
<tr>
<td>oversightboard.com</td>
</tr>
<tr>
<td>liverail.com</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MICROSOFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>acompli.net</td>
</tr>
<tr>
<td>azure.com</td>
</tr>
<tr>
<td>atlantic.nl</td>
</tr>
<tr>
<td>wunderlist.com</td>
</tr>
</tbody>
</table>
Nodes

Indicators

Type: Url
Value: https://cosmopolitan.de/
Attributes:
- Timestamp 2023-03-14
- Exploration Seed

Examples

Identities
Contact
Network Infrastructure
Social Handles
Hashes
Expansions

**Expansion:**
source indicator type + destination indicator type

**Dataset:**
Set of expansions

- Whois
- Certificates
- VirusTotal
- Content Attribution
Pivoting

WhoseDomain

Datasets:
- VirusTotal
- Whois
- Cert
- Content Attribution

Expansion

Filtering

Domain → Identity
Pivoting

WhoseDomain

Datasets: VirusTotal, Whois, Cert, Content Attribution

Domain → Expansion → Filtering → Identity
First Party Identity

WhoseDomain

Datasets: VirusTotal, Whois, Cert

Content Attribution

NLP

Expansion

Filtering

Ranking

Domain

Identity
Contributions

- Novel automatic approach to attribute domains beyond WHOIS.
- 2 novel techniques to identify first-party identities among third parties.
- F1 Score of 0.94 compared to 0.54 from Whois.
- Publicly Available: https://hub.docker.com/r/dianecode/whosedomain
Content Attribution

URL → Download

Regex

Parsers

NLP

Ranking

FIRST PARTY INDICATORS

WhoseDomain
Content Attribution: NLP

NER

WhoseDomain
Entities: Organization, Person, Address, URL.

Broadcom Inc - ORG
1320 Ridder Park Dr, San Jose, CA 95131, USA - ADDR
Entities: Organization, Person, Address, URL.

8 Supported Languages

- Broadcom Inc - ORG
- 1320 Ridder Park Dr, San Jose, CA 95131, USA - ADDR
Content Attribution: NLP

- Entities: Organization, Person, Address, URL.
- 8 Supported Languages
- Not fine tuned → FPs → Ranking.

- Broadcom Inc - ORG
- 1320 Ridder Park Dr, San Jose, CA 95131, USA - ADDR
- European Economic Area - ORG
- Google - ORG
This privacy policy governs your use of the software provided by Khushhal Maheshwari ('khushalm', 'we', 'us', 'ML QUIZ')
This privacy policy governs your use of the software provided by Khushhal Maheshwari (‘khushhalm’, ‘we’, ‘us’, ‘ML QUIZ’)
This privacy policy governs your use of the software provided by Khushhal Maheshwari('khushalm','we','us','ML QUIZ')
This privacy policy governs your use of the software provided by Khushhal Maheshwari ('khushhalm', 'we', 'us', 'ML QUIZ')

Khushhal Maheshwari, khushhalm, ML QUIZ
Content Attribution

URL → Download → Regex → Parsers → NLP → Ranking → FIRST PARTY INDICATORS

WhoseDomain
Content Attribution: Ranking

Ranking

daytondailynews.com [RegEx]

©2021 Cox Media Group. All Rights Reserved [RegEx]

Cox Media Group, Inc. [NER]

cmgcopyright@cmg.com [Link, RegEx]

Legends971 [Link]

countrylegends971 [Link]

American Stock Exchange [NER]

American Arbitration Association [NER]

Google Analytics [NER]

CMG [Syn, NER]

CMG Affiliate [Syn, NER]

New York Stock Exchange [NER]
Content Attribution: Ranking

©2021 Cox Media Group. All Rights Reserved [RegEx]
Cox Media Group, Inc. [NER]
cmgcopyright@cmg.com [Link, RegEx]
CMG [Syn, NER]
CMG Affiliate [Syn, NER]
daytondailynews.com [RegEx]

