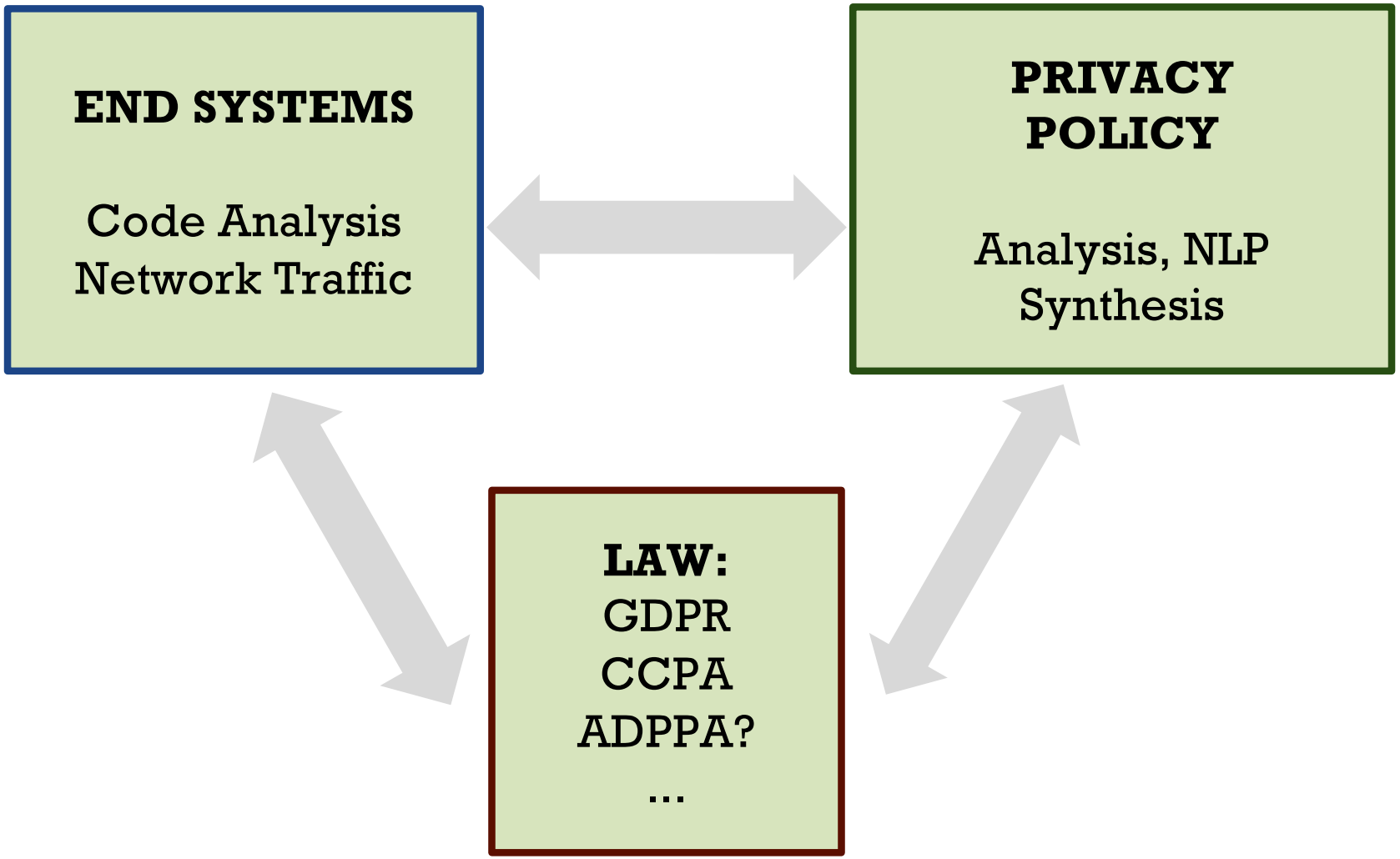
Using the CI tuple for Auditing Data Collection Practices (from the Edge)

Athina Markopoulou
Rahmadi Trimananda, Hao Cui
UC Irvine

UCI University of California, Irvine

ProperData
PROTECTING PERSONAL DATA FLOW ON THE INTERNET
http://properdata.eng.uci.edu

Problem Space



Need for unified/auditable specification: opportunity for CI tuple

Network Point of View

Application Domains

Web/Browsers

Mobile devices & apps

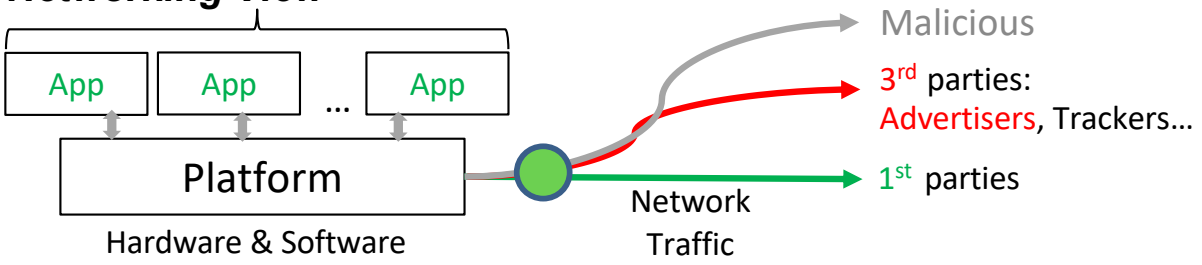
Smart TV & apps

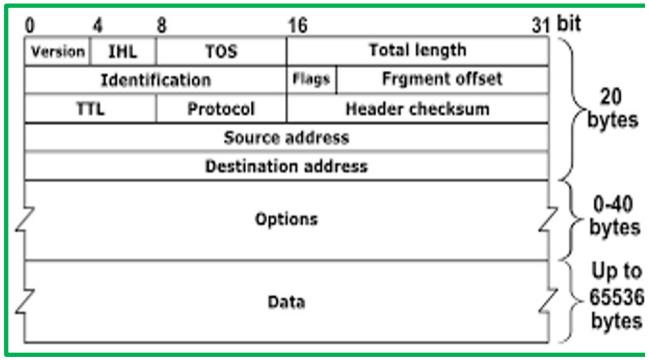
Smart Speakers
Voice Assistants

Smart Home
IoT

VR/AR devices
& apps

Networking View





Implementation Challenges:

- Capture packets in real-time; on-device, on the WiFi router, in the middle of a network
- Encryption→ visibility into protocols: IP, HTTP/HTTPS, DNS, TLS/SNI
- Exercise apps automatically, and at scale
- Low level → difficult to infer high level properties

Network Point of View

Application Domains

Web/Browsers

Mobile devices & apps

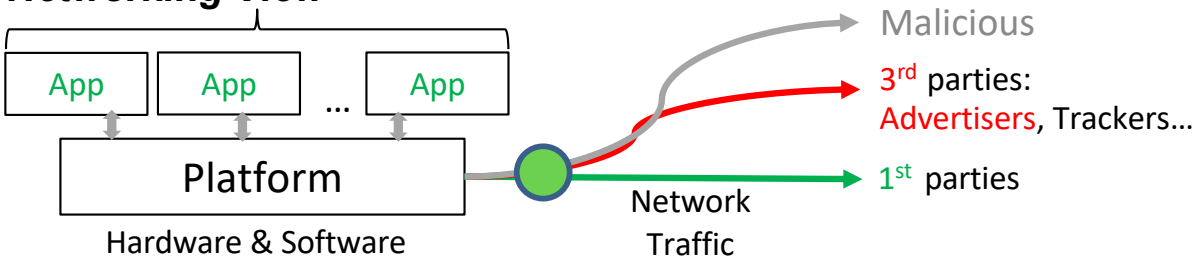
Smart TV & apps

Smart Speakers
Voice Assistants

Smart Home
IoT

VR/AR devices
& apps

Networking View



Goals:

- Diagnosis: who sent it (e.g., app, platform, SDK) what data type (e.g. which PII) goes, to what destination (e.g. ATS), for what purpose?
- Control: can we do something about it (block, obfuscate, add noise etc)?

End Systems/Networking View

CVInspector [NDSS'21]

AntMonitor
NoMoAds [PETS'19]
NoMoATS [PETS'20]

Mobile devices & apps

Smart TV & apps

Rokustic, Firetastic [PETS'20]
SmartTV Fingerprinting [PETS'22]

Your Echos are Heard, [2022]

PingPong [NDSS'20]

OVRSeen [SEC'22]

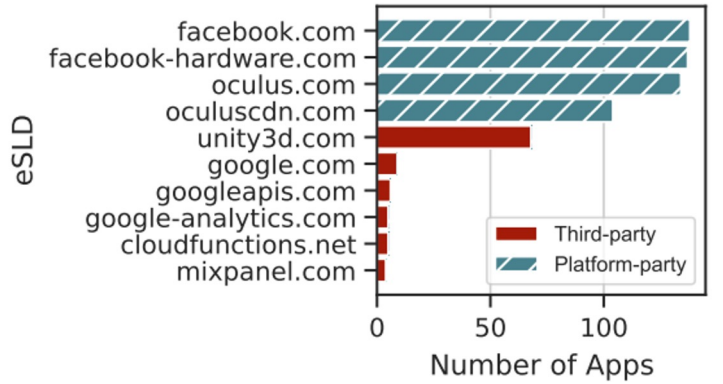
Example: Results from OVRseen

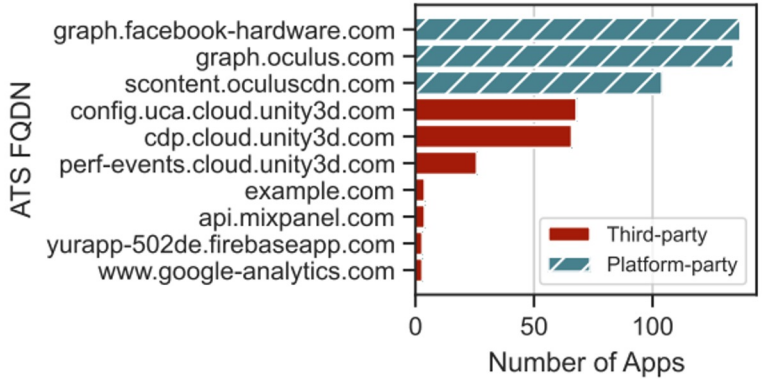
Trimananda et. al., SEC'22

Data type sent out

Data Types (21)	Apps			FQDNs			% Blocked		
PII	1st	3rd	Pl.	1st	3rd	Pl.	1st	3rd	Pl.
Device ID	6	64	2	6	13	1	0	38	100
User ID	5	65	0	5	13	0	20	38	-
Android ID	6	31	18	6	7	2	17	43	50
Serial Number	0	0	18	0	0	2	-	-	50
Person Name	1	7	0	1	4	0	0	50	-
Email	2	5	0	2	5	0	0	20	-
Geolocation	0	5	0	0	4	0	-	50	-
Fingerprint									
SDK Version	23	69	20	34	28	4	6	46	0
Hardware Info	21	65	19	25	23	3	4	39	33
System Version	16	62	19	20	21	3	5	43	33
Session Info	7	66	2	7	13	1	14	46	100
App Name	4	65	2	4	10	1	25	40	100
Build Version	0	61	0	0	3	0	-	100	-
Flags	6	53	2	6	8	1	0	50	100
Usage Time	2	59	0	2	4	0	0	50	-
Language	5	28	16	5	9	1	0	56	0
Cookies	5	4	2	5	3	1	0	33	100
VR Sensory Data									
VR Play Area	0	40	0	0	1	0	-	100	-
VR Movement	1	24	2	1	6	1	0	67	100
VR Field of View	0	16	0	0	1	0	-	100	-
VR Pupillary Distance	0	16	0	0	1	0	-	100	-
Total	33	70	22	44	39	5	5	36	20

