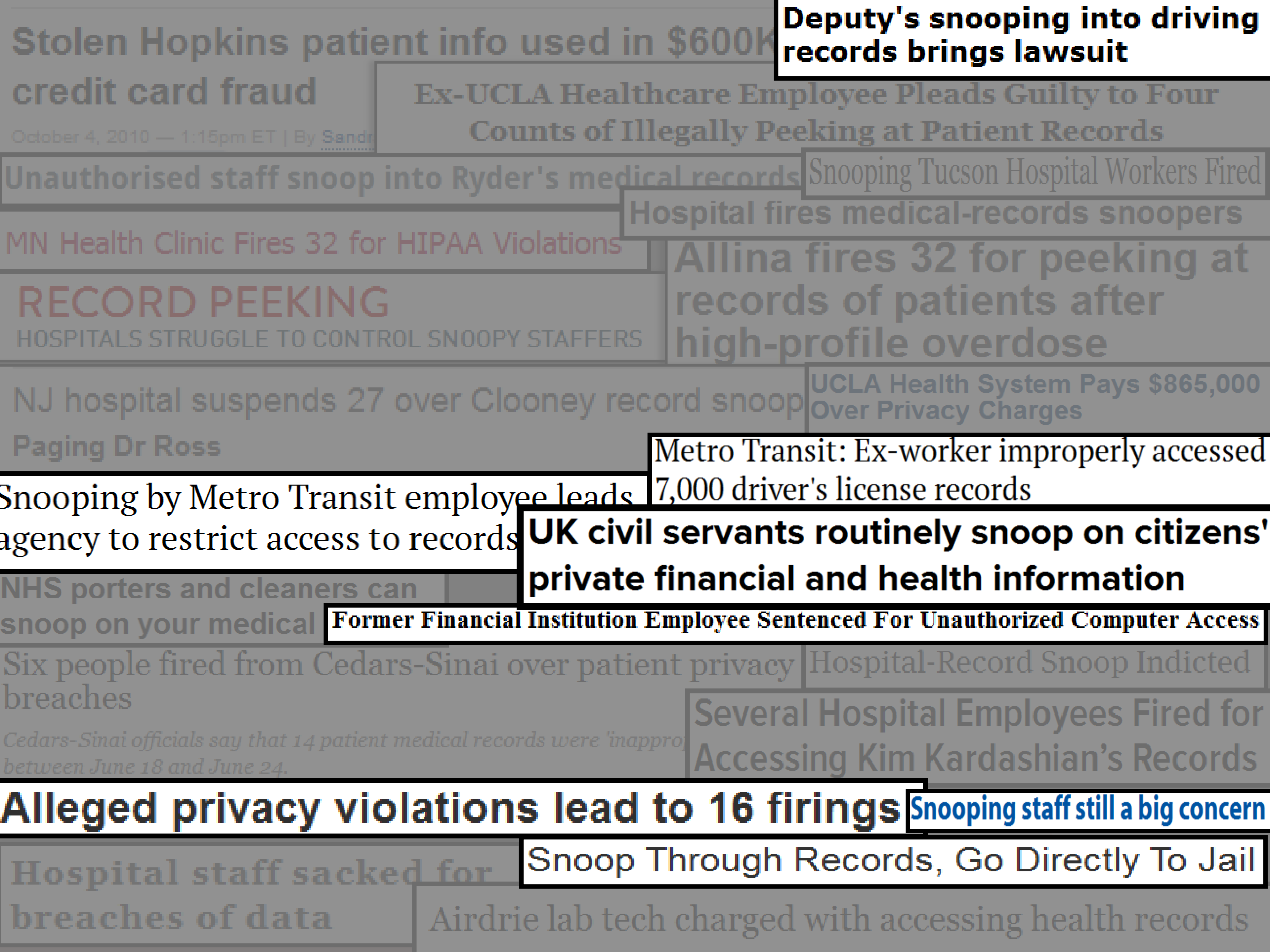
Enabling Forensics by Proposing Heuristics to Identify Mandatory Log Events

Jason King

Rahul Pandita

Laurie Williams

Software Engineering @ NCSU



Stolen Hopkins patient info used in $600K credit card fraud

October 4, 2010 — 1:15pm ET | By Sandr

Deputy's snooping into driving records brings lawsuit

Ex-UCLA Healthcare Employee Pleads Guilty to Four Counts of Illegally Peeking at Patient Records

