Carnegie Mellon University

Grounded Analysis and the Discovery of Composable Security Models

Travis Breaux

July 1, 2014

Science of Security, PI Meeting

institute for SOFTWARE RESEARCH

Carnegie Mellon University

What is validity?

- Coherence theory – a knowledge claim is true, if it belongs to a coherent set of claims; e.g., smoking marijuana causes cancer
- Correspondence theory – a knowledge claim is true, if it corresponds to the world; e.g., it's sunny outside

Schmitt, F.F. (1995) *Truth: A primer.* Boulder, CO: Westview Press.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Coherence by example

Barth et al.'s 2006 formalization of contextual integrity…

- Begins with a philosophical abstraction, hypothesis or stated assumption, i.e., Contextual Integrity defined by Nissenbaum
- Establishes a coherent working example: Alice and Bob exchanging information about Charlie
- Establishes central concepts and rules of inference in Temporal Logic
- Claims fit prevailing assumptions about the world:

  "These norms are interpreted in a model of communicating agents who 'respect' the norms if the trace history of their communication satisfies a temporal formula constructed from the norms by taking the disjunction over positive norms and the conjunction over negative norms."

Barth, Datta, Mitchell, Nissenbaum. "Privacy and Contextual Integrity: Framework and Applications," *IEEE Security and Privacy*, pp. 184-198, 2006.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Correspondence by example

May et al.'s 2006 formalization of regulatory privacy rules…

- Begins with a research claim… that HIPAA consent rules can be expressed using access control matrix operations
- Establishes method with heuristics to translate English legal text into rules expressed in Promela
  - *Heuristic #1:* bi-directional tracing of legal cross-references to logic
  - *Heuristic #2:* distinguish system state and environmental state; latter is only known to human operators (e.g., testimonials)
- Results include select boundary cases and method limitations

May, Gunter, Lee. "Privacy APIs: Access Control Techniques to Analyze and Verify Legal Privacy Policies," IEEE 19th Computer Security Foundations Workshop, pp. 85-97, 2006.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Correspondence by example

Breaux et al.'s 2006 formalization of HIPAA…

- Begins with open coding frame to identify rules and constraints, multiple analysts compared coding result
- Establishes method with heuristics to translate English legal text into rules expressed in first order logic
  - *Heuristic #1:* definitions express transitive hierarchy of concepts
  - *Heuristic #2:* reconciling and prioritizing legal exceptions
  - *Heuristic #3:* conflicts due to rule subsumption
  - *Heuristic #4:* explicate implied rights from stated obligations
- Results include specific technical challenges to formalization

Breaux, Vail, Anton. "Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations," *14th IEEE International Requirements Engineering Conference*, pp. 49-59, 2006.

 5

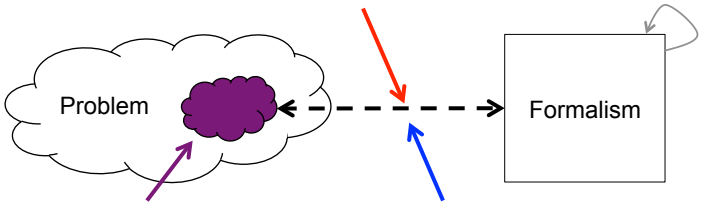
institute for SOFTWARE RESEARCH

Carnegie Mellon University

Threats to validity

Construct Validity: does the formal semantics accurately reflect the problem semantics?

Logical entailment and satisfiability



External Validity: to what extent is the the data representative of the problem at large?

Internal Validity: are the inferences drawn from the dataset consistent and complete?

Reliability: can multiple people apply the method to yield the same results?

Inspired by: Shadish, Cook and Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Miifflin, 2002.

