What Wikipedia Deletes: Characterizing Dangerous Collaborative Content

Andrew G. West and Insup Lee

WikiSym `11 – October 4, 2011

Outline

CONCEPT: Examine “dangerous” (legally-threatening) content on Wikipedia.

- Identifying dangerous content
- Motivations for research
- Corpus construction
- Corpus analysis
- Impacts and takeaway

Deletion Example

Revision history of "Test Page"

- (cur) (prev) 02:06, 14 January 2011 WikiUser (Talk | contribs)
  (38 bytes) *(Add details)* (undo)
- (cur) (prev) 02:01, 14 January 2011 Andrew (Talk | contribs)
  (26 bytes) *Revert vandalism*
- (cur) ~~*00:00, 14 January 2011*~~ SuperVandal (Talk | contribs)
  ~~*(comment removed)*~~ [deleted]
- (cur) (prev) 23:59, 13 January 2011 76.99.208.144 (Talk)
  (26 bytes) *(Minor grammatical fix)* (undo)
- (cur) (prev) 23:59, 13 January 2011 Andrew (Talk | contribs)
  (24 bytes) *(Creating initial content)*

Reasons for Deletion

- Simple vandalism doesn't qualify [3]
- This is "edit deletion"; not "page deletion"

ID	DESCRIPTION	ISSUE
RD1	Blatant copyright violations	Copyright
RD2	Grossly insulting/offensive	Libel/slander
RD3	Purely disruptive material	Security attack
RD4	Edits pending "super delete"	Privacy issues
RD5	Other valid deletion	???
RD6	Non-contentious housekeeping	???

Research Motivations

- Initially: Might deletion functionality be used to "hide" security vulnerabilities (i.e., successful attacks) from public view [9]?

- Other research questions:
  - How prevalent is dangerous content?
  - What types of edits get deleted?
  - How prompt is the administrative response?
  - What is impact in terms of end-user exposure

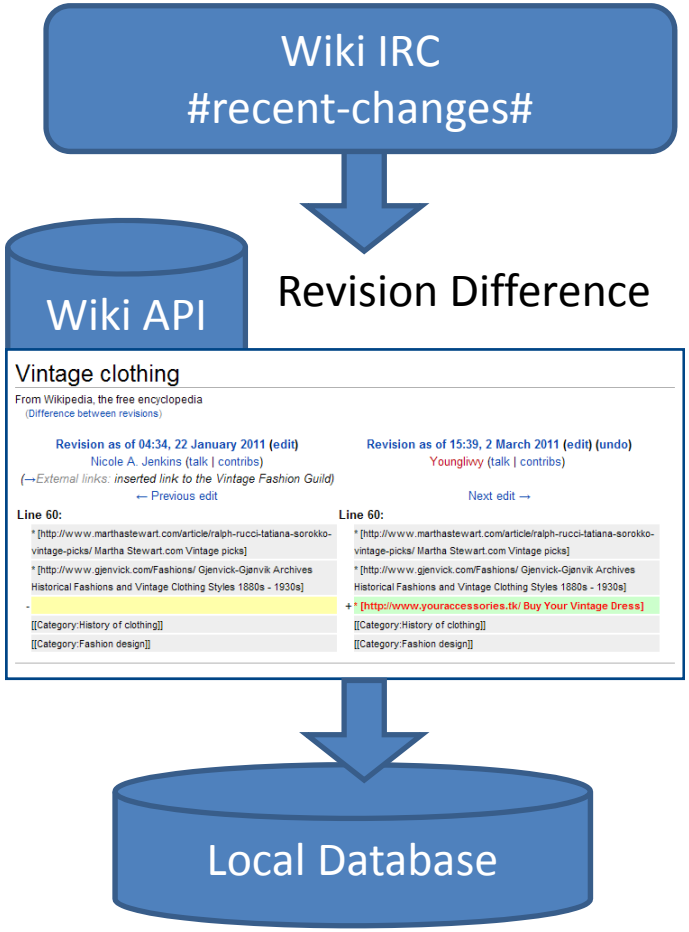
CREATING A CORPUS OF DANGEROUS CONTENT

The RevDelete Tool

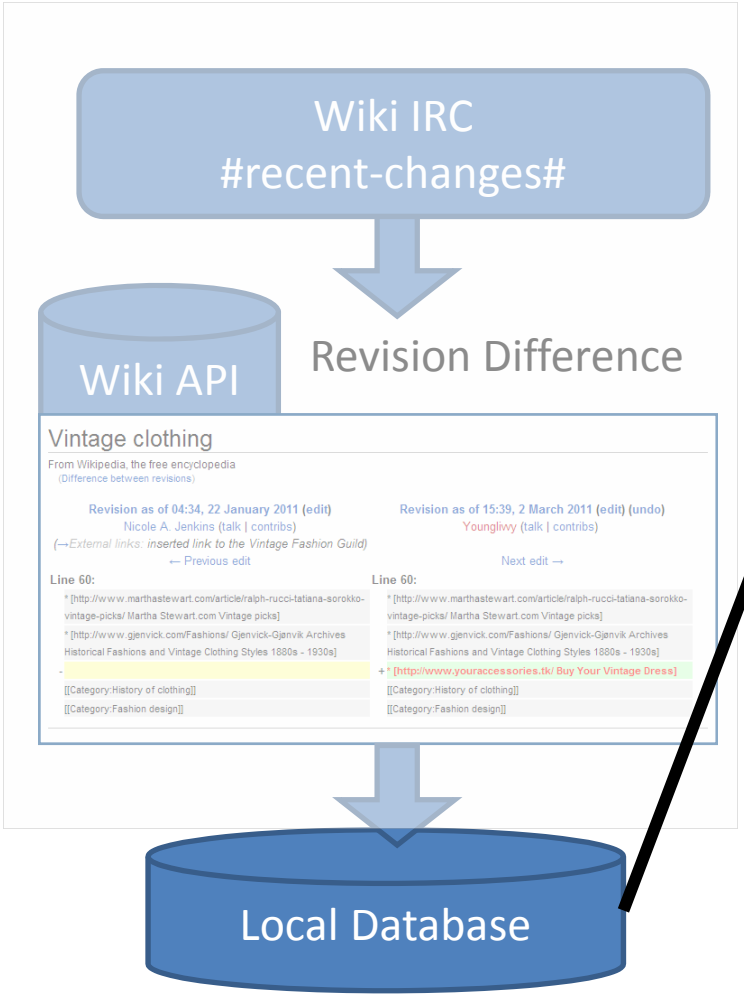
	REDACTION	SUPRESSION
ENABLED TYPES	admin + oversighter	oversighter
ENABLED USERS	≈1800	≈40
LOGGING	Public	Private

- Replaced inelegant DB deletion
- "Redaction" widely enabled since 5/2010
  - History/backlog of incidents being addressed

