

The 17th International Conference on Emerging Security Information,
Systems and Technologies
SECURWARE 2023

September 25, 2023 to September 29, 2023 - Porto, Portugal

Science-Tracker Fingerprinting with Uncertainty: Selected Common Characteristics of Publishers from Network to Application Trackers on the Example of Web, App and Email

Stefan Kiltz, Robert Altschaffel, Jana Dittmann
Email: sec-by-design <at> iti.cs.uni-magdeburg.de



The research from Robert Altschaffel is partly funded by the research project "CyberSec LSA_OVGU-AMSL, Security-by-Design-Orchestration_Booster" under the Grant Number ZS/2015/12/96222.

Researcher at Advanced Multimedia and Security Lab (AMSL), Otto-von-Guericke-University of Magdeburg, Germany

AMSL - Research fields and interests:

- University diploma in computer science in 2006 and his Ph.D. in 2020 at Otto-von-Guericke-University, Magdeburg, Germany
- Work at numerous research projects at local, national and international level since 2005 (teaching since 2006)
- Main research interests: IT-security in general, data protection/privacy issues
- Special interest: digital forensics (process modelling, involved data stream and processing methods)
- Broad coverage of application fields: from whole data centres to individual processor-controlled components
- <https://omen.cs.uni-magdeburg.de/itiamsl/english/home/home.html>

- Introduction
- State of the State / Fundamentals
- Concept of Science-Tracking Fingerprint (STF) with intra- and inter-application assessment using URL and tracking similarities
- Exemplary implementation of the Science-Tracking Fingerprint (STF)
- Evaluation
- Conclusion and future work

- Science-tracking is common practice
- Known impacts ranging from:
 - general privacy violation
 - Data misuse and academic espionage [1]
 - concrete and grave risks to scientists [3]
- Usage of methods from IT-forensics for carefully designed, measured and systematic approaches and estimation of error, loss and uncertainty [5], [6]
- Especially in crime scene forensics: individualization of traces for attribution of actions to entities [7]
- Our goal for science-tracking:
 - looking for hints/leads towards attribution of the science-tracker by means of a Science-Tracking Fingerprint (STF)
 - Enable comparison of science-tracking mechanisms in different services – e.g., selected here: web, app and email
 - Easier access/recognition of changes in tracking over time
 - more effective countermeasures for specific science-trackers
- Caution – no full automatic process (certainly not at this time and not in future)

- User-tracking in general is studied quite extensively (e.g., [8], [9])
- Science-tracking as subset is also part of active research (e.g., [10])
- Primarily, determination of the **extent of user tracking** and resulting **consequences for the tracked** subjects as research goal of existing studies
- Also, employment of IT-forensic techniques to reveal data tracking (e.g.,[4])
- However, according to our knowledge, no publicized fingerprint tracking parties in order to obtain hints/leads toward individualization and attribution of the tracking party

- DCEA Forensic model [6] for our model-based approach provides (among others):
 - Forensic data streams (we use mass-storage DS_T and network DS_N)
 - Forensic data types (we use details about data DT_3 and communication protocol data DT_5)
- Properties of tools used for examination:
 - Operate on the application fields of web, app or email
 - Existing tools with one exception: RA_email_forensics (available as OpenSource on request)
 - Acquire, investigate and analyze URL (DT_5) and tracker data (DT_3) on DS_T and DS_N
 - Off-premises (hosted by a 3rd party) or on-premises (hosted by the examiners)
 - Static operation S (discrete snapshot) or dynamic operation D (continuous examination)

- Data sources (available on request):

Application area	Publisher	Data source	Date	Additional information
Web	ACM	https://dl.acm.org	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacy score on 16/06/2023
	Elsevier	https://www.elsevier.com	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacy score on 16/06/2023
	IEEE	https://www.ieee.org	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacy score on 16/06/2023
	SN-MME	https://www.springernature.com/de/macmillaneducation	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacy score on 16/06/2023
App	ACM	ACM TechNews V1.4.6	31/05/23 12:53	Store URL: https://play.google.com/store/search?q=ACM%20Technews&c=apps
	Elsevier	eReader V8.0.0	31/05/23 11:08	Store URL: https://play.google.com/store/apps/details?id=com.impelsys.elsapac.android.ebookstore
	IEEE	MyXplore V4.0.4	31/05/23 15:21	Store URL: https://play.google.com/store/apps/details?id=org.ieee.mobile.pubs.myxplore
	SN-MME	Macmillan Education eReader V1.0.8.18	01/06/23 09:22	Store URL: https://play.google.com/store/apps/details?id=com.macmillaneducation.ereader
Email	ACM	Subject: Publish Your Work in the ACM Journal on Responsible Computing (JRC)	29/09/22 16:00	Sender: "Kenneth R. Fleischmann, ACM JRC Editor-in-Chief (do not reply)" <call-for-papers@hq.acm.org>
	Elsevier	Subject: Programme announced Register now to join our expert speakers	27/10/22 18:42	Sender: AI in Aging and Age-related Diseases 2022 <conferences@author.email.elsevier.com>
	IEEE	Subject: July 2022 Issue of IEEE Signal Processing w/Content Gazette Is Now Available	09/08/22 17:00	Sender: IEEE Signal Processing Magazine <ieee-pubs@deliver.ieee.org>
	SN-MME	Subject: Springer Nature Editorial Newsletter May 2022	12/05/22 13:11	Sender: Springer Nature <springernature@newsletter.springernature.com>