 6

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Grounded Theory

Defining components of grounded analysis…[1]

- Grounded: Constructing analytical codes and categories from data, not from pre-conceived logically deduced hypotheses
- Iterative: Constant-comparisons to challenge emerging theory
- Reflective: Memo-writing to elaborate categories, specify their properties, define relationships between categories and identify gaps
- Logical: Sampling for theory construction, not for population representativeness
- Disembodied: Conducting the literature review after the independent analysis

[1]Kathy Charmaz, *Constructing Grounded Theory*, SAGE Publications, 2006.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Sampling and replication

- Sampling theory: select units by chance with known probability from a larger population to yield a match between the sample and population distributions
  - Predicts the mean and variance of the sample will match the population (e.g., central limit theorem for normally distributed data)

What if we can't enumerate the population?

- Purposive sampling: classify available population and randomly sample within classes
  - Advantage: forces representation of diverse instances
  - Disadvantage: may overstate the role of outliers
- Replication logic: use comparable datasets

Shadish, Cook and Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Miifflin, 2002.

Yin. *Case Study Research Design and Methods*, SAGE Publications, 2013.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

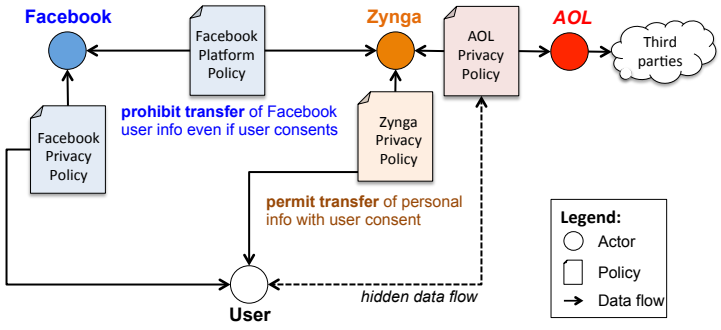
Key concepts in coding method

- Units of analysis: define structure of data elements to code (e.g., sentences, verb or noun phrases, etc.)
- Multi-cycle coding: begin with initial coding frame, iterate to discover new codes; axial coding to find patterns
  - *Mutual Exclusivity*: codes are disjoint; analysts refine definitions
  - *Completeness*: every unit is coded; new codes emerge
  - *Consistency*: new codes tested on previously coded units
- Saturation: Over successive data elements, existing codes consistently explain the dataset (no new codes)
- Reliability: Cohen's or Fleiss' Kappa, or Krippendorf's Alpha used to evaluate inter-rater reliability across multiple coders

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Privacy and data supply chain



Privacy policies contain privacy requirements for data that flow within a data supply chain; conflicts can exist among these requirements; repurposing can be an issue

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Approach and research method

- Exploratory case study design [Yin08]
  - Data: Facebook Platform Policy (for developers)
  - Developed specification language from results
- Extended evaluation
  - Data: Zynga privacy policy, AOL privacy policy
- Applied content analysis [Sal13] to extract phrases to formalize data requirements in logic

R. Yin, *Case Study Research: Design and Methods*, 4th ed. SAGE, 2008.
J. Saldaña, *The Coding Manual for Qualitative Researchers, 2nd ed.* SAGE, 2013

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Mapping policy statements to types

- ***Policy Statements*** *describe events or states outside the app*

  "You must not violate any law or the rights of any individual or entity."

- **Non-data Requirements** describe non-data functionalities

  "You will include your privacy policy URL in the App Dashboard."

- **Data Requirements** describe actions on data

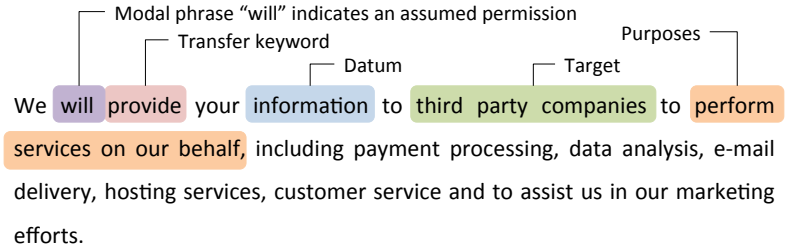
  "You must not include functionality that proxies, requests or collects Facebook usernames or passwords."