Data Collection



Data Collection

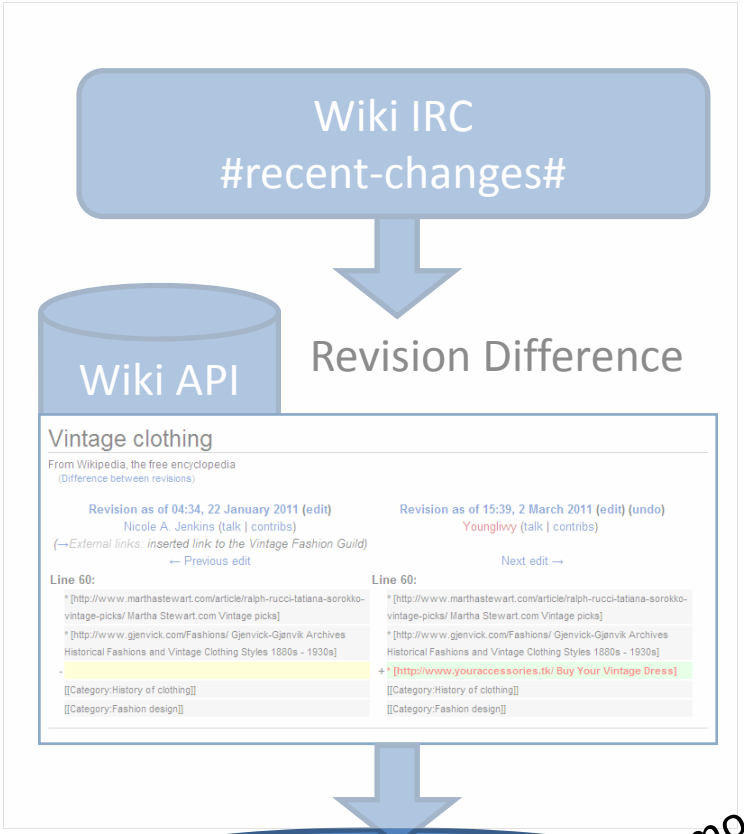


REDACTION

Public Redaction Log

ID	REASON	METADATA
X	[RD1]: Copied from…	11:05AM by AdminA
Y	[RD3]: Spam attack…	10:57AM by AdminB

Data Collection

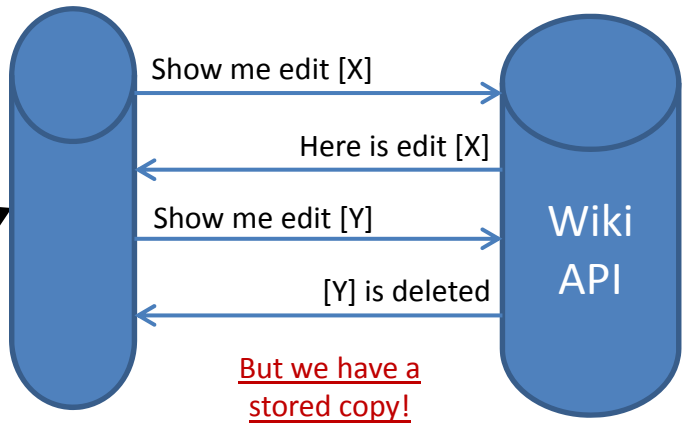


REDACTION

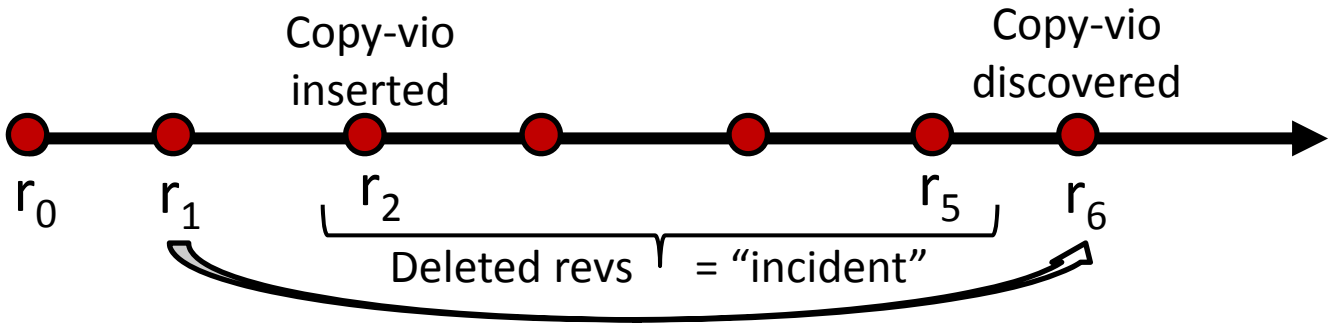
Public Redaction Log

ID	REASON	METADATA
X	[RD1]: Copied from…	11:05AM by AdminA
Y	[RD3]: Spam attack…	10:57AM by AdminB

SUPRESSION



Incident Groupings



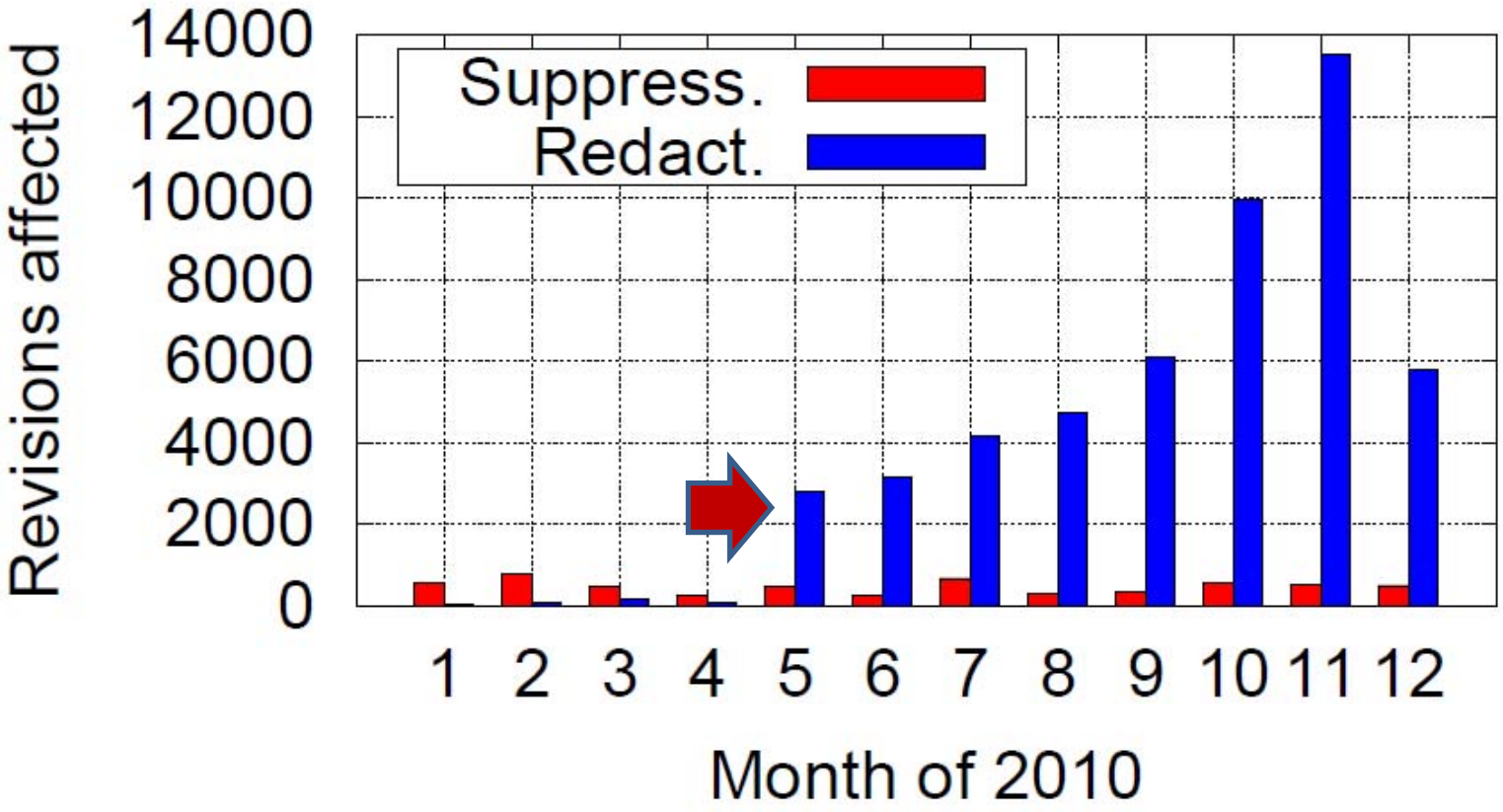
- 89% of incidents have just one revision
  - But copyright issues (RD1) have 12.5 revs @ median
- Collateral damage a real threat
  - Sept. 2010 incident involving 25k RIDs [2]
- Prefer incident-level statistics