- Digital evidence accompanied by **uncertainty** [5], e.g., if tool returns different results in repeated runs with the same input [6]
- Uncertainty encountered regarding (DT₅) and tracker data (DT₃)
- URL data retrieved by tools based on the Domain Name System (DNS) can contain A-Records or CNAME data or a combination thereof [24]:
 - A-Record: host address
 - CNAME: canonical name of alias to the host
- Encountered name resolutions with:
 - Multiple A-Records,
 - Multiple CNAMEs,
 - Combinations of (multiple) A-Records and CNAMEs
- CNAMEs for benign reasons but also to disguise 3rd party involvement (CNAME cloaking, see e.g., [25])

intra- and inter-application assessment

- Science-Tracking Fingerprint (STF) as result orchestrated tool usage whilst adhering to the model-based approach from [6]
- SFT with **semantic** and **syntactic** components formed as an evaluation of tool results when:
 - Accessing the publisher's website
 - Opening an app available from the publisher
 - Processing a publisher's email
- **Certainty** as part of STF semantics when comparing tool results as:
 - Plausible (pl): all tools return the same or comparable result
 - Uncertain (unc): at least one tool with diverging result
 - Non-match (-): no tool returns a meaningful result
- Formalization in the style of Backus-Naur Form (BNF) as matrix of cells
- Each cell with semantics of:
 - Counter: Number of occurrences
 - Certainty: plausible | uncertain | non-match
 - Data stream: Mass-storage (T) | Network (N)
 - Data type: DT₅ (URL) | DT₃ (Detected as Tracker)
 - Discovery mode: list-based (L) and/or manual (M)

intra- and inter-application assessment

- BNF-style representation of the STF

$\langle \text{MATRIX} \rangle ::= \langle \text{ROW} \rangle \ \&\& \ \langle \text{MATRIX} \rangle$

$\langle \text{ROW} \rangle ::= \langle \text{CELL} \rangle \mid /0/ \ \langle \text{CELL} \rangle \mid /0/ \ \langle \text{CELL} \rangle \mid /0/ \ \langle \text{CELL} \rangle \mid /0/$

$\langle \text{CELL} \rangle ::= \langle \text{Counter} \rangle \ \langle \text{EXPR} \rangle$

$\langle \text{EXPR} \rangle ::= \langle \text{EXPR1} \rangle \mid \langle \text{EXPR} \rangle; \langle \text{EXPR1} \rangle$

$\langle \text{EXPR1} \rangle ::= \langle \text{CERTAINTY} \rangle, \langle \text{DATASTREAM} \rangle, \langle \text{DATATYPE} \rangle \mid$
 $\langle \text{CERTAINTY} \rangle, \langle \text{DATASTREAM} \rangle, \langle \text{DATATYPE} \rangle, \langle \text{DISCOVERYMODE} \rangle$

intra- and inter-application assessment

- Discovery mode only relevant for tracker detection (DT₃)
- Full URL data from dynamic investigation D, name resolution from local DNS client
- Dynamic URL investigation contains IP, A-Record, CNAME
- Characteristic particular arrangement of CNAME usage (1st and 3rd Party)
- Quantifiable and qualitative differences between individual publishers
- STF as similarity measure - changes in the application field (Web, App, Email) are to be expected

Concept of Science-Tracking Fingerprint (STF)

intra- and inter-application assessment

- **Syntactically** element/value pairs are formed,
- concatenation forms the matrix as shown below:

	A-Record 1st party	CNAME 1st Party	A-Record Third Party	CNAME Third Party
Web	BNF cell	BNF cell	BNF cell	BNF cell
	BNF cell	BNF cell	BNF cell	BNF cell
App	BNF cell	BNF cell	BNF cell	BNF cell
	BNF cell	BNF cell	BNF cell	BNF cell
Email	BNF cell	BNF cell	BNF cell	BNF cell
	BNF cell	BNF cell	BNF cell	BNF cell

BNF Cell: 0 | Counter_{<pl | unc | none>, <N | T>, <DT5 | DT3>, <M | L>}

- Row contains BNF-style cells where occurrences are counted if:
 - Matching certainty per cell
 - Tracker certainty plausible | uncertain
- Cell entries duplicated if CNAMES are present in 1st and/or 3rd party

Concept of Science-Tracking Fingerprint (STF)

intra- and inter-application assessment

- Properties of all tools used as shown below:

Application	Tool	Location	Input data stream	Input data type	Output data stream	URL output (DT ₅)	IP output (DT ₅)	Tracker Output (DT ₃) based on external data	external data
Web	Privacyscore	Off-premises	DS _N	(DT ₁)	DS _T	x	N/A	x	easelist.to
	Webbkoll	Off-premises	DS _N	(DT ₁)	DS _T	x	x	x	disconnect
	Website Evidence Collector	On-premises	DS _N	DT ₁	DS _T	x	N/A	N/A	N/A
	Wireshark	On-premises	DS _N	DT ₁	DS _T	x	x	N/A	N/A
App	Exodus-Standalone	On-premises	DS _T	DT ₁	DS _T	x	N/A	x	exodus
	AppChecker	On-premises	DS _T	DT ₁	DS _T	x	N/A	x	AppAuthor's list
	Wireshark	On-premises	DS _N	DT ₁	DS _T	x	x	N/A	N/A
Email	emlAnalyze	On-premises	DS _T	DT ₁	DS _T	x	N/A	x	N/A
	RA_email_forensics	On-premises	DS _T	DT ₁	DS _T	x	N/A	x	N/A
	Wireshark	On-premises	DS _N	DT ₁	DS _T	x	x	N/A	N/A