Unauthorised staff snoop into Ryder's medical records

Snooping Tucson Hospital Workers Fired

Hospital fires medical-records snoopers

MN Health Clinic Fires 32 for HIPAA Violations

RECORD PEEKING

HOSPITALS STRUGGLE TO CONTROL SNOOPY STAFFERS

Allina fires 32 for peeking at records of patients after high-profile overdose

NJ hospital suspends 27 over Clooney record snoop

Paging Dr Ross

UCLA Health System Pays $865,000 Over Privacy Charges

Metro Transit: Ex-worker improperly accessed 7,000 driver's license records

Snooping by Metro Transit employee leads agency to restrict access to records

UK civil servants routinely snoop on citizens' private financial and health information

NHS porters and cleaners can snoop on your medical

Former Financial Institution Employee Sentenced For Unauthorized Computer Access

Six people fired from Cedars-Sinai over patient privacy breaches

Cedars-Sinai officials say that 14 patient medical records were 'inappro between June 18 and June 24.

Hospital-Record Snoop Indicted

Several Hospital Employees Fired for Accessing Kim Kardashian's Records

Alleged privacy violations lead to 16 firings

Snooping staff still a big concern

Snoop Through Records, Go Directly To Jail

Hospital staff sacked for breaches of data

Airdrie lab tech charged with accessing health records

Motivation

- Repudiation Threats
- Who, What, When, Where, Why, & How
- 2011 Veriphyr Survey of Patient Privacy Breaches
  - 52%: inadequate tools for monitoring PHI access
- No single resource for log specifications
- Requirements specifications describe user activity

Objective

*…help **software engineers** strengthen **forensic-ability** and user **accountability** by*

*1) systematically **identifying mandatory log events** through processing of unconstrained natural language software artifacts; and*

*2) proposing **empirically-derived heuristics** to help determine whether an event must be logged*

The Science of Security initiative is funded by the National Security Agency.

Research Questions

- **RQ1**: How **often** do descriptions of mandatory log events appear in natural language software artifacts?
- **RQ2**: What similarities and differences exist in the **grammar** and **vocabulary** used in different software artifacts?
- **RQ3**: What factors help **decide** whether an action must be logged?

Software & Artifacts

- **iHRIS: Open Source Human Resources Information Solutions v4.2.** *more than 675,000 health worker records currently supported*
  - Artifact: traditional software requirements specification for Page Builder module

- **iTrust: Open Source Electronic Health Record System v18.** educational, test-bed healthcare software developed & maintained at NCSU
  - Artifact: use-case based requirements specification

- **OCS (Open Conference Systems): Open Source Scholarly Conference Management System v2.3.6.** from Public Knowledge Project (PKP)
  - Artifact: "OCS in an Hour" User Manual

D. Berry, K. Daudjee, J. Dong, I. Fainchtein, M. A. Nelson, T. Nelson, and L. Ou, "User's Manual as a Requirements Specification: Case Studies," Requirements Engineering, vol. 9, pp. 67-82, 2004.

The Science of Security initiative is funded by the National Security Agency.

Identifying Verb-Object Pairs

- For each statement in the software artifact
- Identify each event in terms of the action (verb) and resource (object) acted upon
    - <verb, object>
- Examples:

Classification

- Individually classify verb-object pairs as *mandatory log event* or *not a mandatory log event*
  - – *A mandatory log event* (MLE) *is an action that must be logged in order to hold the software user accountable for performing the action.*
- Reconcile differences, document discussion
- Third individual serves as “tiebreaker”

Inter-rater Agreement

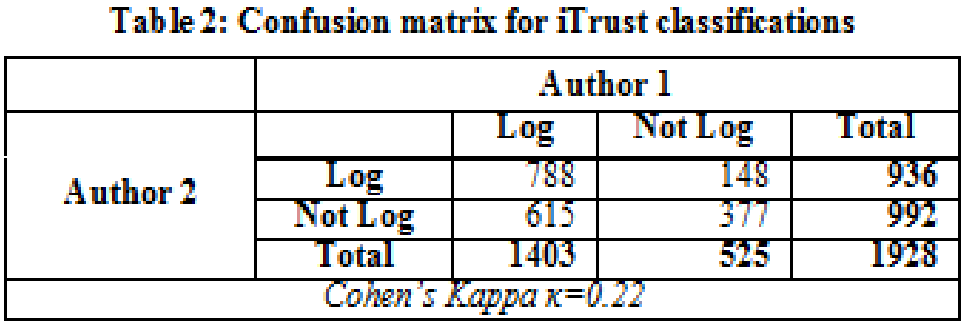


Table 2: Confusion matrix for iTrust classifications

	Author 1			
Author 2		Log	Not Log	Total
	Log	788	148	936
	Not Log	615	377	992
	Total	1403	525	1928
Cohen's Kappa κ=0.22				