ANALYZING REDACTIONS

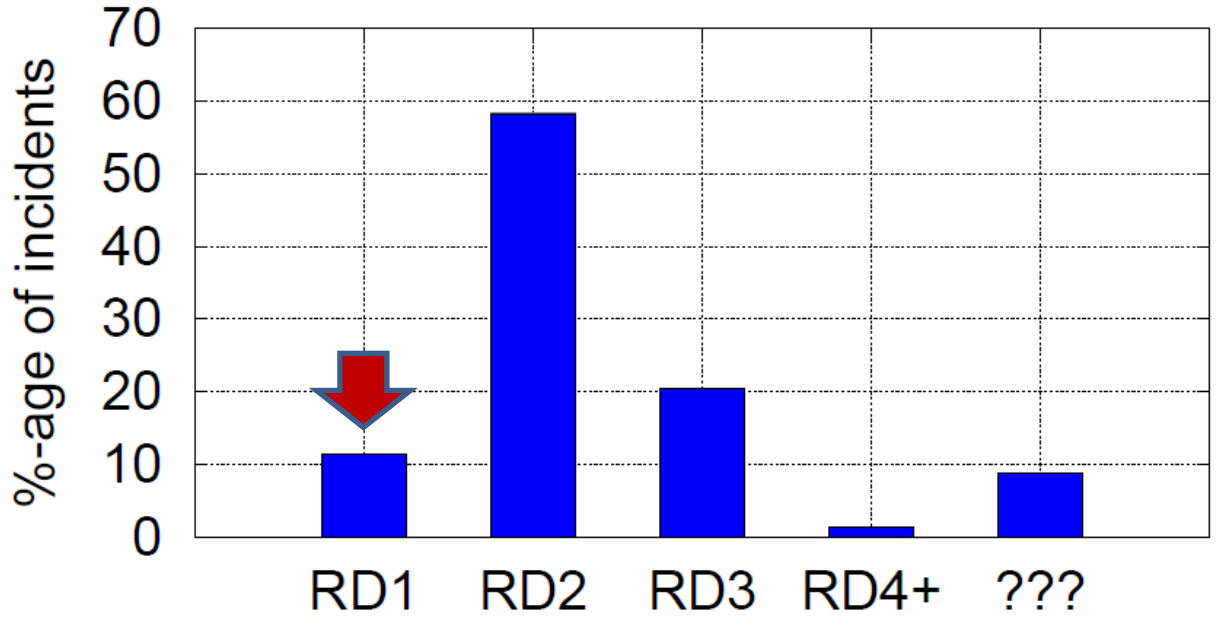
- Prevalence and Descriptions -

Deletion Prevalence

50,500 redactions; 5,600 suppressions



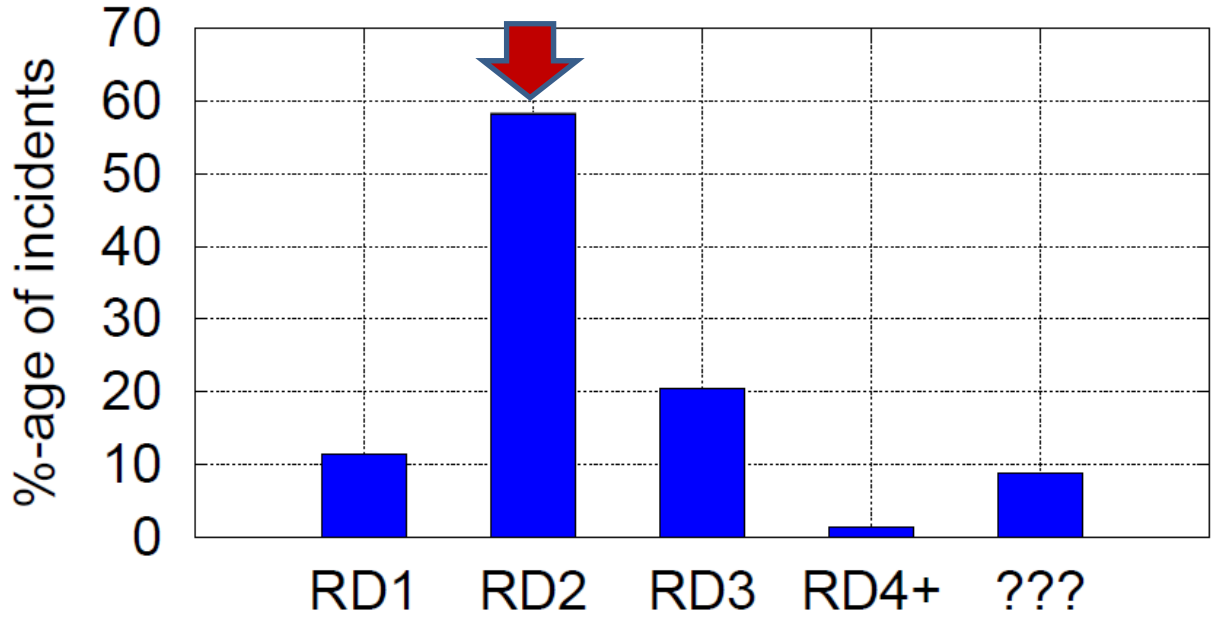
Deletion Reasons



RD1: Copyright violations

- Straightforward but hard to detect
- Often characterized by large insertions

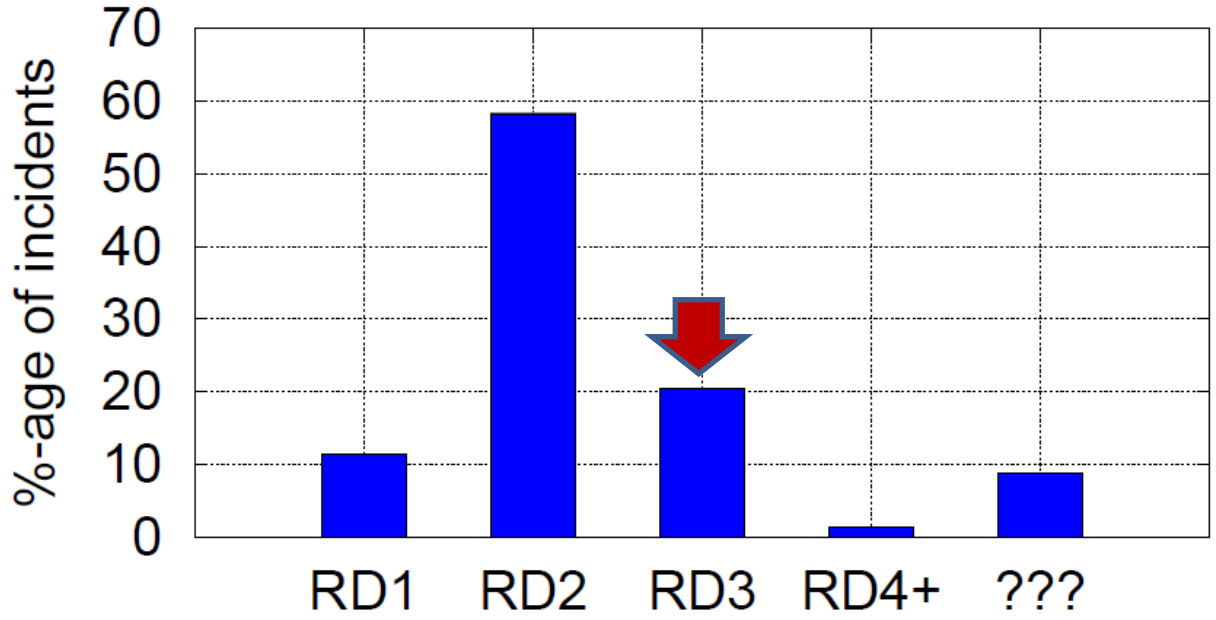
Deletion Reasons



RD2: Grossly insulting/offensive

- Name a specific person
- Un-sourced claims of promiscuity; pedophilia
- Racial slurs and profane language