Top -10 destinations





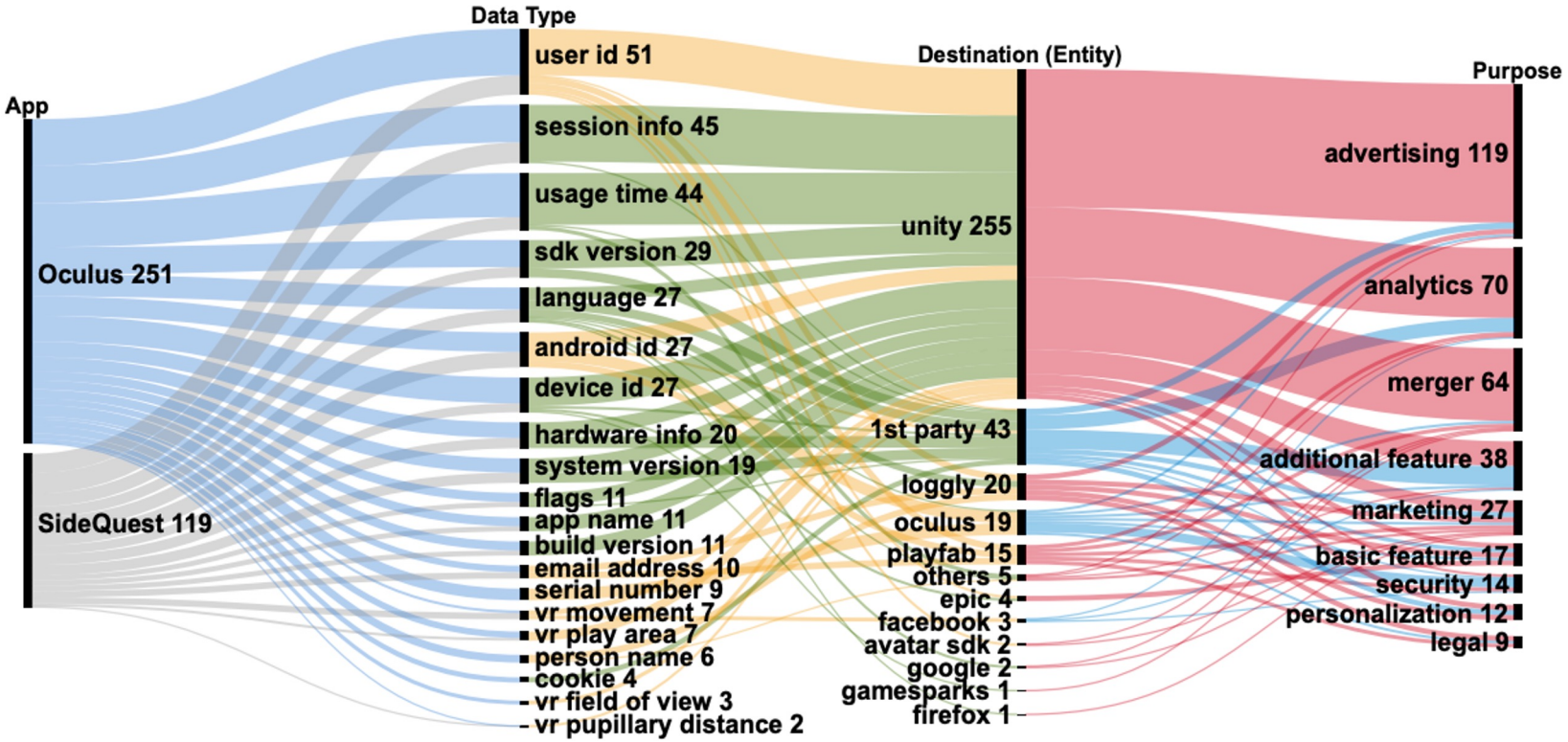
Centralized ecosystem: FB/Oculus, unity

Driven by tracking & social/analytics, not by ads

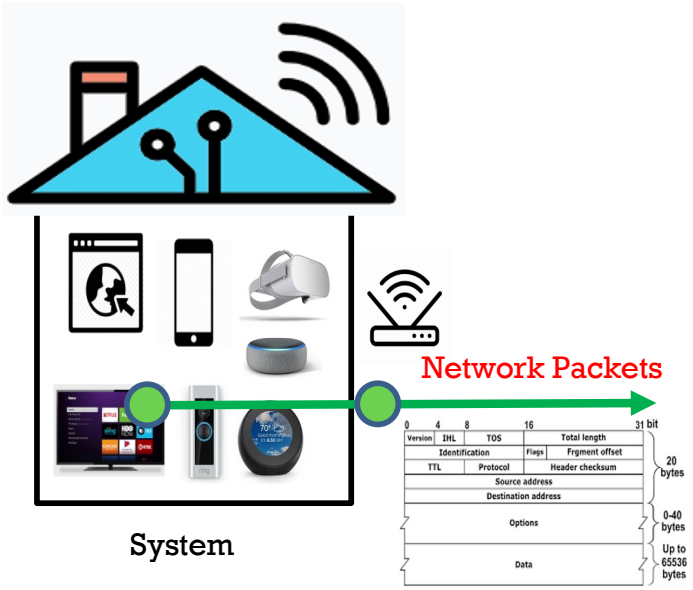
Example: Results from OVRseen

Trimananda et. al., SEC'22

- Purpose can be (partly) inferred from network data
  - Heuristics: key-value pairs (manually labeled purpose based on keys; "adid"→advertising, "passwd" → security etc); also looked at name of apk and compare destination; information about destination domain (organization/ATS: lookup DuckDuckGo, CrunchBase) [Mobipurpose,'19 Purpliance'21]
- Purpose Stated in Privacy Policy
  - Purpose from [Polisis'18] matched for data flows [Policheck] with consistent disclosures [OVRseen'22]



Directly extract CI params



CI tuple: (sender, recipient, data type; [subject]; (purpose; other TP))

"data flow" transmission principle

Sender:
- Application (dev)
- 3rd party library
- Platform, device
- Malware

Recipient:
- 1st, 3rd parties, platform, cloud
- Advertisers & trackers (ATS)
- Organization

Data Type:
- Personally Identifiable Information (PII)
- Fingerprinting
- Activity Data
- Sensor data

[Subject:]
- Typically the user of the app and platform

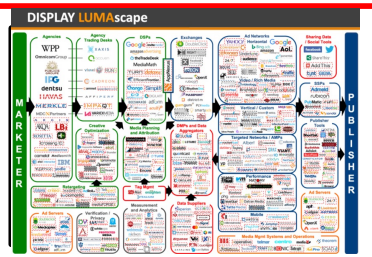
Purpose:
- Functionality
- Analytics
- Tracking
- Ads
- Personalization
- Security

Indirectly: infer tracking/targeting from the edge

2. Observe service on same or different platform

3. Infer Tracking+Targeting

Observe collection by 1st and 3rd parties



Sharing:
unobservable from the edge

amazon aws

Use: unobservable

1. Train Personnas

arXiv > cs > arXiv:2204.10920

Search...