Cohen’s Kappa = 0.22

Results: Verb-Object Pairs

Software	Total Statements	Total Verb-Object Pairs	Total MLE Verb-Object Pairs	Statements with at least One MLE Verb-Object Pair
iHRIS	36	106	96 (91%)	27 (75%)
iTrust	1301	1928	1217 (63%)	802 (62%)
OCS	791	1479	747 (51%)	434 (55%)
TOTAL	**2128**	**3513**	**2060 (59%)**	**1263 (59%)**

- ***RQ1**: How **often** do descriptions of mandatory log events appear in natural language software artifacts?*
    - Over half (59%) of the individual statements studied contain descriptions of MLEs

Results: Differences between Artifacts

- iHRIS: traditional requirements specification
  - “The system shall…”
- iTrust: use-case based requirements specification
  - “A patient chooses to…”
- OCS: user manual
  - “You can also elect to require authors to agree to…”

The Science of Security initiative is funded by the National Security Agency.

Results: Differences between Artifacts

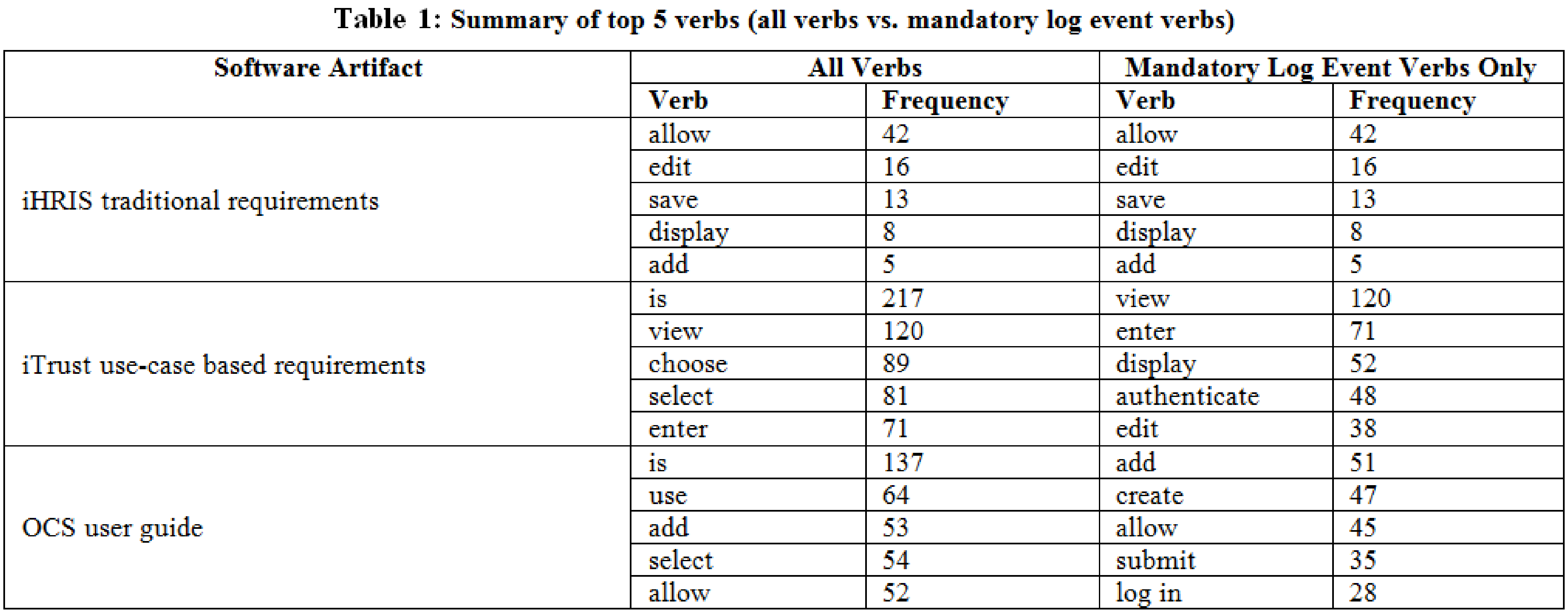


Table 1: Summary of top 5 verbs (all verbs vs. mandatory log event verbs)