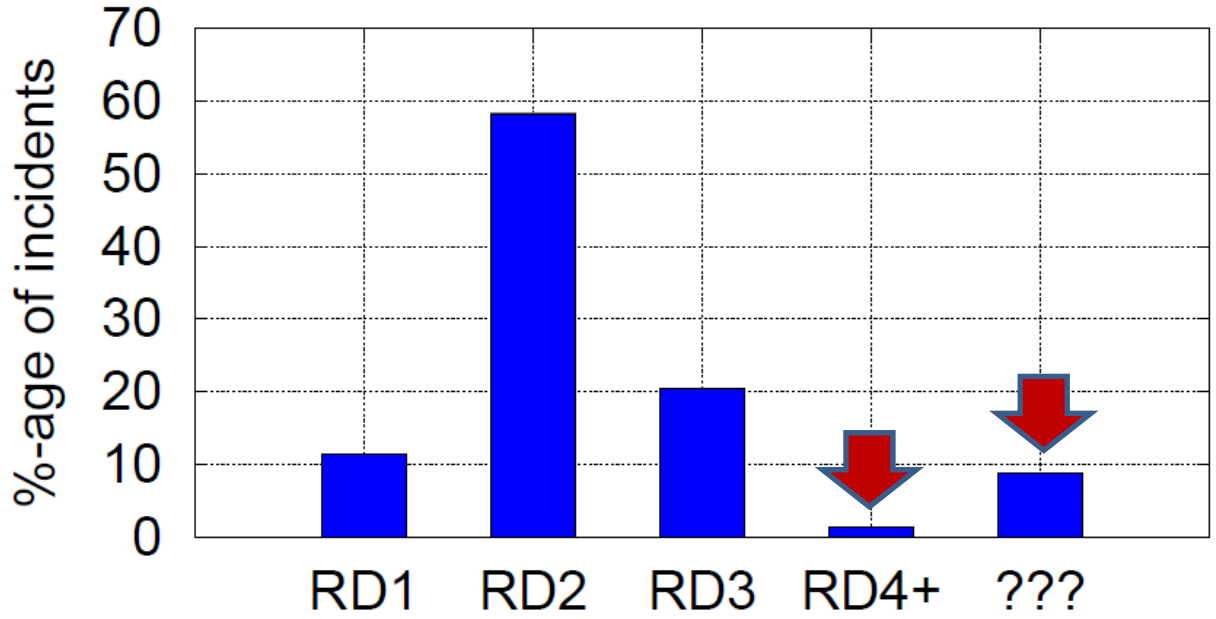
Deletion Reasons



RD3: Purely disruptive material

- Remarkably similar to RD2
- "Appropriately written falsities"; solicitations
- Massive insertions of random content

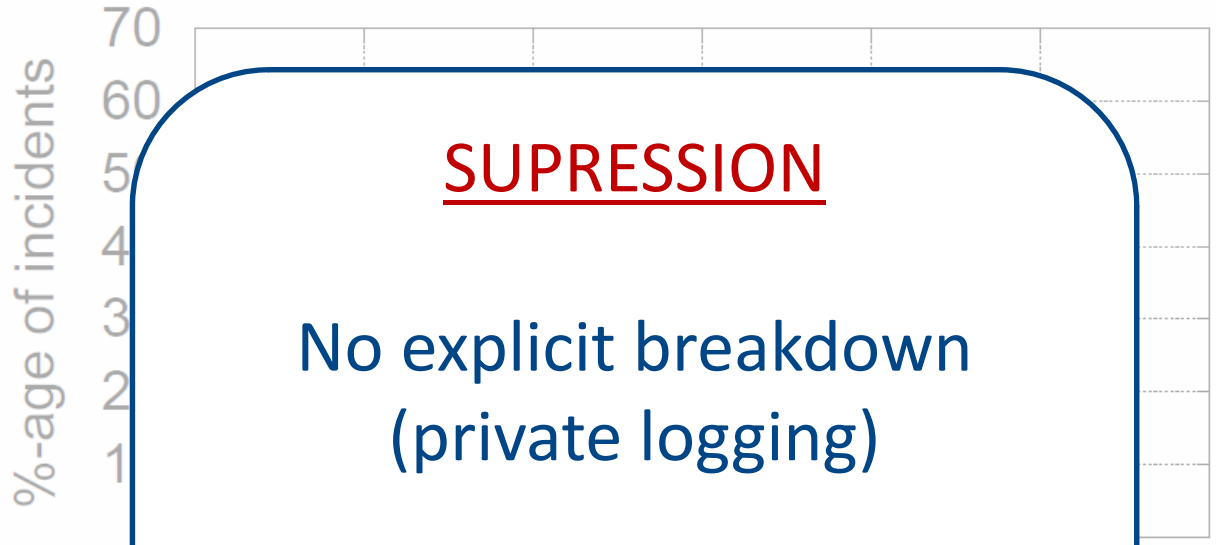
Deletion Reasons



RD4+ (“other”) + unknowns:

- RD4+ rare; given no further attention
- ???: Manual inspection reveals mix of types
- ???: Revisions awaiting suppression

Deletion Reasons



SUPRESSION

No explicit breakdown
(private logging)