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Identifying actions on data

Step 3: Annotate policy text to identify action and role values

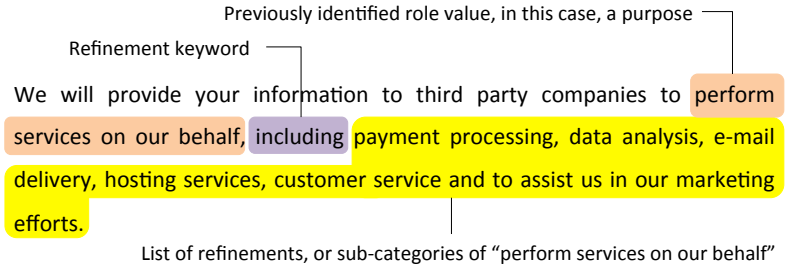


institute for SOFTWARE RESEARCH

Carnegie Mellon University

Identifying definitions, elaborations

Step 4: Annotate policy text to identify other subsumption relations

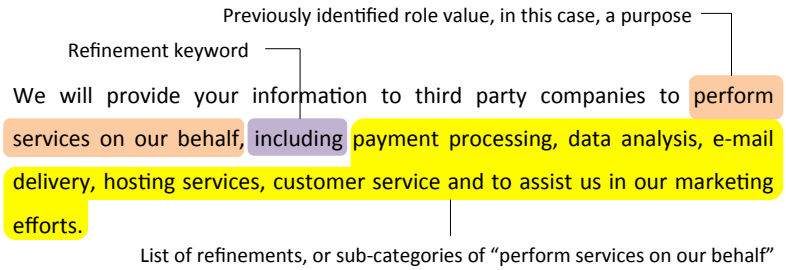


institute for SOFTWARE RESEARCH

Carnegie Mellon University

Identifying definitions, elaborations

Step 4: Annotate policy text to identify other subsumption relations



Step 5: Write expression in specification language

SPEC HEADER
P performing-services > payment-processing, e-mail-delivery, hosting-services, customer-service, marketing
SPEC POLICY
P TRANSFER information TO third-party-companies FOR performing-services

Carnegie Mellon University

Formal semantics to compile logic

Step 5: Write expression in specification language

SPEC HEADER
P performing-services > payment-processing, e-mail-delivery, hosting-services, customer-service, marketing
SPEC POLICY
P TRANSFER information TO third-party-companies FOR performing-services

Step 6: Compile language into Description Logic (OWL)

payment-processing ⊑ performing-services
e-mail-delivery ⊑ performing-services
…
Z-92 ≡ TRANSFER ⊓ ∃hasObject.information ⊓
∃hasTarget.third-party-companies ⊓ ∃hasPurpose.performing-services
Z-92 ⊑ Permission

isr institute for SOFTWARE RESEARCH

Carnegie Mellon University

Using context in annotation

- [Zynga] *"may access and store some or all of the following information, as allowed by you, the SNS and your preferences"*

  Action is COLLECT

- [AOL] *"Personal information such as name, address and phone number is never accessed for this purpose"*

  Action is USE

- [AOL] *"In that the case, the acquiring (or merging) company will have access to your information"*