Help | Advanced

Computer Science > Cryptography and Security

[Submitted on 22 Apr 2022 (v1), last revised 11 May 2022 (this version, v3)]

Your Echos are Heard: Tracking, Profiling, and Ad Targeting in the Amazon Smart Speaker Ecosystem

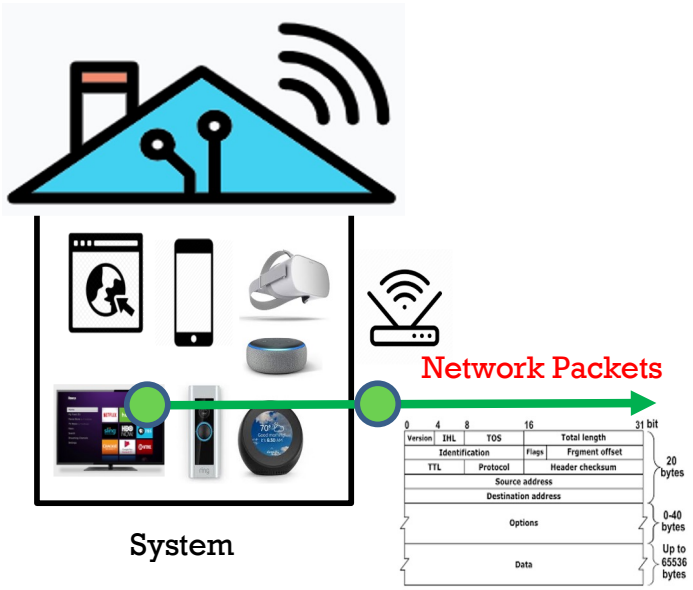
Umar Iqbal, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, Zubair Shafiq

We find that Amazon processes voice data to infer user interests and uses it to serve targeted ads on-platform (Echo devices) as well as off-platform (web). Smart speaker interaction leads to as much as 30X higher ad bids from advertisers. Finally, we find that Amazon's and skills' operational practices are often not clearly disclosed in their privacy policies.

Org.	Domains	Skills
Amazon	*(11).amazon.com	895
	prod.amcs-tachyon.com	305
	api.amazonalexa.com	173
	*(7).cloudfront.net	144
	device-metrics-us-2.amazon.com	123
	*(4).amazonaws.com	52
	acsechocaptiveportal.com	27
	fireoscaptiveportal.com	20
	ingestion.us-east-1.prod.arteries.alexa.a2z.com	7
	ffs-provisioner-config.amazon-dss.com	2
Skills	*(2).youversionapi.com	2
	static.garmincdn.com	1
Third party	dilillabs.com	9
	*(2).megaphone.fm	9
	cdn2.voiceapps.com	7
	*(2).podtrac.com	7
	*(2).pod.npr.org	4
	chtbl.com	3
	1432239411.rsc.cdn77.org	3
	*(2).libsyn.com	3
	*(3).streamtheworld.com	3
	discovery.meethue.com	2
	turnernetworksales.mc.tritondigital.com	1
	traffic.omny.fm	1

TABLE 1: Amazon, skill vendors, and third-party domains contacted by skills. "Org." column refers to organization. "Skills" column represents the count of skills. Advertising and tracking domains are shaded with grey. Subdomains counts are represented with *(#), e.g., *(11).amazon.com represents requests to 11 subdomains of amazon.com.

Auditing Network Traffic



CI tuple: (sender, recipient, data type; [subject]; (purpose; other TP))

"data flow" transmission principle

Sender:
- Application (dev)
- 3rd party library
- Platform, device
- Malware

Recipient:
- 1st, 3rd parties, platform, cloud
- Advertisers & trackers (ATS)
- Organization

Data Type:
- Personally Identifiable Information (PII)
- Fingerprinting
- Activity Data
- Sensor data

[Subject:]
- Typically the user of the app and platform

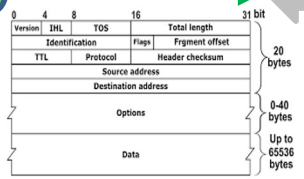
Purpose:
- Functionality
- Analytics
- Tracking
- Ads
- Personalization
- Security

`Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])`

Auditing Network Traffic vs. Privacy Policy

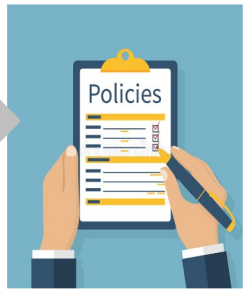
System

Network Packets



Consistent ?

Privacy Policy



CI tuple: (sender, recipient, data type; [subject]; (purpose; other TP))

"data flow" transmission principle

Sender:
- Application (dev)
- 3rd party library
- Platform, device
- Malware

Recipient:
- 1st, 3rd parties, platform, cloud
- Advertisers & trackers (ATS)
- Organization

Data Type:
- Personally Identifiable Information (PII)
- Fingerprinting
- Activity Data
- Sensor data

[Subject:]
- Typically the user of the app and platform

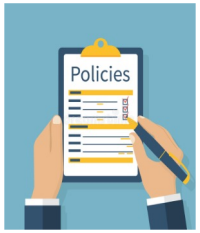
Purpose:
- Functionality
- Analytics
- Tracking
- Ads
- Personalization
- Security

Other Aspects:
- With Notice?
- With Consent?
- Consistent Disclosure?

Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])

Policy: (sender, destination=entity, data type, purpose, [other])