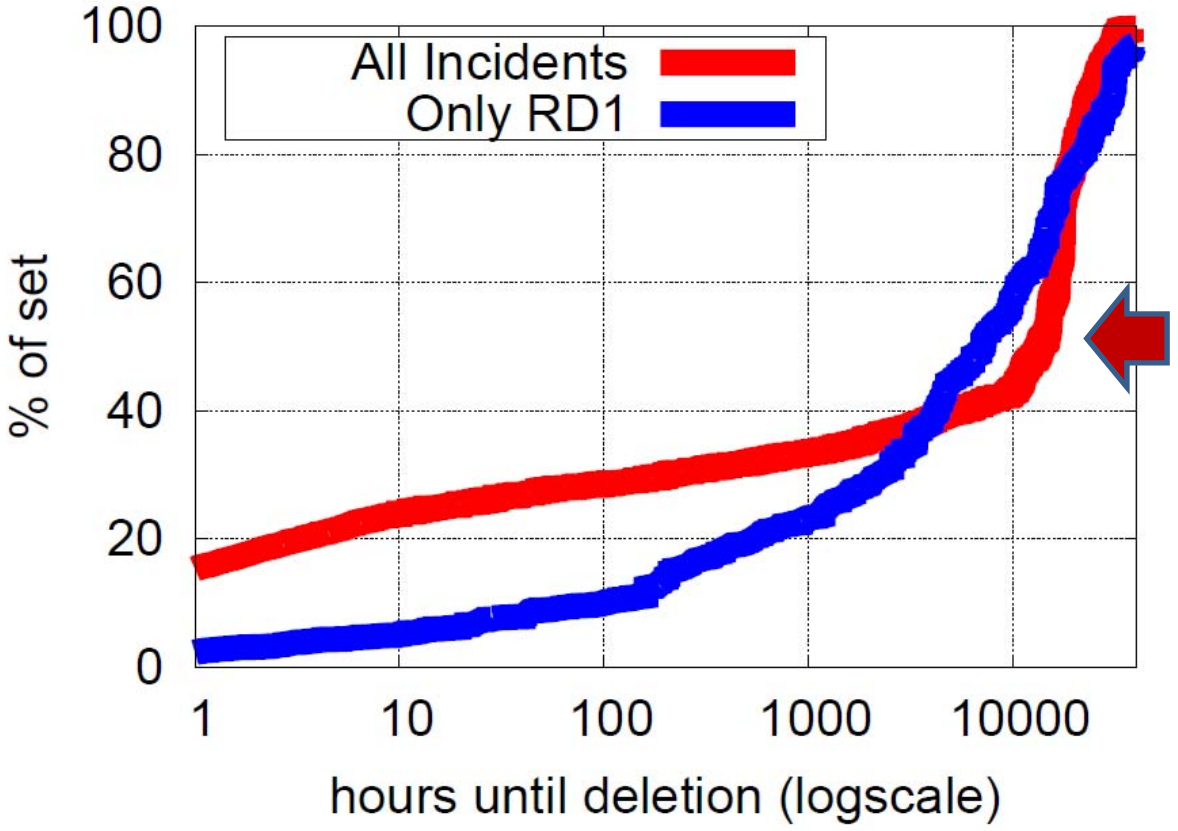
Informally:
- Phone numbers
- Physical addresses

RD4+ (
- R
- ???: Manual inspection reveals mix of types
- ???: Revisions awaiting suppression