  Action is TRANSFER

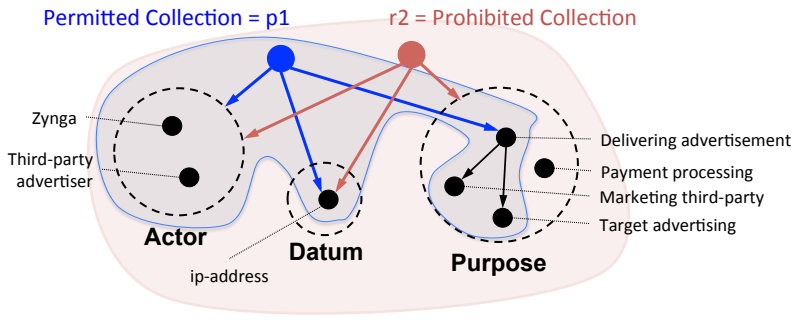
institute for SOFTWARE RESEARCH

Carnegie Mellon University

How we identify conflicts – 1

p1: Permitted to collect IP address from anyone for advertising

r2: Prohibited from collecting IP address from anyone for anything



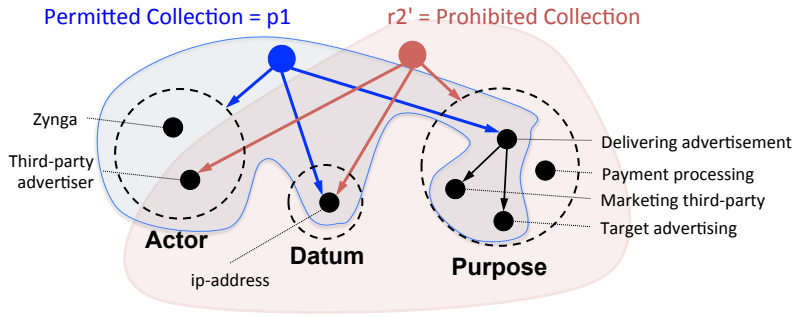
institute for SOFTWARE RESEARCH

Carnegie Mellon University

How we identify conflicts – 2

p1: Permitted to collect IP address from anyone for advertising

r2': Prohibited from collecting IP address from third-party advertisers for anything



institute for SOFTWARE RESEARCH

Carnegie Mellon University

How we trace data

- Characterizing data flows using subsumption
  - *Underflow*, occurs when the data target subsumes the source
  - *Overflow*, occurs when the data source subsumes the target
  - *Exact flow*, occurs when the data source and target are equivalent
  - Identify repurposing, visualize dependencies etc.

AOL-16: Collect name, contact information, payment method from site visitor for business purposes

AOL-48: Transfer personally identifiable information to key partners

$$contact\_info \sqsubseteq personally\_identifiable\_info$$
$$business\_purposes \sqsubseteq anything$$

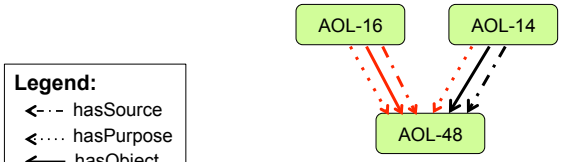
institute for SOFTWARE RESEARCH

Carnegie Mellon University

How we trace data

AOL-16: Collect name, contact information, payment method from site visitor for business purposes

AOL-14: Collect personally identifiable information (from anyone) for contacting you to discuss our products and services



Legend:
- hasSource
- hasPurpose
- hasObject

Colors:
- **red** underflow
- **black** exact flow

AOL-48: Transfer personally identifiable information (from anyone) to key partners (for any purpose)

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Characterizing data flows in DL

- *Underflow*, occurs when the data source *Fs* is subsumed by the target *Ft*, if and only if,
  $T \models Fs,j \sqsubseteq Ft,k$
- *Overflow*, occurs when the data target is subsumed by the source, if and only if,
  $T \models Ft,j \sqsubseteq Fs,k$
- *Exact flow*, occurs when the data source and target are equivalent, if and only if,
  $T \models Fs,j \equiv Ft,k$
- *No flow,* otherwise.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

RESULTS OF CASE STUDY

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Results of extended evaluation

Policy	S	D	Modality			Action		
			P	O	R	C	U	T
Facebook	105	39	15	4	25	6	15	14
Zynga	195	64	58	1	8	22	8	15
AOL	74	41	43	0	4	12	15	10

Extracted: (S)tatements, (D)ata requirements
Modalities: (P)ermission, (O)bligation, (R) prohibition
Actions: (C)ollection, (U)se), (T)ransfer

Breaux, Hibshi, Rao. "Eddy, A Formal Language for Specifying and Analyzing Data Flow Specifications for Conflicting Privacy Requirements," To Appear: *Requirements Engineering Journal,* 2014.