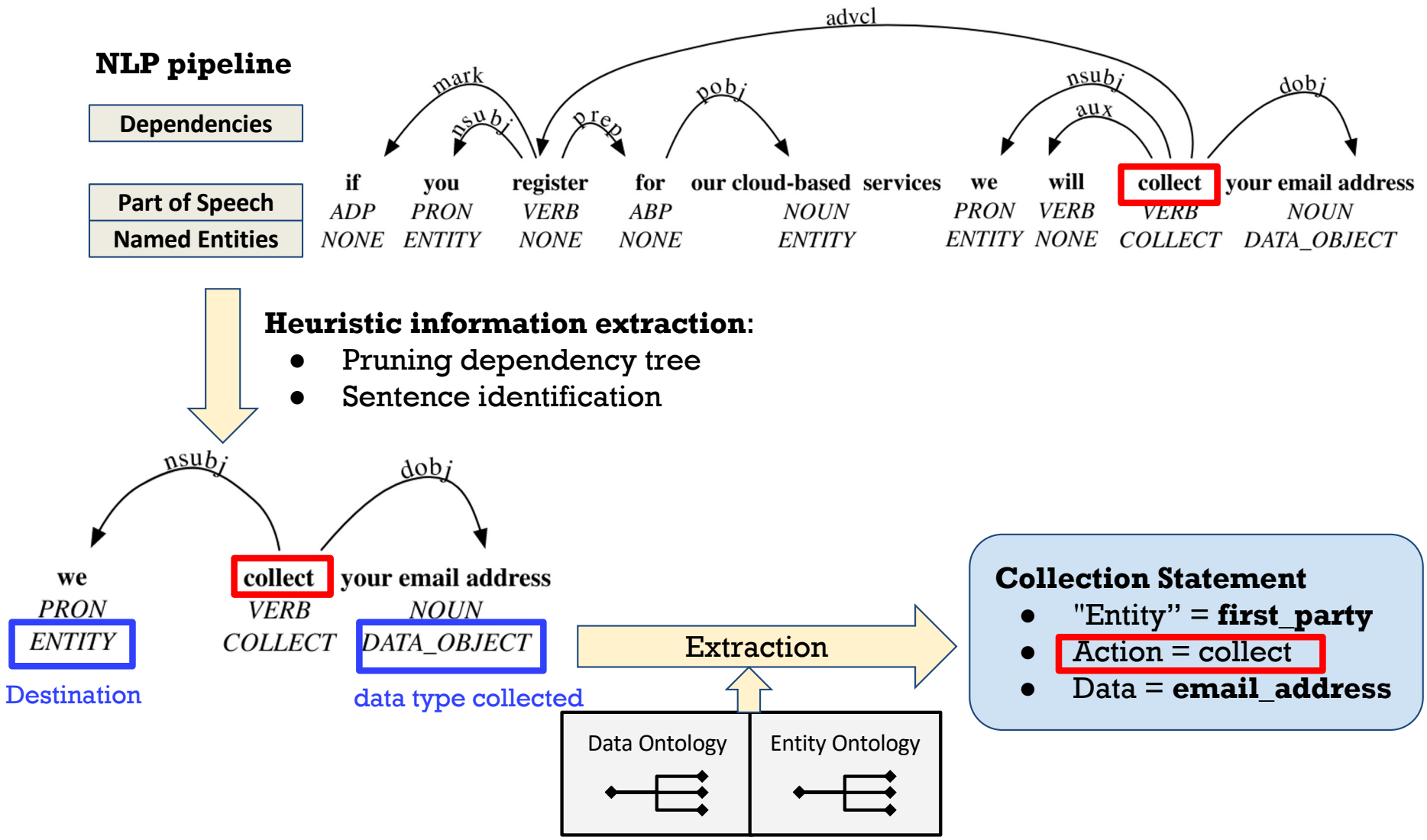
Privacy Policy Analysis



- Early trend: analysis by experts
  - within CI: [Shvartzshnaider et al. '19]

- Recent development: automated via NLP
- Today, NLP-based privacy policy analysis can successfully:
  - extract (data types, recipient=entity) [PolicyLint'19]
  - extract purpose [Polisys'18], [MobiPurpose'19], [Purpliance'21]
    - Although this is more tricky; and unclear how to connect to other parameters
  - check the consistency of network vs. policy side [PoliCheck'20, OVSeen'22]
    - Ontologies are necessary for that
  - be applied at scale to a large number of policies and application domains
    - Mobile, amazon skills, VR apps

Example: Extracting Collection Statements [PolicyLint+]



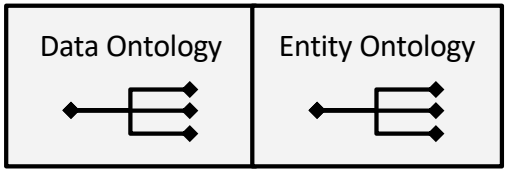
Collection Statements vs. Data Flows

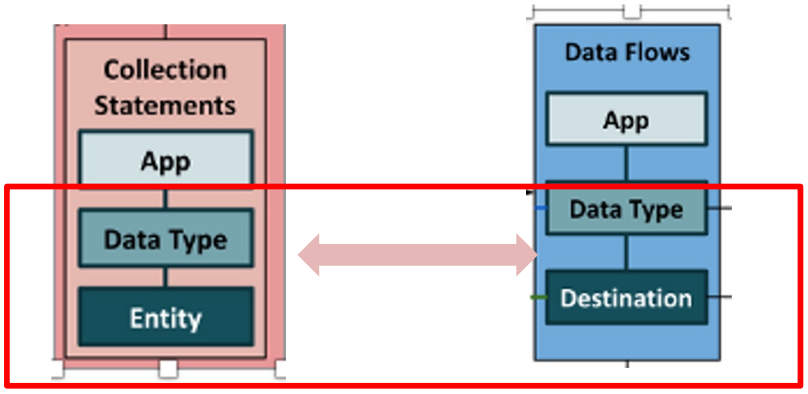
From privacy policy:
P=(data type, entity)

From network traffic:
F=(data type, destination)

	Disclosure Type	Privacy Policy Text	Action : Data Collection Statement (P)	Data Flow (F)
Consistent	Clear	"For example, we collect information ..., and *a timestamp for the request.*"	*collect* : ⟨*com.cvr.terminus, usage time, we*⟩	⟨*usage time, we*⟩
	Vague	"We will share *your information (in some cases personal information)* with third-parties, ..."	*collect* : ⟨*com.HomeNetGames.WW1oculus, pii, third party*⟩	⟨*serial number, oculus*⟩ ⟨*android id, oculus*⟩
Inconsistent	Omitted	-	*collect* : ⟨*com.kluge.SynthRiders, -, -*⟩	⟨*system version, oculus*⟩ ⟨*sdk version, oculus*⟩ ⟨*hardware information, oculus*⟩
	Ambiguous	"..., Skydance *will not disclose any Personally Identifiable Information to third parties ...* your *Personally Identifiable Information will be disclosed to such third parties* and ..."	*collect* : ⟨*com.SDI.TWD, pii, third party*⟩	⟨*serial number, oculus*⟩ ⟨*android id, oculus*⟩
	Incorrect	"We *do not share our customer's personal information with unaffiliated third parties ...*"	*not_collect* : ⟨*com.downpourinteractive. onward, pii, third party*⟩	⟨*device id, unity*⟩ ⟨*user id, oculus*⟩