ANALYZING REDACTIONS

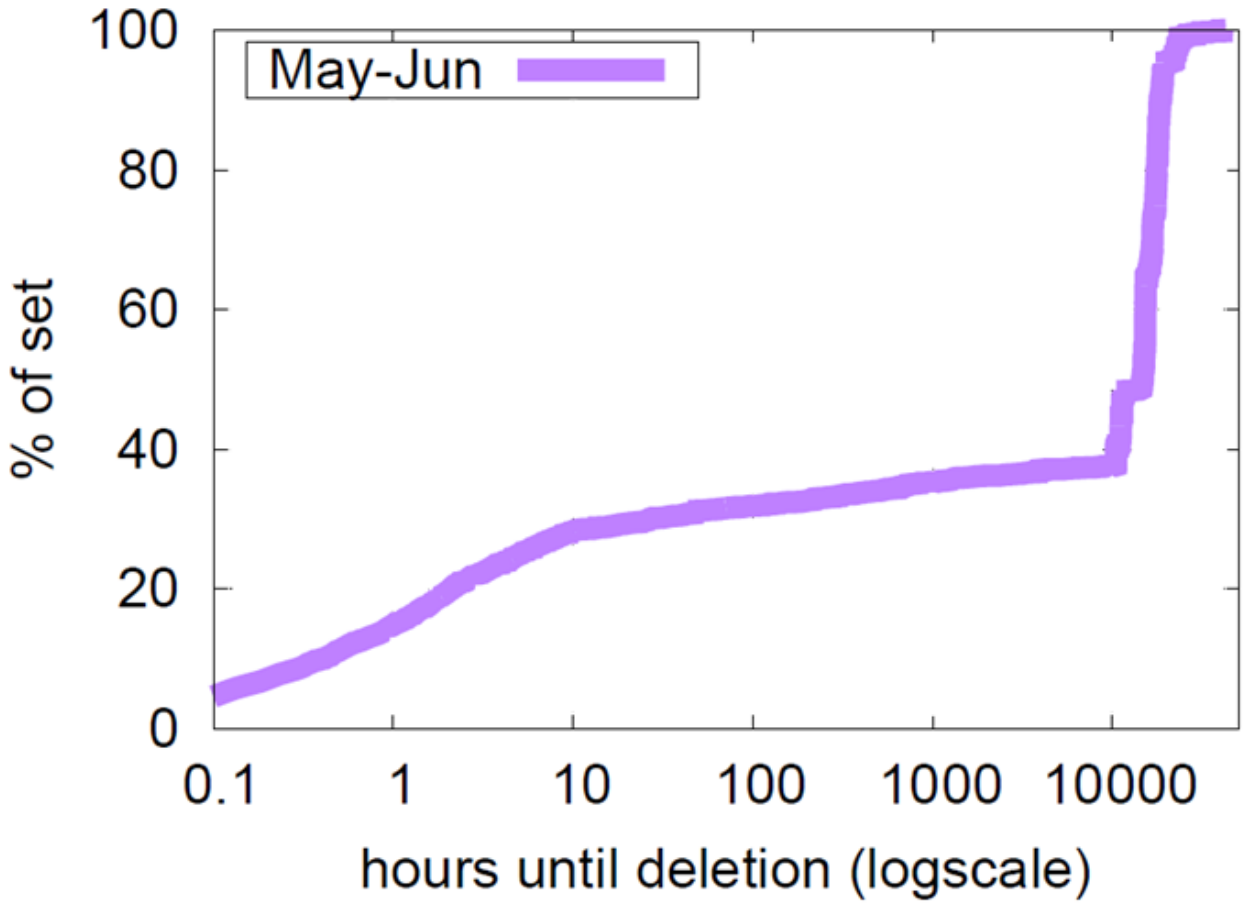
- Time to Deletion -

Time to Deletion

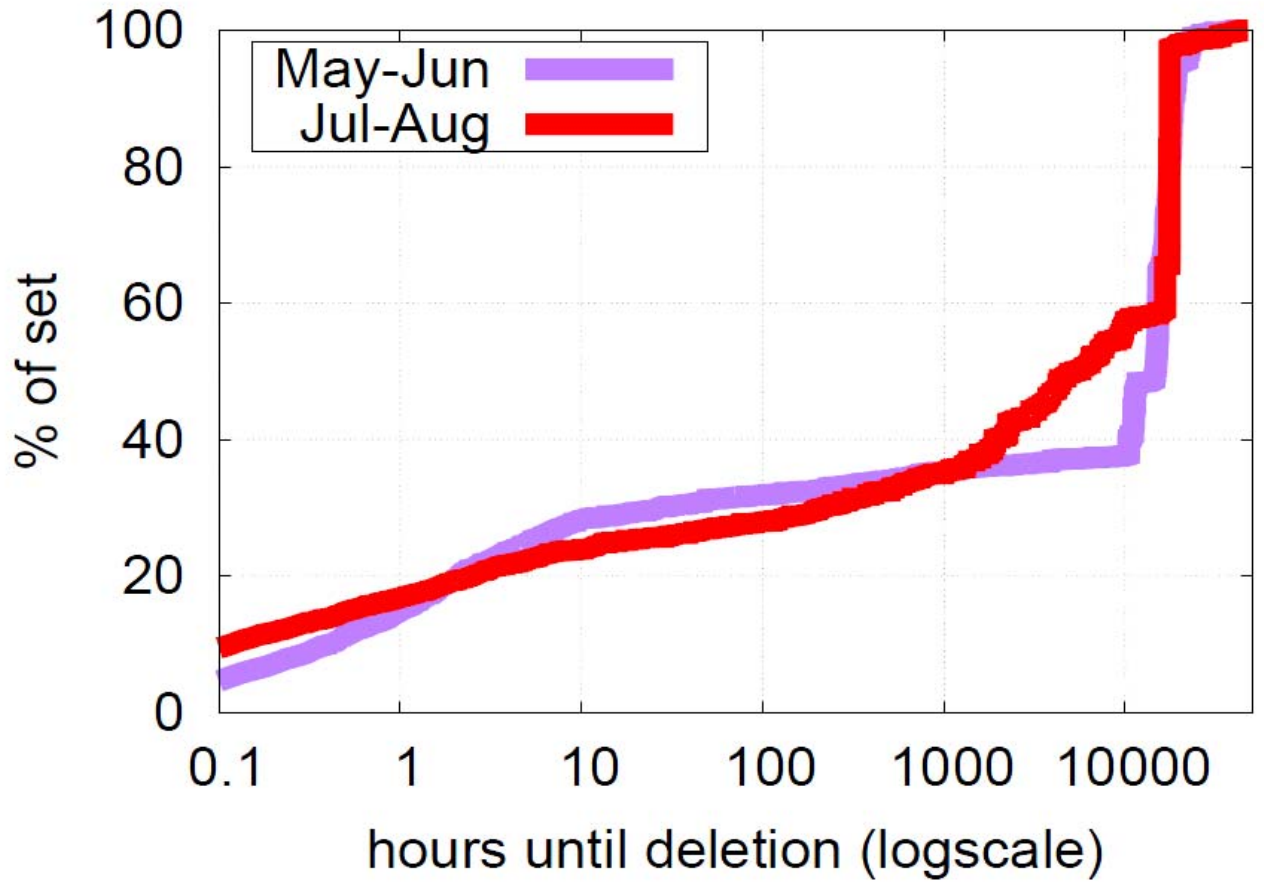


@ Median: 13,000 hours = 1.5 years

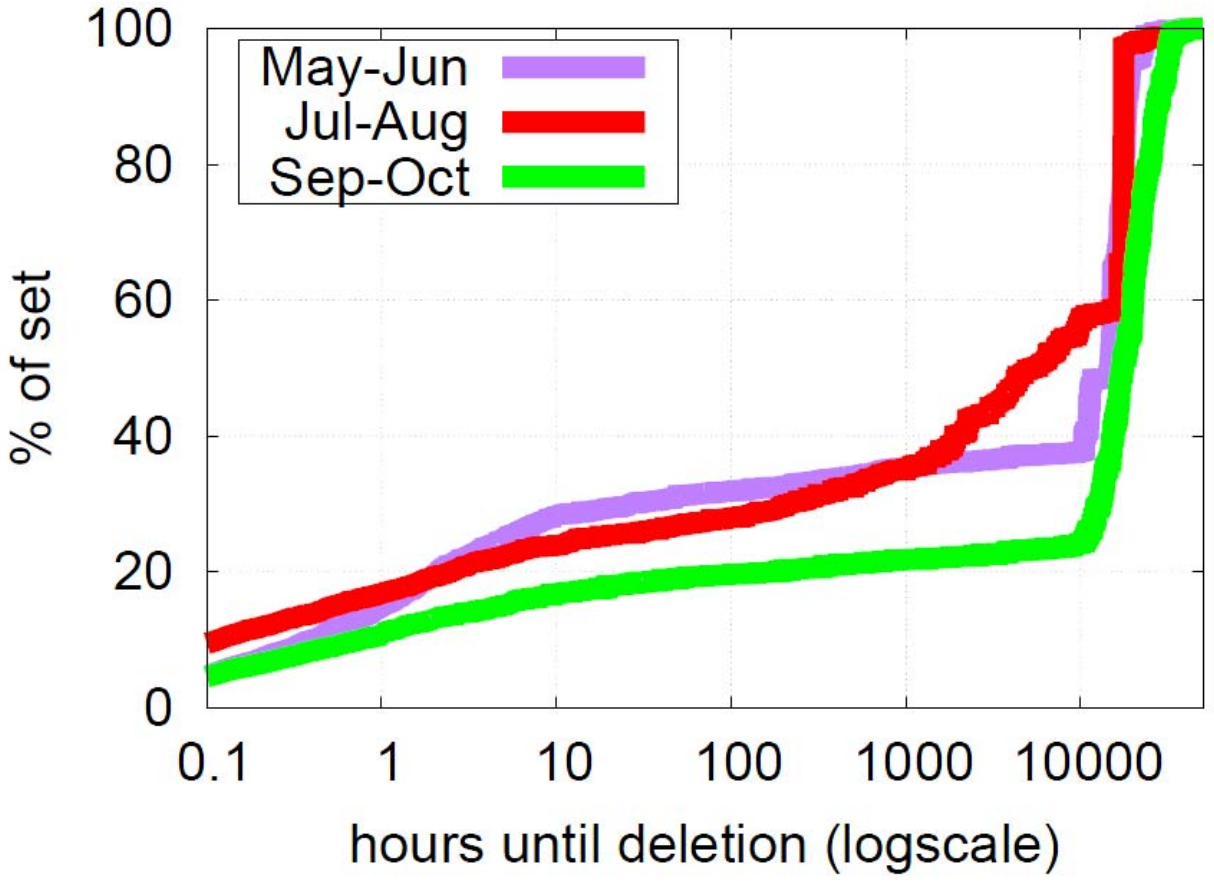
Time to Deletion (2)