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Results of extended evaluation

Policy	S	D	Modality			Action		
			P	O	R	C	U	T
Facebook	105	39	15	4	25	6	15	14
Zynga	195	64	58	1	8	22	8	15
AOL	74	41	43	0	4	12	15	10

Extracted: (S)tatements, (D)ata requirements
Modalities: (P)ermission, (O)bligation, (R) prohibition
Actions: (C)ollection, (U)se), (T)ransfer

Breaux, Hibshi, Rao. "Eddy, A Formal Language for Specifying and Analyzing Data Flow Specifications for Conflicting Privacy Requirements," To Appear: *Requirements Engineering Journal,* 2014.

isr institute for SOFTWARE RESEARCH

Carnegie Mellon University

Results of extended evaluation

Policy	S	D	Modality			Action		
			P	O	R	C	U	T
Facebook	105	39	15	4	25	6	15	14
Zynga	195	64	58	1	8	22	8	15
AOL	74	41	43	0	4	12	15	10

Extracted: (S)tatements, (D)ata requirements
Modalities: (P)ermission, (O)bligation, (R) prohibition
Actions: (C)ollection, (U)se), (T)ransfer

Breaux, Hibshi, Rao. "Eddy, A Formal Language for Specifying and Analyzing Data Flow Specifications for Conflicting Privacy Requirements," To Appear: *Requirements Engineering Journal,* 2014.

isr institute for SOFTWARE RESEARCH

Carnegie Mellon University

Indicative keywords in coding

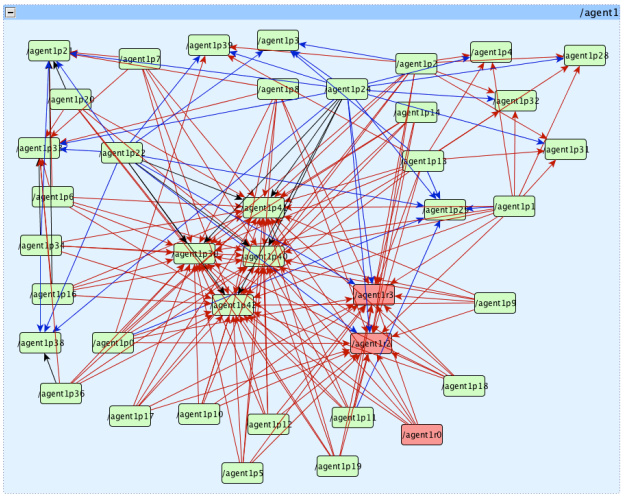
Action keywords indicate when a statement was coded as a collection, use or transfer requirement

DL Action	Action keywords
COLLECT	Access, assign, collect, collected, collection, collects, give you, import, keep, observes, provide, receive, record, request, share, use
USE	Access, accessed, communicate, delivering, include, matches, send, use, used, uses, using, utilized
TRANSFER	Access, disclose, disclosed, disclosure, give, in partnership with, include, make public, on behalf of, provide, see, share, shared, transfer, use, used with, utilized by

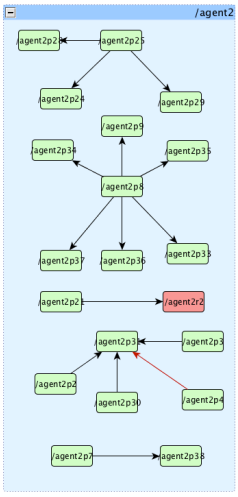
institute for SOFTWARE RESEARCH

Carnegie Mellon University

Zynga*



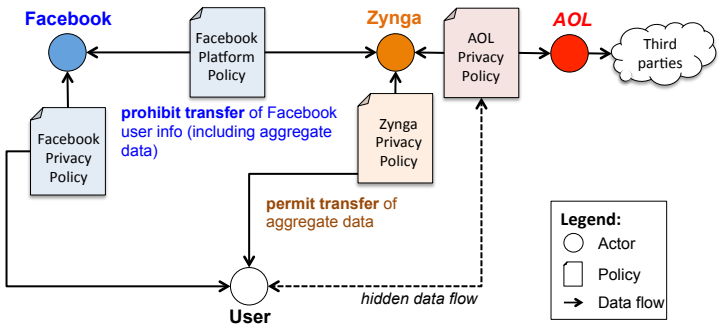
AOL*



Legend: Permission Prohibition → Exact flow → Underflow → Overflow * Arrows point from collections to transfers