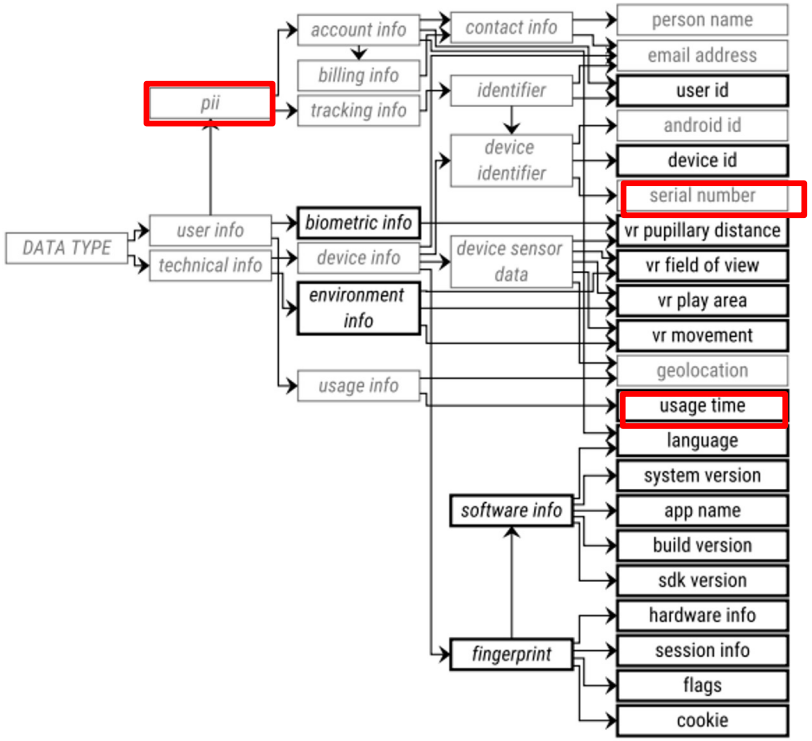
Ontologies taken into account when checking for consistency of collection statements vs. data flows



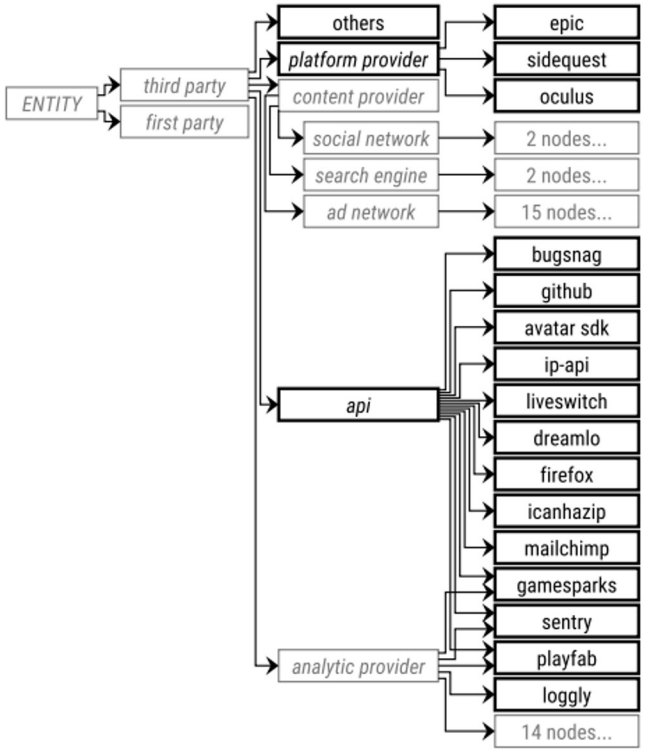


Example: VR Ontologies

Data Types Collected



Destinations



- Ontologies are necessary to check consistency
- Today heuristically defined: a combination of data-driven and expert curation

Extracting Purpose

Polisis [Sec'2018]

Older NLP models: DNN, RNN; trained on OPP-115
Segmented text in paragraphs, used 9 categories
Granularity usually per phrase, e.g., for first/third-party collection/use
Service available online

MobiPurpose [Ubicomp'2019]:

Network Traffic of android apps:

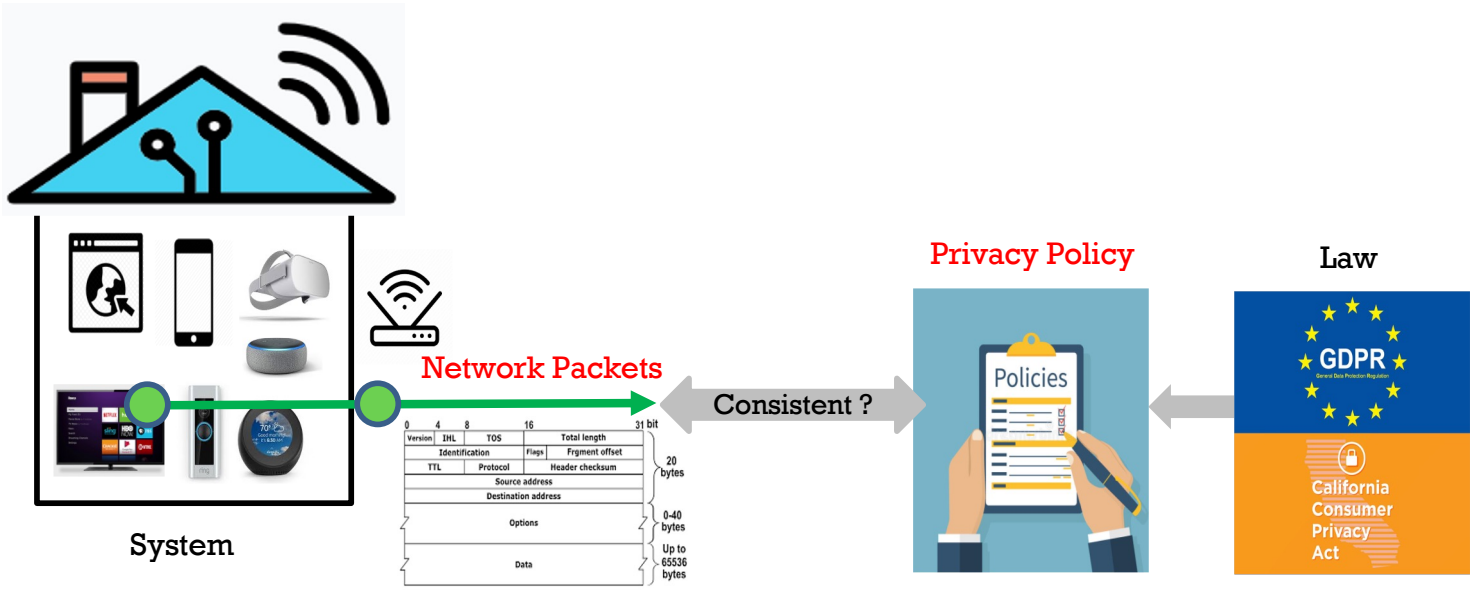
key-value pairs: manually labeled purpose based on keys; "adid"→advertising, "passwd"-->security etc.