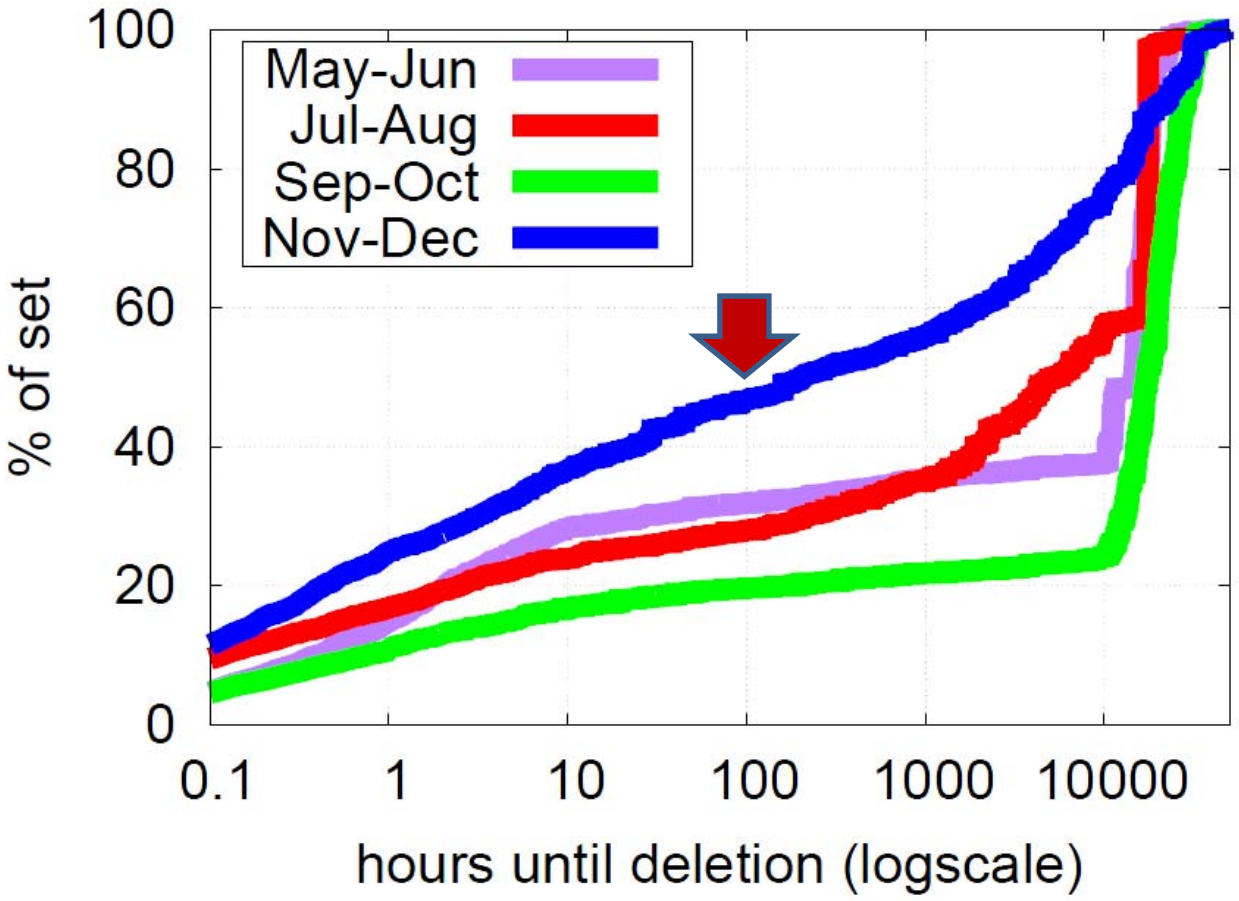
Time to Deletion (2)



Time to Deletion (2)



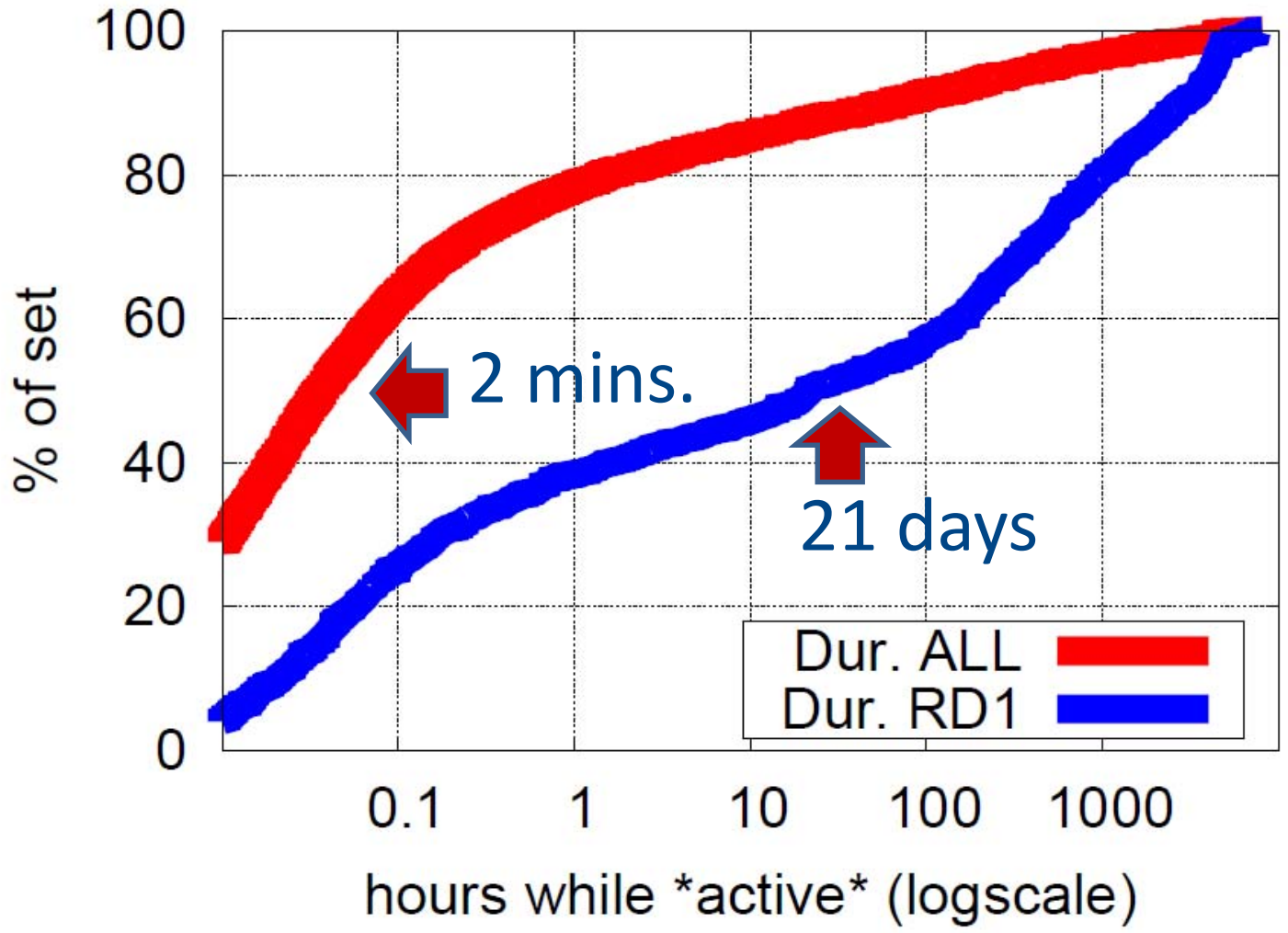
Time to Deletion (2)