American Arbitration Association [NER]
American Stock Exchange [NER]
New York Stock Exchange [NER]
Legends971 [Link]
countrylegends971 [Link]
Google Analytics [NER]

WhoseDomain
Content Attribution: Ranking

1. daytondailynews.com [RegEx]
2. countrylegends971 [Link]
3. American Arbitration Association [NER]
   American Stock Exchange [NER]
   New York Stock Exchange [NER]

CMG [Syn, NER]
CMG Affiliate [Syn, NER]

©2021 Cox Media Group. All Rights Reserved [RegEx]
Cox Media Group, Inc. [NER]
cmgcopyright@cmg.com [Link, RegEx]
Evaluation

- WhoseDomain Attribution Accuracy.
- Dataset Impact:
  - Sources Attribution Accuracy
  - Chaining Sources Accuracy.
- Updating Tracking Lists.
Evaluation: Attribution Accuracy - F1 Score

739 Domains

4 Datasets
Evaluation: Chaining Sources

WHOIS

F1 Score

WhoseDomain
Evaluation: Chaining Sources

F1 Score

CONTENT

VIRUSTOTAL
0.91

WHOIS
0.59

WhoseDomain
Evaluation: Chaining Sources

F1 Score

CERTS
CONTENT
0.94
0.89

VIRUSTOTAL
WHOIS
0.91
0.59

WhoseDomain
Evaluation: Updating Tracker Domain Lists

3,001 Trackers from Disconnect
Evaluation: Updating Tracker Domain Lists

3,001 Trackers from Disconnect → 77% → WhoseDomain Attributes → 60% → WhoseDomain and Disconnect Agreement
Evaluation: Updating Tracker Domain Lists

- 3,001 Trackers from Disconnect
- WhoseDomain Attributes
- WhoseDomain and Disconnect Agreement

- 74 TPs
- 7 FPs
- 19 FNs

77% F1 Score

0.85

60%
Evaluation: Updating Tracker Domain Lists

3,001 Trackers from Disconnect

WhoseDomain Attributes

74 TPs
7 FPs
19 FNs

0.85 F1 Score

77%

60% WhoseDomain and Disconnect Agreement
Novel automatic approach to attribute domains beyond WHOIS.

2 novel techniques to identify first-party identities among third parties.

F1 Score of 0.94 compared to 0.54 from Whois.

Publicly Available:  https://hub.docker.com/r/dianecode/whosedomain
Domain and Website Attribution beyond WHOIS

https://silsebastian.github.io

https://hub.docker.com/r/dianecode/whosedomain

Speaker: Silvia Sebastián
Authors: S. Sebastián, R. Diugan, J. Caballero, I. Sanchez-Rola, L. Bilge
BACKUP SLIDES

Indicators
Case Study: VT and Certs
Synonym Definition
Extraction
Ranking Case Study
Ranking Evaluation
Accuracy Evaluation
Incremental Evaluation
Content Attribution
Evaluation
<table>
<thead>
<tr>
<th>Indicator</th>
<th>Class</th>
<th>Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
<td>identity</td>
<td>NLP</td>
</tr>
<tr>
<td>organization</td>
<td>identity</td>
<td>NLP</td>
</tr>
<tr>
<td>personName</td>
<td>identity</td>
<td>NLP</td>
</tr>
<tr>
<td>email</td>
<td>contact</td>
<td>Regex</td>
</tr>
<tr>
<td>facebookHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>githubHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>instagramHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>linkedinHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>pinterestHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>skypeHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>telegramHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>twitterHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>whatsappHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>youtubeHandle</td>
<td>social</td>
<td>Regex</td>
</tr>
<tr>
<td>copyright</td>
<td>ipr</td>
<td>Regex</td>
</tr>
<tr>
<td>trademark</td>
<td>ipr</td>
<td>Regex</td>
</tr>
<tr>
<td>aboutUrl</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>contactUrl</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>fqdn</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>esld</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>privacyUrl</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>securityUrl</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>tosUrl</td>
<td>network</td>
<td>Regex</td>
</tr>
<tr>
<td>url</td>
<td>network</td>
<td>Regex</td>
</tr>
</tbody>
</table>