- All tools operate on raw data DT₁
- DT₁ and other intermediate results inaccessible for off-premises tools
- CNAME data only acquired by wireshark [14] (dynamic investigation D)
- Wireshark and webkoll [12] als acquire IP, helpful in finding matches
- List-based tracker detection relies on external data
- External data change over time -> difficulties to repeatability for off-premises tools

Concept of Science-Tracking Fingerprint (STF)

intra- and inter-application assessment

- Table-based Intra- / Inter-Application comparison (Web, App, Email)

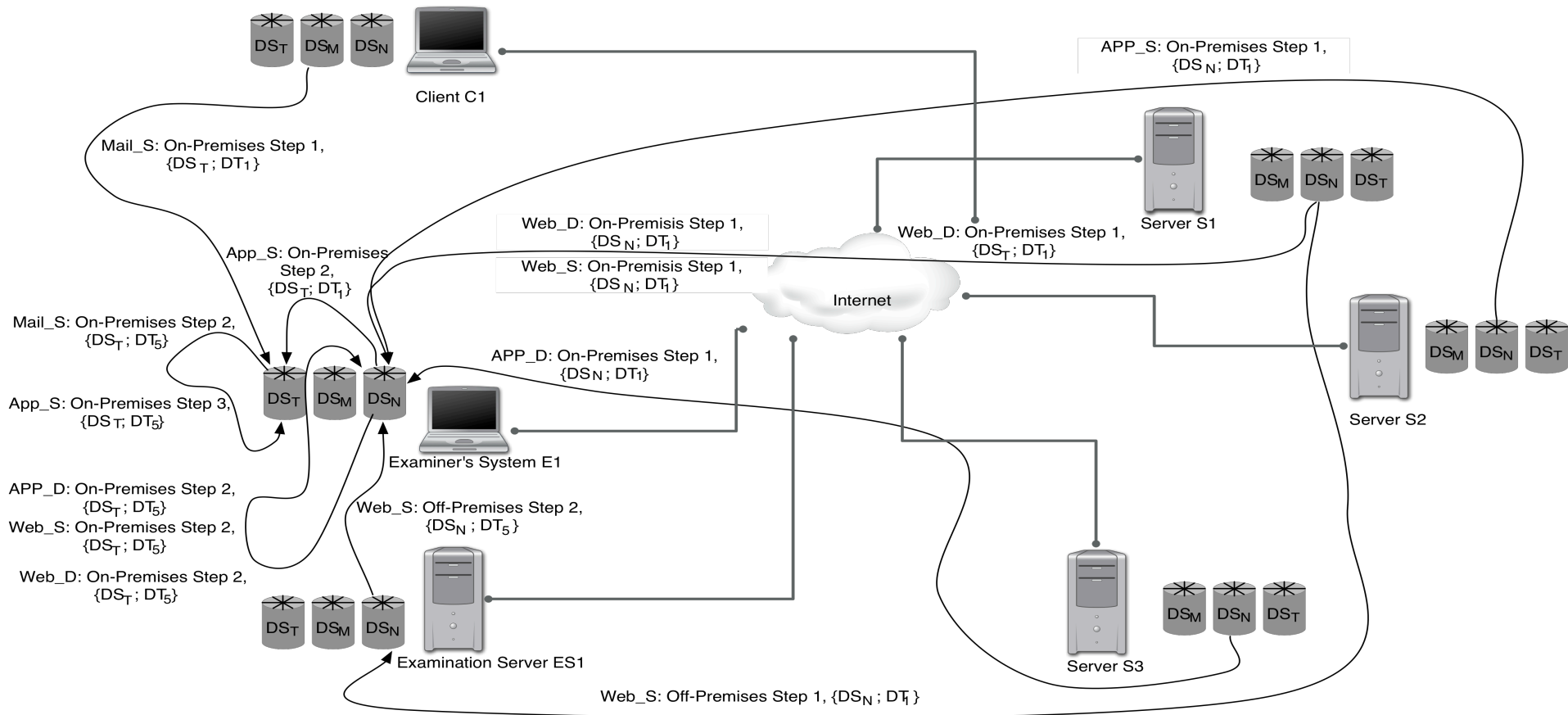
Static examination					Dynamic examination				
Off-Premesis					On-Premesis				
Privacyscore Web_S: Off Premesis $\{DS_N, DT_5, DT_3\}$		Webbkoll Web_S: Off Premesis $\{DS_N, DT_5, DT_3\}$			Website Evidence Collector_S: On Premesis $\{DS_N, DT_5\}$		Wireshark Web_D: On Premesis $\{DS_N, DT_5, DT_3\}$		Detailed Intra-Application Test result (Comparison)
3 rd Parties	Tracker Requests	Domain/ Host	IP	Detected as Tracker	Third party hosts	IP	Address (based on A-Record [A] Or CNAME [C])	Intra- Application DT ₅ match	Intra- Application DT ₃ Known tracker match

- One table for each application field, using certainty categories
- Aggregation of URL data (DT₅) and tracker detection (DT₃) per publisher
- Best-fit approach for Intra- / Inter-Application comparison for (DT₅)
- Inter-Application matches as hints for cross-application tracking (shared URL and tracker channels)

Concept of Science-Tracking Fingerprint (STF)

intra- and inter-application assessment

- System landscape analysis for an understanding of the opportunities and limitations of forensic examination results



- Simplified landscape to depict connections between components of interests and data flows