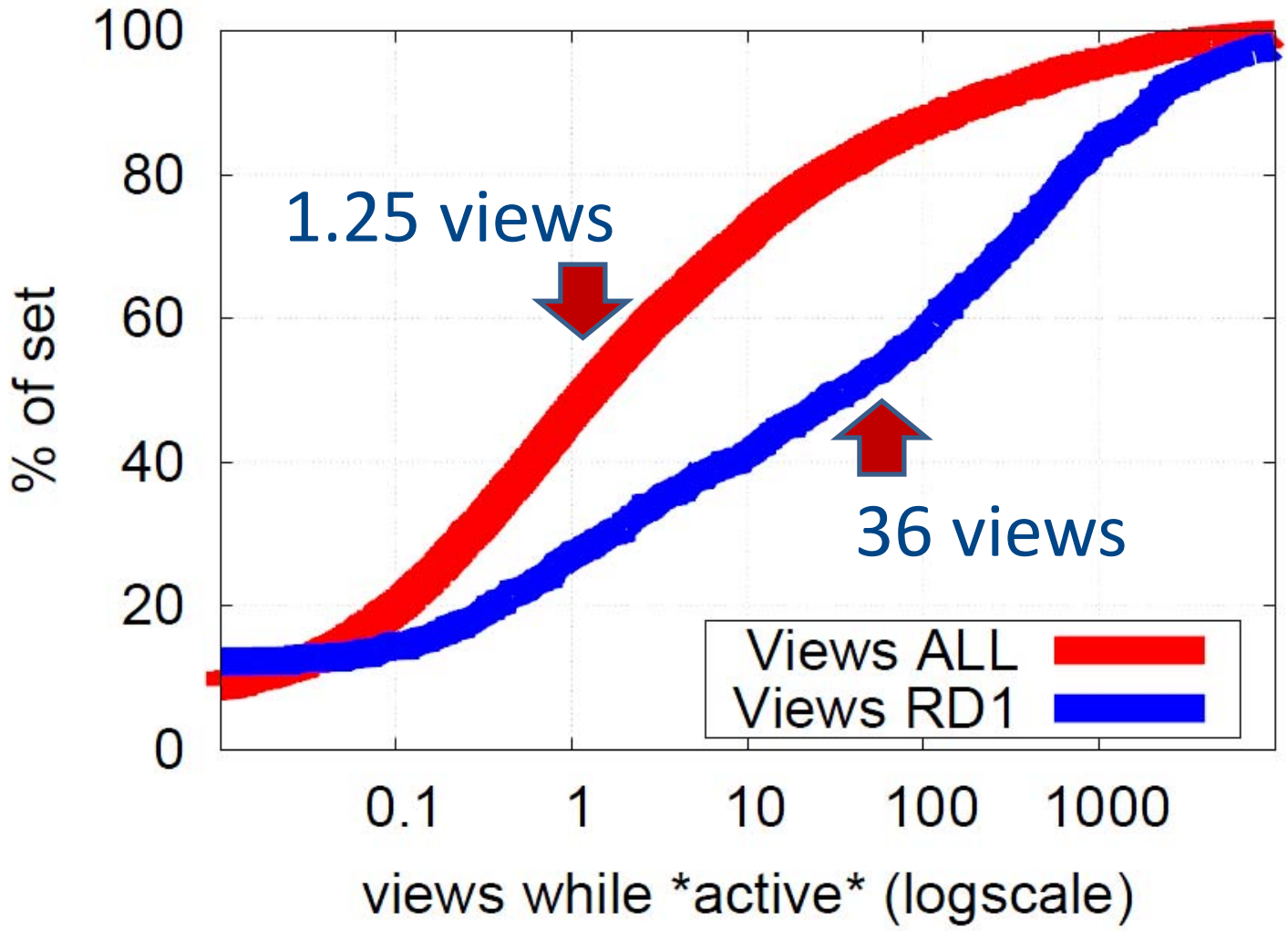
ANALYZING REDACTIONS

- Active Lifespan and Views -

Active Lifespan



Revision Views



RELATED WORK, IMPACTS, AND CONCLUSIONS

Related Work

- Deletion on Wikipedia
  - New tool; first writing on `RevDelete`
  - Article deletion much different (but well studied!)
- YouTube deletion study [4]
  - 0.4% video deletion; just 5% due to copyright
  - We concentrate on text, not multimedia
- Redaction-capable wiki [6] (*i.e.*, Intellipedia)
  - Proactive while Wikipedia's is *ex post facto*
- Writings about regulatory need for deletion capability to keep UGC sites from blacklisting [5, 9]

Impact/Conclusions

- EN.WP served 85 billion pages in 2010
  - ≈ 5.9 million views of dangerous revisions
  - Just 0.007% of views contained redacted content
- Copyright issues appear most challenging
  - Possible areas of future focus (*e.g.*, anti-plagiarism)
  - Disruption incidents, meanwhile, very reactive
- Open questions
  - Is the scope of deletion sufficient on legal grounds?
  - What percentage of eligible content is *not* deleted?
  - Non-textual content [8, 11] (*e.g.*, child porn)?

References

[01] Wikimedia API. http://en.wikipedia.org/w/api.php

[02] WP: Long-term abuse. http://en.wikipedia.org/wiki/WP:LTA

[03] WP: Revision deletion. http://en.wikipedia.org/wiki/WP:RVDL

[04] M. Cha, *et al*. **I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system**. In *IMC*, 2007.

[05] L. Edwards. **Content filtering and the new censorship**. In *ICDS '10: Proceedings of the Conference on Digital Society*, 2010.

[06] P. Gehres, N. Singleton, G. Louthan, and J. Hale**. Toward sensitive information redaction in a collaborative, multilevel security environment**. In *WikiSym*, 2010.

[07] E. Goldman. **Wikipedia's labor squeeze and its consequences**. *Journal of Telecommunications and High Technology Law*, 8, 2009.

[08] J. Merante. UK Natl. **Portrait Gallery threatens Wikipedia user over public domain images**. http://creativecommons.org/weblog/entry/15764 , 7/142009.

[09] B. Stone. **Policing the Web's lurid precincts**. *The New York Times*, B1, 7/18/2010.

[10] A. G. West, J. Chang, K. Venkatasubramanian, and I. Lee. **Link spamming Wikipedia for profit**. In *CEAS`11 (Collaboration, Electronic Messaging, Anti-Abuse, Spam)*.

[11] J. Winter. **Wikipedia distributing child porn, co-founder tells FBI**. FoxNews.com.

Backup Slides (1)

MO	RD1	RD2	RD3	RD4+	OTH	SUM
Jan.	2	11	0	1	9	23
Feb.	3	23	10	2	4	42
Mar.	25	31	3	1	27	87
Apr.	1	17	5	0	18	41
May	17	697	1006	2	97	1819
Jun.	37	913	427	37	101	1515
Jul.	88	718	1695	6	158	2665
Aug.	167	840	103	51	313	1474
Sep.	129	1846	161	18	193	2347
Oct.	252	5067	179	19	165	5682
Nov.	1087	535	112	14	215	1963
Dec.	338	323	152	84	352	1249
SUM	2146	11021	3853	235	1652	18907

Table 3: Deletion incidents (month × rationale)

Backup Slides (2)

REDACTED	NUM	%
content	13616	72.0%
summary	4082	21.6%
user	832	0.8%
summary + content	151	4.4%
user + content	51	0.3%
user + summary	14	0.1%
all fields	161	0.8%
TOTAL	18907	100.0%

Table 4: Redacted fields for incidents

Backup Slides (3)

CHANGES	#	%
Visibility increased	563	69%
Visibility decreased	188	23%
No visibility changes	40	5%
Orthogonal changes	25	3%
TOTAL	816	100%

Table 2: Visibility changes