Table 5.1: Indicators extracted by WhoseDomain, their class, and whether extracted using a regular expression or NLP.
Case Study

- WHOIS is anonymized: “REDACTED FOR PRIVACY”
Case Study

- WHOIS is anonymized: “REDACTED FOR PRIVACY”.
- Guesser could not find an URL for this domain (404 NOT FOUND).
Case Study

- WHOIS is anonymized: “REDACTED FOR PRIVACY”.
- Guesser could not find an URL for this domain (404 NOT FOUND). VirusTotal return 44 subdomains.
Evaluation: Combining Sources

Case Study

- WHOIS is anonymized: “REDACTED FOR PRIVACY”.
- Guesser could not find an URL for this domain (404 NOT FOUND).
- VirusTotal return 44 subdomains.
- Certificates finds just in one of them the identity in the Subject CN.

acuityplatform.com

origin.acuityplatform.com

Acuity Ads, Inc.
This privacy policy governs your use of the software provided by Khushhal Maheshwari (‘khushhalm’, ‘we’, ‘us’, ‘ML QUIZ’)

1. Segment text into sentences.
2. Remove sentences without context terms:
   English: “we”, “us”, “provider” | Spanish: “a partir de ahora”, “con domicilio en”
3. Extract Nominal groups defining a grammar.
4. Dependency Parsing to find elements depending on context terms.
5. All nominal groups depending on the context terms or quoted synonyms are considered identities

Khushhal Maheshwari, khushhalm, ML QUIZ

GRAMMAR:

  {<JJ>+<NN.*>+}
  {<NN.*><CD>}
  {<NN.*>+}
## Content Attribution: Ranking

### (a) Winmer cluster

<table>
<thead>
<tr>
<th>Cluster 0: 22</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
</tr>
<tr>
<td>identity</td>
</tr>
<tr>
<td>identity</td>
</tr>
<tr>
<td>organization</td>
</tr>
<tr>
<td>email</td>
</tr>
<tr>
<td>esld</td>
</tr>
</tbody>
</table>

### (b) Expanded and filtered indicators

<table>
<thead>
<tr>
<th>Email</th>
<th><a href="mailto:cmgcopyright@cg.com">cmgcopyright@cg.com</a></th>
<th>['link', 'regexp']</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esld</td>
<td>cmg.com</td>
<td>['regexp']</td>
</tr>
<tr>
<td>Facebook Handle</td>
<td>countrylegends971</td>
<td>['link']</td>
</tr>
<tr>
<td>Identity</td>
<td>Google Analytics</td>
<td>['ner']</td>
</tr>
<tr>
<td>Identity</td>
<td>American Stock Exchange</td>
<td>['ner']</td>
</tr>
<tr>
<td>Identity</td>
<td>CMG</td>
<td>['dependency', 'ner']</td>
</tr>
<tr>
<td>Identity</td>
<td>CMG Affiliate</td>
<td>['dependency', 'ner']</td>
</tr>
<tr>
<td>Organization</td>
<td>Cox Media Group, Inc.</td>
<td>['ner']</td>
</tr>
<tr>
<td>Organization</td>
<td>Cox Media Group</td>
<td>['regexp']</td>
</tr>
<tr>
<td>Twitter Handle</td>
<td>Legends971</td>
<td>['link']</td>
</tr>
</tbody>
</table>

### (c) Clustered indicators

<table>
<thead>
<tr>
<th>Cluster 0: 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Organization</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Esld</td>
</tr>
<tr>
<td>Cluster 1: 3</td>
</tr>
<tr>
<td>Facebook Handle</td>
</tr>
<tr>
<td>Twitter Handle</td>
</tr>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Organization</td>
</tr>
<tr>
<td>Organization</td>
</tr>
</tbody>
</table>