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- Exemplary selected **publishers** for the STF research chosen from a variety:
 - ACM
 - Elsevier
 - IEEE
 - Springer Nature Macmillan Education
- For Inter-application research selection of fitting smartphone apps and website data
- General principle: Choose websites/apps that are likely to be used by scientists for literature research (digital libraries)
- More detail on the following table

Exemplary implementation of the Science-Tracking Fingerprint (STF)

Application area	Publisher	Data source	Date	Additional information
Web	ACM	https://dl.acm.org	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacyscore on 16/06/2023
	Elsevier	https://www.elsevier.com	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacyscore on 16/06/2023
	IEEE	https://www.ieee.org	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacyscore on 16/06/2023
	SN-MME	https://www.springernature.com/de/macmillaneducation	06/06/23	due to unavailable 3rd party and tracker detection re-run for privacyscore on 16/06/2023
App	ACM	ACM TechNews V1.4.6	31/05/23 12:53	Store URL: https://play.google.com/store/search?q=ACM%20Technews&c=apps
	Elsevier	eReader V8.0.0	31/05/23 11:08	Store URL: https://play.google.com/store/apps/details?id=com.impelsys.elsapac.android.ebookstore
	IEEE	MyXplore V4.0.4	31/05/23 15:21	Store URL: https://play.google.com/store/apps/details?id=org.ieee.mobile.pubs.myxplorer
	SN-MME	Macmillan Education eReader V1.0.8.18	01/06/23 09:22	Store URL: https://play.google.com/store/apps/details?id=com.macmillaneducation.ereader
Email	ACM	Subject: Publish Your Work in the ACM Journal on Responsible Computing (JRC)	29/09/22 16:00	Sender: "Kenneth R. Fleischmann, ACM JRC Editor-in-Chief (do not reply)" <call-for-papers@hq.acm.org>
	Elsevier	Subject: Programme announced Register now to join our expert speakers	27/10/22 18:42	Sender: AI in Aging and Age-related Diseases 2022 <conferences@author.email.elsevier.com>
	IEEE	Subject: July 2022 Issue of IEEE Signal Processing w/Content Gazette Is Now Available	09/08/22 17:00	Sender: IEEE Signal Processing Magazine <ieee-pubs@deliver.ieee.org>
	SN-MME	Subject: Springer Nature Editorial Newsletter May 2022	12/05/22 13:11	Sender: Springer Nature <springernature@newsletter.springernature.com>

Exemplary implementation of the Science- Tracking Fingerprint (STF)

- Exemplary chosen **platform**: instantiation of the Examiner's System E1 of the exemplary system landscape (Lenovo E15, Intel Core i7-1255U, 16GB RAM, 256GB SSD)
- Low-noise (especially regarding network traffic) Debian 11 [21] base system
- Exemplary chosen **tools** for investigation:
 - Ungoogled chromium browser V. 95.0.4638.54 [23] for off-premises investigations (privacyscore [11] and webbkoll V. ec39808 [12]) and manual tracker verification for the Email application field
 - On-premises static web investigation: website evidence collector V1.0.0 [13]
 - On-premises dynamic web investigation: wireshark V. 3.4.10 [14]
 - On-premises static app investigation: exodus privacy V 1.3.1 [16] and appchecker V. 2020.05 [17]
 - On-premises dynamic app investigation as virtualized Android 7.1 [19] environment run inside Virtual Box 6.1 [20] and wireshark V. 3.4.10 [14] on the host-side
 - On-premises static email investigation using emlAnalyzer (unversioned) [22] and self-implemented RA_email_forensics V 0.5 (available on request)
 - On-premises dynamic email investigation by following embedded links: wireshark V. 3.4.10 [14]

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- Intra- application comparison Table on the example of the IEEE publisher for the Web application field (others available as open data on request):