Software Artifact	All Verbs		Mandatory Log Event Verbs Only	
	Verb	**Frequency**	**Verb**	**Frequency**
iHRIS traditional requirements	allow	42	allow	42
	edit	16	edit	16
	save	13	save	13
	display	8	display	8
	add	5	add	5
iTrust use-case based requirements	is	217	view	120
	view	120	enter	71
	choose	89	display	52
	select	81	authenticate	48
	enter	71	edit	38
OCS user guide	is	137	add	51
	use	64	create	47
	add	53	allow	45
	select	54	submit	35
	allow	52	log in	28

Results: Differences between Artifacts

- **RQ2**: What similarities and differences exist in the **grammar** and **vocabulary** used in different software artifacts?
  - iHRIS traditional requirements specification has more constrained & consistent grammar/vocabulary
  - *States* of a system are commonly described in iTrust and OCS
    - "A patient is a registered user"
    - "OCS is designed to be a multilingual system"

Results: Heuristics

- **RQ3:** *What factors help decide whether an action is a mandatory log event?*

Heuristics

- **Heuristic H1**: If the verb involves creating, reading, updating, or deleting resource data in the software system, then the event must be logged.
  - 134 <verb, object> pairs explicitly used *create, read, update, delete* terminology

- **Heuristic H2**: If the verb can be accurately rephrased in terms of creating, reading, updating, or deleting resource data in the software system, then the event must be logged.
  - Example: "A patient *designates* a representative"
    "A patient *creates* a representative for…"
  - 1,243 <verb,object> pairs can be rephrased as CRUD actions

Heuristics

- **Heuristic H3**: If the verb implies the system displaying or printing resource data that is capable of being viewed in the user interface or on paper, then the event must be logged.
  - Examples: view, see, print, display, appear
  - 397 <verb, object> pairs relate to viewing data
- **Heuristic H4**: If the verb expresses the intent to perform an action, such as "choose to", "select to", "plan to", or "wish to", then the intent event is not a mandatory log event.
  - 351 <verb, object> pairs express intent

Heuristics

- **Heuristic H5**: If the verb expresses the granting or revocation of access privileges in the software system, then the event must be logged.
  - Example: “Doctors are allowed to…”
  - 126 <verb, object> pairs relate to access privileges

- **Heuristic H6**: If the verb is ambiguous, such as “provide” or “order”, context must be considered when determining if the event must be logged.
  - Example: “A doctor *orders* lab procedures…”
“Lab procedures are *ordered* alphabetically in a list…”
  - 158 <verb, object> pairs contain ambiguous verbs

Software Engineering @ NCSU

Heuristics

- **Heuristic H7**: If the verb describes an action that takes place outside the scope of the functionality of the software, then the event is not a mandatory log event.
  - Example: “To enable electronic payment, register for a PayPal business account”
  - 314 <verb, object> pairs were outside the scope

- **Heuristic H8**: If the verb involves the creation or termination of a user session, then the event must be logged.
  - Example: “A patient authenticates…”
  - 100 <verb, object> pairs relate to user sessions

The Science of Security initiative is funded by the National Security Agency.

Heuristics

- **Heuristic H9**: If the verb describes a state or quality within the system, then the event is not a mandatory log event.
  - Example: “A patient is a registered user”
    “A list of pages is available”
  - 253 <verb, object> pairs describe system states or qualities