Carnegie Mellon University

Identifying conflicting requirements



In a multi-tier application, conflicts can exist between privacy requirements in policies governing data flow in a data supply chain

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Conflicts identified in our study

- Conflicts between Facebook and Zynga (3 conflicts)
  - sharing of aggregate or anonymous data
  - transfer of unique user IDs to third party advertisers
  - sharing data for the purposes of merger and acquisition by a third-party
- Conflict within AOL Advertising (1 conflict)
  - collection and use of personally identifiable information

institute for SOFTWARE RESEARCH

Carnegie Mellon University

Completeness with respect to dataset

Policy	S	D	Formalized
Facebook	105	39	.371
Zynga	195	64	.328
AOL	74	41	.554

- **Functional requirements**: "You may cache data you receive through use of the Facebook API"
- **Missing semantics**: "Information collected on AOL Advertising Sites may be combined with information collected from other sources."
- **Testimonials**: "You must ensure that you own or have secured all rights necessary to copy, display, distribute… all content of or within your application"

institute for SOFTWARE RESEARCH

Carnegie Mellon University

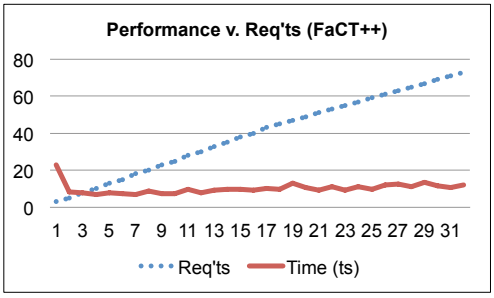
Reliability of extraction

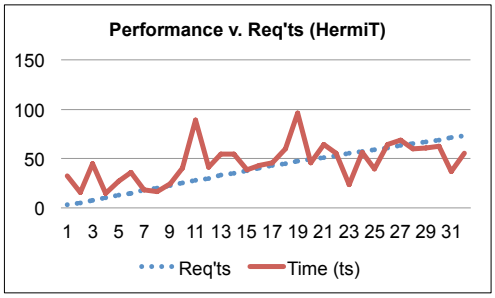
Policy	Missing Codes	Disagree-ments	Cohen's Kappa	Krippendorf's Alpha	
				Replication (2 Raters)	All 3 Raters
Facebook	5	3/54	.884	.941	.937
Zynga	13	4/76	.919	.922	.906
AOL	5	5/35	.800	.828	.849

- *Missing Codes* refers to the number of units that were coded by only one coder, which were not used to compute Cohen's Kappa but were factored into Krippendorf's Alpha
- *Disagreements* reports the number of units where the coders disagreed out of the total number of mutually-coded units

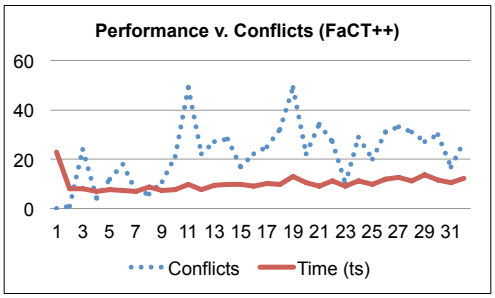
institute for SOFTWARE RESEARCH

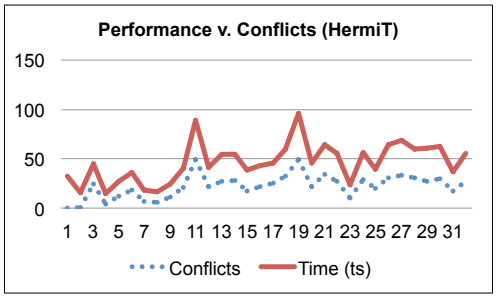
Carnegie Mellon University





Carnegie Mellon University





Carnegie Mellon University

Questions?

Research funded by:

- ONR Award #N00244-12-1-0014
- National Security Agency
- NSF Award #1330596

isr institute for SOFTWARE RESEARCH