Application Web									
Static examination					Dynamic examination				
Off-Premesis					On-Premesis				
Privacyscore Web_S: Off Premesis {DS _N , DT _s , DT ₃ }		Webbkoll Web_S: Off Premesis {DS _N , DT _s , DT ₃ }			Website Evidence Collector_S: On Premesis {DS _N , DT ₃ }	Wireshark Web_D: On Premesis {DS _N , DT _s , DT ₃ }			Detailed Intra-Application Test result (Comparison)
3 rd Parties	Tracker Requests	Domain/Host	IP	Detected as Tracker	3 rd party hosts	IP	Address (based on A-Record [A] or CNAME [C])	Intra-Application DT _s match	Intra-Application DT ₃ known tracker match
insight.adsrvr.org	x	insight.adsrvr.org	15.197.193.217	x	insight.adsrvr.org	52.223.40.198	insight.adsrvr.org [A]	pl	pl
js.adsrvr.org	x	js.adsrvr.org	18.165.129.129	x	js.adsrvr.org	18.64.82.184	js.adsrvr.org [A]	pl	pl
						23.54.103.160	www.ieeee.org [A]	unc	-
s3.amazonaws.com		s3.amazonaws.com	52.216.38.0		s3.amazonaws.com	52.216.44.24	s3.amazonaws.com [A]	pl	-
s3-us-west-2.amazonaws.com		s3-us-west-2.amazonaws.com	52.218.217.72		s3-us-west-2.amazonaws.com	52.218.245.56	s3-us-west-2.amazonaws.com [A]	pl	-
cdnjs.cloudflare.com	x	cdnjs.cloudflare.com	2a06:98c1:3123:e000::		cdnjs.cloudflare.com	104.17.25.14	cdnjs.cloudflare.com [A]	pl	unc
4490791.fl.s.doubleclick.net	x	4490791.fl.s.doubleclick.net	142.250.74.166	x	4490791.fl.s.doubleclick.net	142.250.181.198	4490791.fl.s.doubleclick.net [A]	pl	pl
googleads.g.doubleclick.net	x	googleads.g.doubleclick.net	2a00:1450:400f:80b::2002	x	googleads.g.doubleclick.net	142.251.209.130	googleads.g.doubleclick.net [A]	pl	pl
stats.g.doubleclick.net	x	stats.g.doubleclick.net	2a00:1450:4010:c0d::9c	x	stats.g.doubleclick.net	142.250.147.155	stats.g.doubleclick.net [A]	pl	pl
						142.250.181.194	adservice.google.de [A]	unc	-
connect.facebook.net	x	connect.facebook.net	2a03:2880:f013:d:face:b00c:0:3	x	connect.facebook.net	157.240.223.15	connect.facebook.net [A]	pl	pl
www.facebook.com	x	www.facebook.com	2a03:2880:f113:81:face:b00c:0:25de	x	www.facebook.com	157.240.223.35	www.facebook.com [A]	pl	pl
kit.fontawesome.com		kit.fontawesome.com	2606:4700::6812:1734		kit.fontawesome.com	104.18.22.52	kit.fontawesome.com [A]	pl	-
adservice.google.com	x	adservice.google.com	2a00:1450:400f:801::2002	x	adservice.google.com	172.217.19.66	adservice.google.com [A]	pl	pl
www.google.com	x	www.google.com	2a00:1450:400f:802::2004	x	www.google.com	142.250.181.196	www.google.com [A]	pl	pl
www.google.de	x	www.google.de	2a00:1450:400f:801::2003	x	www.google.de	142.251.209.131	www.google.de [A]	pl	pl
region1.google-analytics.com		region1.google-analytics.com	2001:4860:4802:32::36	x	region1.google-analytics.com	216.239.32.36	region1.analytics.google.com [A]	pl	unc
www.google-analytics.com	x	www.google-analytics.com	2a00:1450:400f:802::200e	x	www.google-analytics.com			unc	pl
www.googletagmanager.com	x	www.googletagmanager.com	2a00:1450:400f:803::2008		www.googletagmanager.com	142.250.181.200	www.googletagmanager.com [A]	pl	unc
					securesso.ieeee.org	140.98.193.42	securesso.ieeee.org [A]	unc	-
code.jquery.com		code.jquery.com	2001:4de0:ac18::1:a:2b		code.jquery.com	69.16.175.42	code.jquery.com [A]	pl	-
app-ab24.marketo.com	x	app-ab24.marketo.com	104.16.96.80	x	app-ab24.marketo.com	104.16.96.80	app-ab24.marketo.com [A]	pl	pl
munchkin.marketo.net	x	munchkin.marketo.net	23.61.220.209	x	munchkin.marketo.net			unc	pl
756-gph-899.mktorep.com	x	756-gph-899.mktorep.com	192.28.144.124		756-gph-899.mktorep.com			unc	unc
up.pixel.ad	x	up.pixel.ad	95.140.228.46		up.pixel.ad	178.79.242.16	up.pixel.ad [A]	pl	unc
di.ricdn.com	x	di.ricdn.com	35.244.174.68	x	di.ricdn.com	35.244.174.68	di.ricdn.com [A]	pl	pl
		6045067.global.siteimproveanalytics.io	18.185.183.56		6045067.global.siteimproveanalytics.io	18.185.183.56	6045067.global.siteimproveanalytics.io [A]	unc	-
		siteimproveanalytics.com	2606:4700:e4::ac40:ad0c		siteimproveanalytics.com	172.64.173.12	siteimproveanalytics.com [A]	unc	-
pixel.sitescout.com	x	pixel.sitescout.com	98.98.134.241	x	pixel.sitescout.com	98.98.134.243	pixel.sitescout.com [A]	pl	pl
tags.tiqcdn.com		tags.tiqcdn.com	2600:9000:2375:8a00:7:2bfb:7c00:93a1		tags.tiqcdn.com	18.64.79.94	tags.tiqcdn.com [A]	pl	-
www.youtube.com		www.youtube.com	2a00:1450:400f:80a::200e	x	www.youtube.com	142.251.209.142	www.youtube.com [A]	pl	unc

- Best-fit selection from the respective A-Records/CNAMEs for dynamic investigation, unavailable data marked in grey, empty tool results marked (-)

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- Aggregated Intra-/Inter application results for all 4 exemplary chosen publishers (highlighted entry covers previous detailed table):

	aggregated DT ₅ matches			aggregated DT ₃ matches		
	pl	unc	none	pl	unc	none
Intra-Web (ACM)	19	5	0	5	9	10
Intra-App (ACM)	0	8	0	2	0	6
Intra-Email (ACM)	3	1	0	1	0	3
Inter-Application (ACM)	0	1	35	1	0	35
Intra-Web (Elsevier)	5	12	0	2	4	11
Intra-App (Elsevier)	0	18	0	4	2	12
Intra-Email (Elsevier)	9	0	0	9	0	0
Inter-Application (Elsevier)	0	1	43	0	1	43
Intra-Web (IEEE)	23	8	0	15	6	10
Intra-App (IEEE)	0	15	0	1	3	11
Intra-Email (IEEE)	5	0	0	5	0	0
Inter-Application (IEEE)	2	3	46	1	3	47
Intra-Web (SN-MME)	11	11	0	2	6	14
Intra-App (SN-MME)	0	7	0	1	1	5
Intra-Email (SN-MME)	31	0	0	1	0	30
Inter-Application (SN-MME)	0	1	59	0	1	59