also looked at name of apk and destination domain/entity

Purpliance [CCS'2021]:

Network side: built classifiers based on MobiPurpose
Policy side: New NLP models (BERT); looking for: "to", "in order to", "for…–ing"
Hierarchy in purposes → 5 main purposes
Applications: (1) contradictions, including purposes; tuple 5 vs. nested; (2) consistency of data flow-to-privacy policy, with purpose

Auditing Consistency of Network Traffic vs. Privacy Policies



CI tuple: (sender, recipient, data type; [subject]; (purpose; other))

"data flow" | transmission principle

Sender:
- Application (dev)
- 3rd party library
- Platform, device
- Malware

Recipient:
- 1st, 3rd parties, platform, cloud
- Advertisers & trackers (ATS)
- Organization

Data Type:
- Personally Identifiable Information (PII)
- Fingerprinting
- Activity Data
- Sensor data

[Subject:]
- Typically the user of the app and platform

Purpose:
- Functionality
- Analytics
- Tracking
- Ads
- Personalization
- Security

Other Aspects:
- With Notice?
- With Consent?
- Consistent Disclosure?

```
Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])

Policy: (sender, destination=entity, data type, purpose, [other])
```

Privacy Laws

Examples of disclosure requirements (right to know)

Category	CCPA Sections	GDPR Articles
Categories of personal information collected, used, or shared	1798.130(a)(5)(B-C)	14(1)(d)
Source (GDPR) / Categories of sources (CCPA) of the personal information	1798.110(c)(2)	14(2)(f)
Purposes for the collection, use, and sharing of personal information	1798.110(c)(3)	13(1)(c), 14(1)(c)
Categories of third parties with whom personal information is shared	1798.110(c)(4)	13(1)(e), 14(1)(e)

This affects how policies are written:

- "Collect" (1st party) vs. "share" (3rd party);
- "Use" vs. "collect" vs. purpose? How does it fit in information flow?
- Parameters of CI tuple can be "bloated"

Implication: "bloating" CI parameters

Example of good privacy policies following that format of sections

Information We Collect ...

Unity

- ***Name***
- ***Age or date of birth*** *...*

How We Use the Information We Collect or Receive

- *To **create, administer** and **troubleshoot** accounts, ...*
- *To **credit** or **accept payments**; ...*

Sharing Information ...

- ***Our affiliates** located all over the world ...*
- ***Third-party service providers**: ...*

We **collect** the following categories of *personal information*:

- ***Device information***… such as ***IP address***…
- ***Location***. ***We*** use ***this information*** to ***provide features***…

We **use** your *personal information*… to:

Kayak

- ***Provide the Services***...
- ***Authenticate your account***...

We **disclose** the *personal information*… as follows:

- With our ***travel partners***...
- With ***social networking services***...

Example of bad FB privacy policy [Shvartzshnaider et al. 2019]

[Advertisers, app developers and publishers]^{senders} can send **[us]**^{recipient} information **[through Facebook Business Tools that they use, including our social plug-ins (such as the Like button), Facebook Login, our APIs and SDKs or the Facebook pixel]**^{TP}. These partners provide information about **[your]**^{subject} **[activities off Facebook including information about your device, websites you visit, purchases you make, the ads you see and how you use their services]**^{attributes}**[whether or not you have a Facebook account or are logged in to Facebook]**^{TP}.

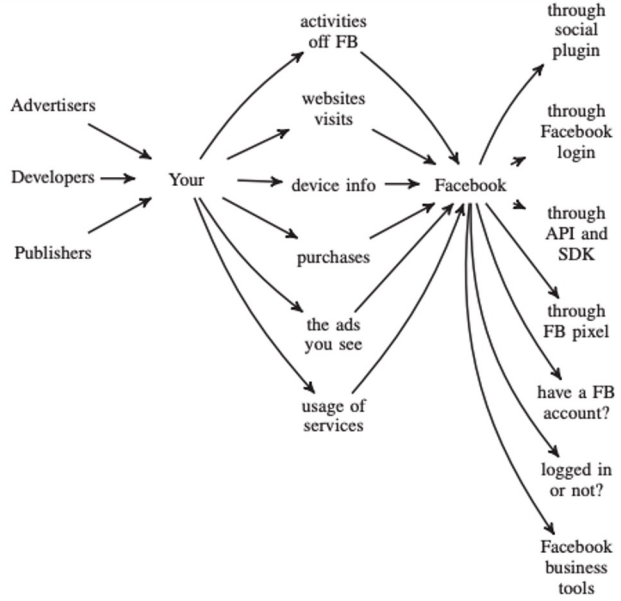


Figure 3: Example of CI parameter bloating in privacy policy text (*top*) and mapped into possible interpretations (*bottom*).

Open Problems and Directions

Q1: Auditing Network traffic vs Policies (NLP)

```
Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])

Policy: (sender, destination=entity, data type, purpose, [other])
```

- Dealing with parameters obtained from different sources
- Propose the full CI tuple to be used for this particular auditing.

Q2: Rethink the tuple data structure

- Hierarchy
- Purpose

Q3: Be proactive

- Beyond assessing/evaluating practices as appropriate or not
- Participate in defining laws, standards and open interfaces

Q2. Extend/Refine the Tuple data structure?