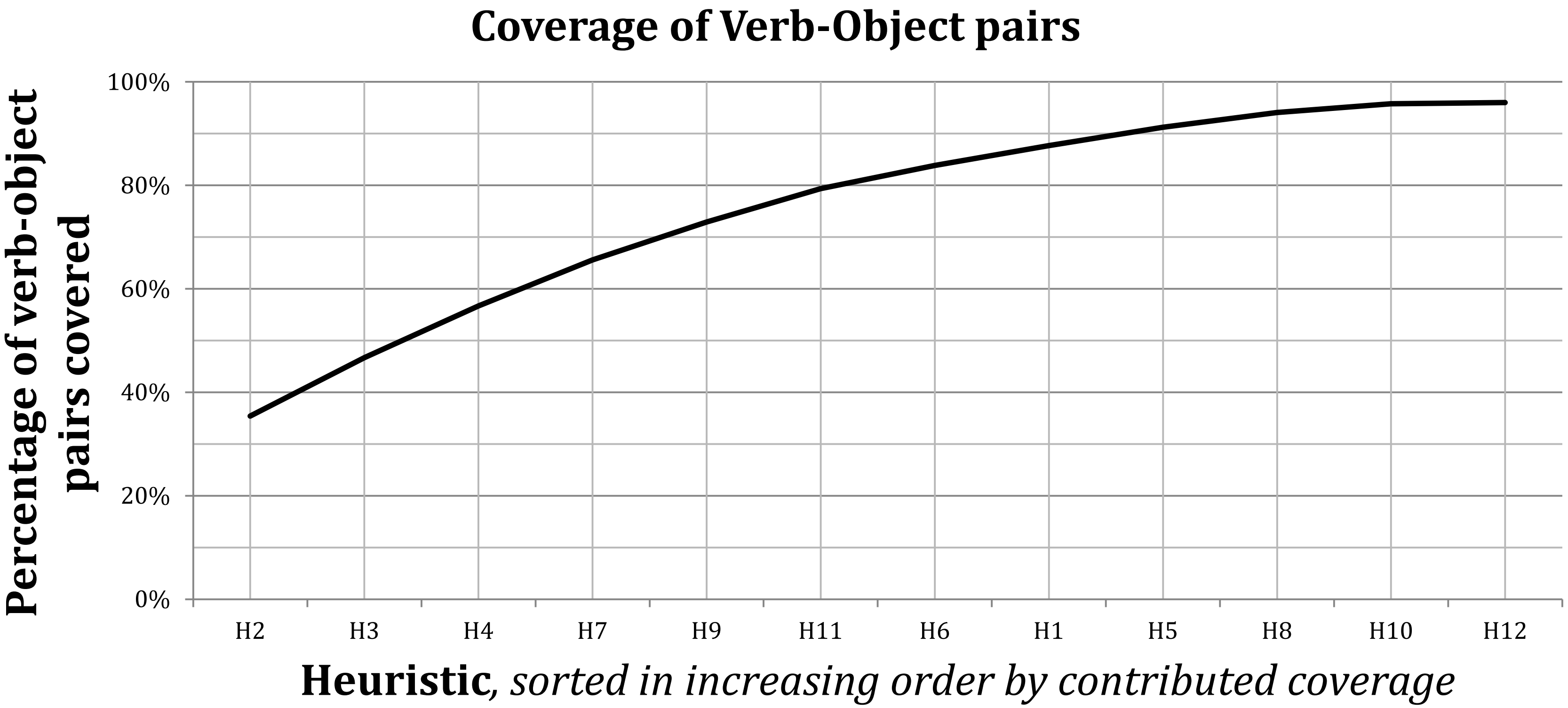
- **Heuristic H10**: If the verb describes possession or composition of a resource or quality, then the event is not a mandatory log event.
  - Example: “The patient has a known interaction”
    “The row contains the doctor’s comments”
  - 57 <verb, object> pairs describe possession or composition

Heuristics

- **Heuristic H11**: If the verb describes navigation or mechanical interaction with the software interface, then the event is not a mandatory log event.
  - Example: “The doctor remains on the allergy page”
    “The author needs to click on ‘Active Submissions’”
  - 226 <verb, object> pairs relate to navigation or interface actions

- **Heuristic H12**: If the verb describes initialization of the software or configuration of the software, then the event must be logged.
  - Example: “The administrator can install additional locales”
  - 8 <verb, object> pairs relate to initialization/configuration

Heuristics: Summary



- Our 12 heuristics were helpful for ~96% of verb-object pairs

Lessons Learned

- **Use Consistent Terminology**
  - *View, display, present, show, see, provide, read…*
- **Use Consistent Perspective**
  - The *user* reads, views…
  - The *system* shows, displays, presents…
  - The *data* appears…
- **Use CRUD Terminology**
  - *Manage? Make? Indicate? Merge?*

Future Work

- Include additional types of domains, artifacts
- Tool-assisted approach using the heuristics
- User study:
  - Controlled experiment
  - Relevance, quantity, efficiency of identifying MLEs using heuristics
- Metric for *forensic-ability*

Summary

- A set of **empirically-derived heuristics** to assist software engineers in determining whether a given user action described in a software artifact must be logged.
- A **set of considerations for requirements engineers** to help clearly and unambiguously document mandatory log events in software artifacts.
- An **oracle of mandatory log event classifications** for three open-source software systems. The oracle is publically available on the project website.
  - http://go.ncsu.edu/NLPLogging