- Results form basis for STF, hints for potential cross-application tracking through inter-application matches (needs further investigation)

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- STF of the ACM publisher with semantics (BNF-style description) and syntactical vector (element-value pairs):

	A-Record 1 st party	CNAME 1 st Party	A-Record 3 rd Party	CNAME 3 rd Party
Web	0	0	3 _{PL,N,DT5;PL,N,DT3,L}	0
	0	0	2 _{PL,N,DT5;PL,N,DT3,L}	2 _{PL,N,DT5;PL,N,DT3,L}
	0	0	6 _{PL,N,DT5;UNC,N,DT3,L}	0
	0	0	2 _{PL,N,DT5;UNC,N,DT3,L}	2 _{PL,N,DT5;UNC,N,DT3,L}
	0	0	1 _{UNC,N,DT5;UNC,N,DT3,L}	1 _{UNC,N,DT5;UNC,N,DT3,L}
App	0	0	1 _{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}	1 _{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}
	0	0	1 _{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}	0
Email	0	0	1 _{PL,T,DT5;PL,T,DT3,M, 1PL,N,DT5;PL,N,DT3,M}	0



- Highlighted Entry: Example of fully plausible intra-app tool match
- Note: for each horizontal line the cell contains either a 0 or the detected key-value occurrences according to the presence of A-Records/CNAMEs from 1st and/or 3rd party

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- STF of the Elsevier publisher with semantics (BNF-style description) and syntactical vector (element-value pairs):

	A-Record 1 st party	CNAME 1 st Party	A-Record 3 rd Party	CNAME 3 rd Party
Web	0	0	2 _{PL,N,DT5;PL,N,DT3,L}	2 _{PL,N,DT5;PL,N,DT3,L}
	0	0	2 _{PL,N,DT5;UNC,N,DT3,L}	0
	0	0	1 _{PL,N,DT5;UNC,N,DT3,L}	1 _{PL,N,DT5;UNC,N,DT3,L}
	0	0	1 _{UNC,N,DT5;UNC,N,DT3,L}	0
App	4 _{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}	0	0	0
	2 _{UNC,T,DT5;UNC,N,DT3,L; UNC,N, DT5; UNC, S, DT3, L}	0	0	0
Email	1 _{PL,T,DT5;PL,T,DT3,M, 1PL,N,DT5;PL,N,DT3,M}	0	0	1 _{PL,T,DT5;PL,T,DT3,M, 1PL,N,DT5;PL,N,DT3,M}
	0	0	8 _{PL,T,DT5;PL,T,DT3,L, 8PM,N,DT5;PL,N,DT3,M}	8 _{PL,T,DT5;PL,T,DT3,M, 8PL,N,DT5;PL,N,DT3,M}

- Highlighted entry: Only uncertain results for app URL data and no detected app 3rd party involvement
- Note: for each horizontal line the cell contains either a 0 or the detected key-value occurrences according to the presence of A-Records/CNAMEs from 1st and/or 3rd party

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- STF of the IEEE publisher with semantics (BNF-style description) and syntactical vector (element-value pairs):

	A-Record 1 st party	CNAME 1 st Party	A-Record 3 rd Party	CNAME 3 rd Party
Web	0	0	8 _{PL,N,DT5;PL,N,DT3,L}	0
	0	0	5 _{PL,N,DT5;PL,N,DT3,L}	5 _{PL,N,DT5;PL,N,DT3,L}
	0	0	3 _{PL,N,DT5;UNC,N,DT3,L}	0
	0	0	2 _{PL,N,DT5;UNC,N,DT3,L}	2 _{PL,N,DT5;UNC,N,DT3,L}
	0	0	2 _{UNC,N,DT5;PL,N,DT3,L}	0
	0	0	2 _{UNC,N,DT5;UNC,N,DT3,L}	0
App	0	0	1 _{UNC,N,DT5;PL,N,DT3,L; UNC,T, DT5; PL, S, DT3, L}	0
	0	0	3 _{UNC,N,DT5;UNC,N,DT3,L; UNC,T, DT5; UNC, S, DT3, L}	0
Email	5 _{PL,T,DT5;PL,T,DT3,M} , 5 _{PL,N,DT5;PL,N,DT3,M}	0	0	5 _{PL,T,DT5;PL,T,DT3,M} , 5 _{PL,N,DT5;PL,N,DT3,M}

- Highlighted entry: manual detection based CNAMEs and [28] instead of visible signs
- Note: for each horizontal line the cell contains either a 0 or the detected key-value occurrences according to the presence of A-Records/CNAMEs from 1st and/or 3rd party

Exemplary implementation of the Science-Tracking Fingerprint (STF)

- STF of the Springernature-Macmillaneducation (SN-MME) publisher with semantics (BNF-style description) and syntactical vector (element-value pairs):

	A-Record 1 st party	CNAME 1 st Party	A-Record 3 rd Party	CNAME 3 rd Party
Web	0	0	1 _{PL,N,DT5;PL,N,DT3,L}	0
	0	0	1 _{PL,N,DT5;PL,N,DT3,L}	1 _{PL,N,DT5;PL,N,DT3,L}
	0	0	3 _{PL,N,DT5;UNC,N,DT3,L}	0
	0	0	3 _{PL,N,DT5;UNC,N,DT3,L}	3 _{PL,N,DT5;UNC,N,DT3,L}
App	0	0	1 _{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, N, DT3, L}	0
	0	0	1 _{UNC,T,DT5;UNC,T,DT3,L; UNC,N, DT5; UNC, N, DT3, L}	0
Email	0	0	1 _{PL,T,DT5;PL,T,DT3,ML, 1PL,N,DT5;PL,N,DT3,ML}	1 _{PL,T,DT5;PL,T,DT3,ML, 1PL,N,DT5;PL,N,DT3,ML}