Hierarchy is necessary for consistency checking

Hierarchy is happening already

Hierarchy is more scalable

Hierarchies are hard to define: local (within a policy) vs global (by experts)

Law: currently encourages "bloating"; ongoing discussion in CPPA

Law: distinguishes collect/share/use

Re-consider purpose:

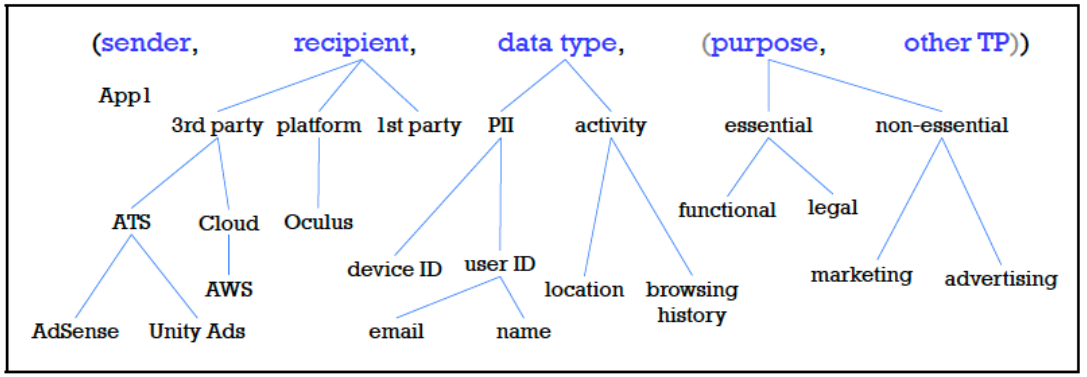
- part of another parameter (data type, purpose), (entity, purpose)?
- or part of TP?
- its own parameter?

(sender,	recipient,	data type,	(purpose,	other TP))
(App1,	1st party,	PII,	(functionality,	.....))
(App1,	1st party,	device ID,	(functionality,	.....))
(App1,	1st party,	user ID,	(functionality,	.....))
(App1,	ATS,	PII,	(advertising,	.....))
(App1,	ATS,	device ID,	(advertising,	.....))
(App1,	ATS,	user ID,	(advertising,	.....))
(App1,	AdSense,	PII,	(advertising,	.....))
(App1,	.....	.....	(.....,	.....))

(a) "Flat" CI tuples

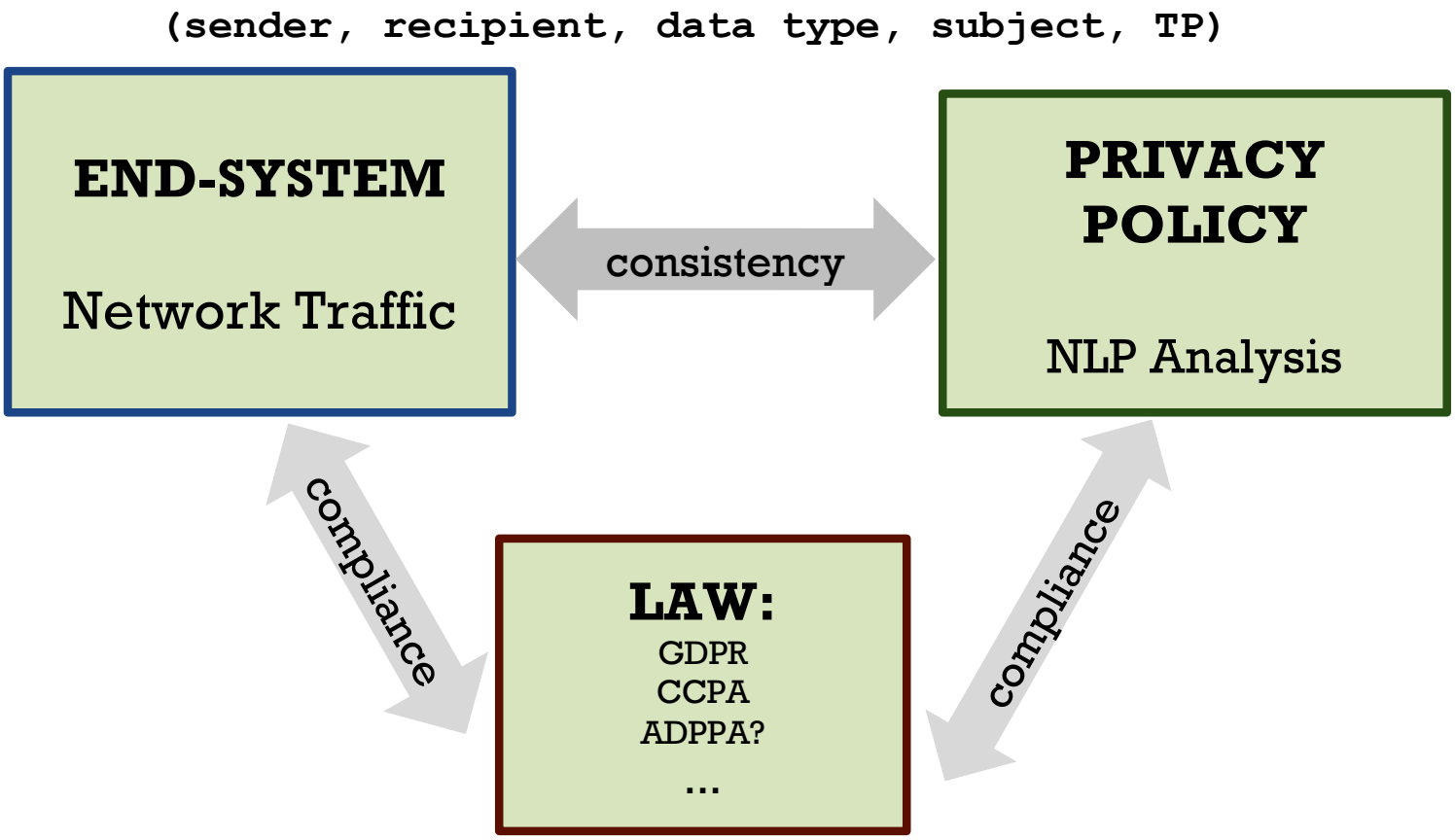
(sender,	recipient,	data type,	(purpose,	other TP))
App1	[1st party]	[PII]	functionality	.....
	[Platform]	device ID	marketing	
	Oculus	user ID	advertising	
	[ATS]	email	legal requirement	
	AdSense	name	.....	
	Unity Ads	[Activity]		
	[Cloud]	browsing history		
	AWS	location		

(b) "Bloated" CI tuples



(c) "Hierarchical" CI-tuples (*i.e.*, with ontologies).

CI tuple for Auditing from the Edge



Need for unified/auditable specification:

- opportunity for CI tuple to define the data structure for auditing and data rights requests

Thank you!

CI for Auditing (from the edge)

athina@uci.edu

http://properdata.eng.uci.edu
https://athinagroup.eng.uci.edu/projects/ovrseen/

UCI University of California, Irvine

ProperData
PROTECTING PERSONAL DATA FLOW ON THE INTERNET