### (d) Ranked clusters

<table>
<thead>
<tr>
<th>Cluster 0: 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Organization</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Esld</td>
</tr>
<tr>
<td>Cluster 1: 3</td>
</tr>
<tr>
<td>Facebook Handle</td>
</tr>
<tr>
<td>Twitter Handle</td>
</tr>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Identity</td>
</tr>
<tr>
<td>Organization</td>
</tr>
<tr>
<td>Organization</td>
</tr>
</tbody>
</table>

---

(c) 2021 Cox Media Group. All Rights Reserved.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>GT</th>
<th>Web</th>
<th>WhoseDomain</th>
<th>WHOIS</th>
<th>Certificates</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>tranco_top</td>
<td>0.96</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.64</td>
<td>0.77</td>
</tr>
<tr>
<td>tranco_100K</td>
<td>0.89</td>
<td>0.95</td>
<td>0.92</td>
<td>0.93</td>
<td>0.29</td>
<td>0.44</td>
</tr>
<tr>
<td>brands</td>
<td>0.93</td>
<td>1.00</td>
<td>0.96</td>
<td>0.91</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td>trackers</td>
<td>0.94</td>
<td>0.80</td>
<td>0.86</td>
<td>0.96</td>
<td>0.39</td>
<td>0.55</td>
</tr>
<tr>
<td>All</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.95</td>
<td>0.43</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 5.2: Attribution accuracy for WhoseDomain with all sources and ablation study using only one source at a time.
Table 5.3: Attribution accuracy increase as expansions are incrementally added from left (WHOIS only) to right (all expansions).
<table>
<thead>
<tr>
<th>Ranking All</th>
<th>Ranking Winners</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind.</td>
<td>Iden.</td>
<td>Clust.</td>
</tr>
<tr>
<td>6,386</td>
<td>2,325</td>
<td>2,102</td>
</tr>
</tbody>
</table>

Table 5.4: Document ranking evaluation on 100 GT privacy policies. Ind. are all indicators, Iden. identities, and Clust. clusters.
## Content Attribution Evaluation Details

<table>
<thead>
<tr>
<th></th>
<th>Regexp</th>
<th>NLP-NER</th>
<th>NLP-Synonym</th>
<th>DP-Metadata</th>
<th>DP-Schema</th>
<th>DP-Link</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Before Ranking</strong></td>
<td>1,069 (71)</td>
<td>2,377 (2,255)</td>
<td>110 (110)</td>
<td>117 (10)</td>
<td>87 (17)</td>
<td>2,494 (0)</td>
</tr>
<tr>
<td><strong>After Ranking</strong></td>
<td>326 (65)</td>
<td>625 (609)</td>
<td>104 (104)</td>
<td>83 (7)</td>
<td>47 (12)</td>
<td>1,138 (0)</td>
</tr>
</tbody>
</table>

Table 5: Number of indicators (identity indicators in parenthesis) extracted by the different content attribution techniques on the dataset of 100 GT privacy policies.

<table>
<thead>
<tr>
<th>GT</th>
<th>Content</th>
<th>Regexp</th>
<th>NLP-NER</th>
<th>NLP-Synonym</th>
<th>DP-Metadata</th>
<th>DP-Schema</th>
<th>DP-Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>tranco_top</td>
<td>250</td>
<td>42</td>
<td>26</td>
<td>8</td>
<td>0</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>tranco_100k</td>
<td>250</td>
<td>151</td>
<td>118</td>
<td>14</td>
<td>1</td>
<td>54</td>
<td>43</td>
</tr>
<tr>
<td>brands</td>
<td>100</td>
<td>80</td>
<td>74</td>
<td>19</td>
<td>3</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>trackers</td>
<td>139</td>
<td>82</td>
<td>69</td>
<td>40</td>
<td>11</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>739</td>
<td>355</td>
<td>287</td>
<td>81</td>
<td>15</td>
<td>132</td>
<td>87</td>
</tr>
</tbody>
</table>

Table 6: Number of indicators extracted by the different content attribution techniques that end up in the final attribution results for each GT dataset.