- Highlighted entry: List based detection after code revision of RA_email_forensics software (open source, available on request)
- Note: for each horizontal line the cell contains either a 0 or the detected key-value occurrences according to the presence of A-Records/CNAMEs from 1st and/or 3rd party

- Research shows necessity to also look manually for semantic connections between A-Records and CNAMEs across 1st and 3rd parties based on domain-owners (often particular relevant for cross-application evaluations, e.g., web | app)
- Proposed usage of whois [26] queries and partially backing owners documentation [27] in similar URLs (e.g., www.google-analytics.com | firebase.google.com as part of publisher SN-MME investigation)
- Particular relevant for comparison from whois output:
 - Registrant
 - Admin
 - Tech
- Can uncover connections between seemingly sure mismatches (e.g., tiqcdn.com | tealium.com, as part of investigating the IEEE publisher)
- Ad cloud constituents with very dissimilar URLs (e.g., marketo.com | omtrdc.net | mktossl.com | mktoweb.com all part of Adobe Ad Cloud with matching whois entries, partly backed by [29] found during IEEE publisher investigation)
- However, comparison results based on the above generally marked as uncertain!

- Some tools (e.g. for tracker detection) with reliance on dynamically changing external data (e.g., lists such as [15]) with additions, modifications, deletions between t_i and t_{i+1}
- Requires on-premises tool access for repeatability of examination
- First results on 4 selected publishers on the application fields of web, app and Email to show easy individualization
- STF to provide leads for further investigation towards attribution
- STF also in support for energy efficiency research by revealing unnecessary (from the user's perspective) tracking
- However, STF alone **not** deemed **sufficient** for attribution!

Evaluation

- Additional evaluation over time for 3 selected call-for-paper Emails from the publisher IEEE:

Fingerprint a) Mail from August, 9th, 2022

	A-Record First party	CNAME First Party	A-Record Third Party	CNAME Third Party
Email	3 _{PL,T,DT5;PL,T,DT3,M} , 3 _{PL,N,DT5;PL,N,DT3,M}	0	0	3 _{PL,T,DT5;PL,T,DT3,M} , 3 _{PL,N,DT5;PL,N,DT3,M}

Fingerprint b) Mail from September,13th, 2022

	A-Record First party	CNAME First Party	A-Record Third Party	CNAME Third Party
Email	0	0	3 _{PL,T,DT5;PL,T,DT3,M} , 3 _{PL,N,DT5;PL,N,DT3,M}	3 _{PL,T,DT5;PL,T,DT3,M} , 3 _{PL,N,DT5;PL,N,DT3,M}

Fingerprint c) Mail from August, 28th, 2023

Email	5 _{PL,T,DT5;PL,T,DT3,M} , 5 _{PL,N,DT5;PL,N,DT3,M}	0	0	5 _{PL,T,DT5;PL,T,DT3,M} , 5 _{PL,N,DT5;PL,N,DT3,M}
	1 _{PL,T,DT5;PL,T,DT3,M} , 1 _{PL,N,DT5;PL,N,DT3,M}	0	0	1 _{PL,T,DT5;UNC,T,DT3,M} , 1 _{PL,N,DT5;UNC,N,DT3,M}

- Similarities: Usage of CNAMEs 1st party, A-Record 3rd party for 2 mails
- Differences: 1 mail completely 3rd party
- Notable observation: list-based tracker detection of 3rd party image links, though arguably added by conference organizers
- Positive news:** Research into Science-Tracking can lead to changes - **progressive privacy-preserving agreement** with a publisher [30] based on similar research

Conclusion and Future Work

- Science-Tracking Fingerprint (STF) as tool-independent similarity measure to individualize and compare science publishers and to give leads/hints towards attribution based on existing forensic model
- Idea: compare different tool results with certainty category by measuring aspects of Science-Tracking (URL data from 1st / 3rd party with A-Record/CNAME information, tool-based tracker detection)
- Semantic component based on BNF-style representation with syntactic element as element-value pairs
- Tested with 4 exemplary chosen publishers on the application areas of Web, App, Email and existing tools
- In addition, self-implemented Open Source RA_email_forensics tool adherent to existing forensic model
- Intra-/Inter-application matching to give hints/leads to potential cross-application tracking
- STF also supports energy efficiency evaluations for sustainability through analyzing data packets used for science tracking by the publishers
- **Future work:** Test on broad scale (supported by BNF-style representation), reduced reliance on 3rd parties to provide off-premises tools by using existing local installation for better repeatability and version control



Thank you very much for your attention!

(Reminder: our Open Source Software and our Open data
is available on email request to
sec-by-design <at> iti.cs.uni-magdeburg.de)

References

- [1] Deutsche Forschungsgemeinschaft, "Data tracking in research: aggregation and use or sale of usage data by academic publishers" [Online] https://www.dfg.de/download/pdf/foerderung/programme/lis/datentracking_papier_en.pdf (2023.09.05).
- [2] E. Bettinger, and M. Bursic, and A. Chandler, "Disrupting the Digital Status Quo: Why and How to Staff for Privacy in Academic Libraries" [Online] <https://publish.illinois.edu/licensingprivacy/files/2023/06/Whitepaper-on-Privacy-Staffing-Licensing-Privacy.pdf> (2023.09.05).
- [3] R. Siems, "When your journal reads you – user tracking on science publisher platforms", Elephant in the Lab. <https://doi.org/10.5281/zenodo.4683778>, 2021.
- [4] R. Altschaffel, and S. Kiltz, and T. Lucke, and J. Dittmann, "Introduction to Being a Privacy Detective: Investigating and Comparing Potential Privacy Violations in Mobile Apps Using Forensic Methods", in Proceedings of the Fourteenth International Conference on Emerging Security Information, Systems and Technologies (Securware), Valencia, Spain, 21-25/09/2020, ISBN 978-1-61208-821-1, pp 60-68, 2020.
- [5] E. Casey, "Error, Uncertainty and Loss in Digital Evidence", In International Journal of Digital Evidence, Volume 1, Issue 2, pp. 1-45, 2002.
- [6] S. Kiltz, "Data-Centric Examination Approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics", PhD Thesis, Faculty of Computer Science, Otto-von-Guericke-University Magdeburg, Germany, September, 2020.
- [7] K. Inman and N. Rudin, "Principles and Practises of Criminalistics: The Profession of Forensic Science", CRC Press LLC Boca Raton Florida, USA, ISBN 0-8493-8127-4, 2001.
- [8] W. Christl, "Corporate Surveillance in Everyday Life" [Online] https://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf (2023.09.05).
- [9] H. Mildebrath "Unpacking 'commercial surveillance': The state of tracking" [Online] [https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/739266/EPRS_BRI\(2022\)739266_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/739266/EPRS_BRI(2022)739266_EN.pdf) (2023.09.05).
- [10] C. Hanson, "User Tracking on Academic Publisher Platforms" [Online] <https://www.codyh.com/writing/tracking.html> (2023.09.05).
- [11] D. Herrmann, "Welcome - PrivacyScore"[Online] <https://privacyscore.org/> (2023.09.05).
- [12] Dataskydd.net Sverige, "Analyze | Webb koll - dataskydd.net" [Online] <https://webbkoll.dataskydd.net/en> (2023.09.05).
- [13] European Data Protection Supervisor, "EDPS Inspection Software | European Data Protection Supervisor" [Online] https://edps.europa.eu/edps-inspection-software_en (2023.09.05).
- [14] Wireshark Foundation, "Wireshark · Go Deep" [Online] <https://www.wireshark.org/> (2023.09.05).
- [15] Disconnect Inc., "GitHub - disconnectme/disconnect-tracking-protection: Canonical repository for the Disconnect services file" [Online] <https://github.com/disconnectme/disconnect-tracking-protection> (2023.09.05).

References

- [16] Exodus Privacy, "GitHub - Exodus-Privacy/exodus-standalone: exodus CLI client for local analysis" [Online] <https://github.com/Exodus-Privacy/exodus-standalone> (2023.09.05).
- [17] J. Alemann, and N. Baier, and M. Streuber, and T. Nam, and L. Peters, "GitHub - Tienisto/AppChecker" [Online] <https://github.com/Tienisto/AppChecker> (2023.09.05).
- [18] Exodus Privacy, "exodus" [Online] <https://reports.exodus-privacy.eu.org/en/trackers/> (2023.09.05).
- [19] cwhung, "Android-x86 - Porting Android to x86" [Online] <https://www.android-x86.org/> (2023.09.05).
- [20] Oracle Inc., "Oracle VM VirtualBox" [Online] <https://www.virtualbox.org/> (2023.09.05).
- [21] Software in the Public Interest, Inc. , "Debian -- News -- Debian 11 "bullseye" released" [Online] <https://www.debian.org/News/2021/20210814> (2023.09.05).
- [22] F. Wahl, "GitHub - wahlflo/eml_analyzer: A cli script to analyze an E-Mail in the EML format for viewing the header, extracting attachments, etc." [Online] https://github.com/wahlflo/eml_analyzer (2023.09.05).
- [23] ungoogled-chromium Authors, "GitHub - ungoogled-software/ungoogled-chromium: Google Chromium, sans integration with Google" [Online] <https://github.com/ungoogled-software/ungoogled-chromium> (2023.09.05).
- [24] P. Mockapetris, "Domain names - concepts and facilites" [Online] <https://datatracker.ietf.org/doc/pdf/rfc1034> (2023.09.05).
- [25] Palo Alto Networks, Inc., "CNAME Cloaking: Disguising Third Parties Through the DNS" [Online] <https://unit42.paloaltonetworks.com/cname-cloaking/> (2023.09.05).
- [26] M. d'Itri, "whois(1) — whois — Debian bullseye — Debian Manpages" [Online] <https://manpages.debian.org/bullseye/whois/whois.1.en.html> (2023.09.05).
- [27] Google Inc., "What is Google Analytics for Firebase? - Firebase Help" [Online] <https://support.google.com/firebase/answer/7388022?hl=EN> (2023.09.05).
- [28] Adobe Inc., "trackingServer | Adobe Analytics" [Online] <https://experienceleague.adobe.com/docs/analytics/implementation/vars/config-vars/trackingserver.html?lang=en-US> (2023.09.05).
- [29] Adobe Inc., "Get started with tracking | Adobe Campaign" [Online] <https://experienceleague.adobe.com/docs/campaign-classic/using/sending-messages/tracking-messages/about-message-tracking.html?lang=en> (2023.09.05)
- [30] Elsevier B.V., & MPDL Services gGmbH, Max Planck Society, "Projekt DEAL – Elsevier Publish and Read Agreement", doi:10.17617/2.3523659 [Online] https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3523659 (2023.09.14).