

Towards Vandalism Detection in Knowledge Bases: Corpus Construction and Analysis

Stefan Heindorf Martin Potthast Benno Stein Gregor Engels

Motivation

Context

Information systems use structured knowledge bases

Problems

- Some people vandalize those knowledge bases
- Patrollers are busy
- Vandalism is not detected in time

Solution Idea

ML Machine learning to detect vandalism

Vandalism corpus

Contributions

- Wikidata Vandalism Corpus WDCV-2015
- Corpus analysis

Related work

Concentrates on *unstructured* knowledge bases

Corpus Construction

Automatic Revision Labeling

- Wikidata revision history
- Only non-bot revisions considered
- Goal: Automatic labeling as vandalism/non-vandalism

Option 1: Rollback (by administrators)
103,205 rollbacked revisions

Option 2: Undo/Restore (by all users)
64,820 undone/reverted revisions

Manual Validation

1,000 rollbacked revision
 1,000 undone/reverted revisions
 1,000 inconspicuous revisions

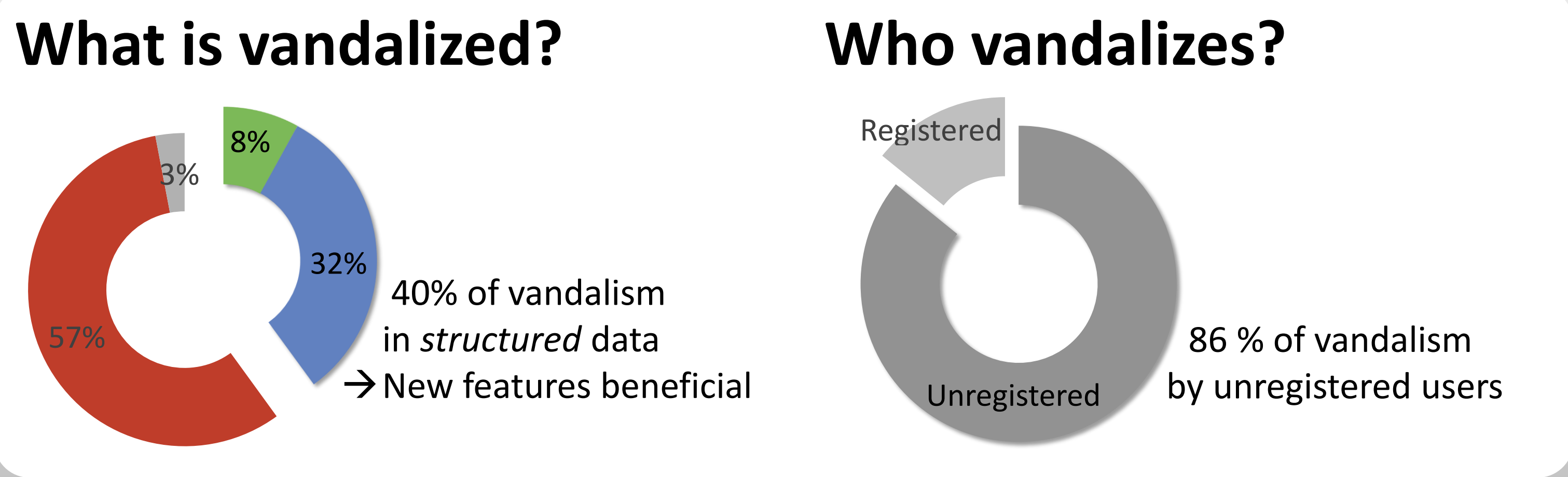
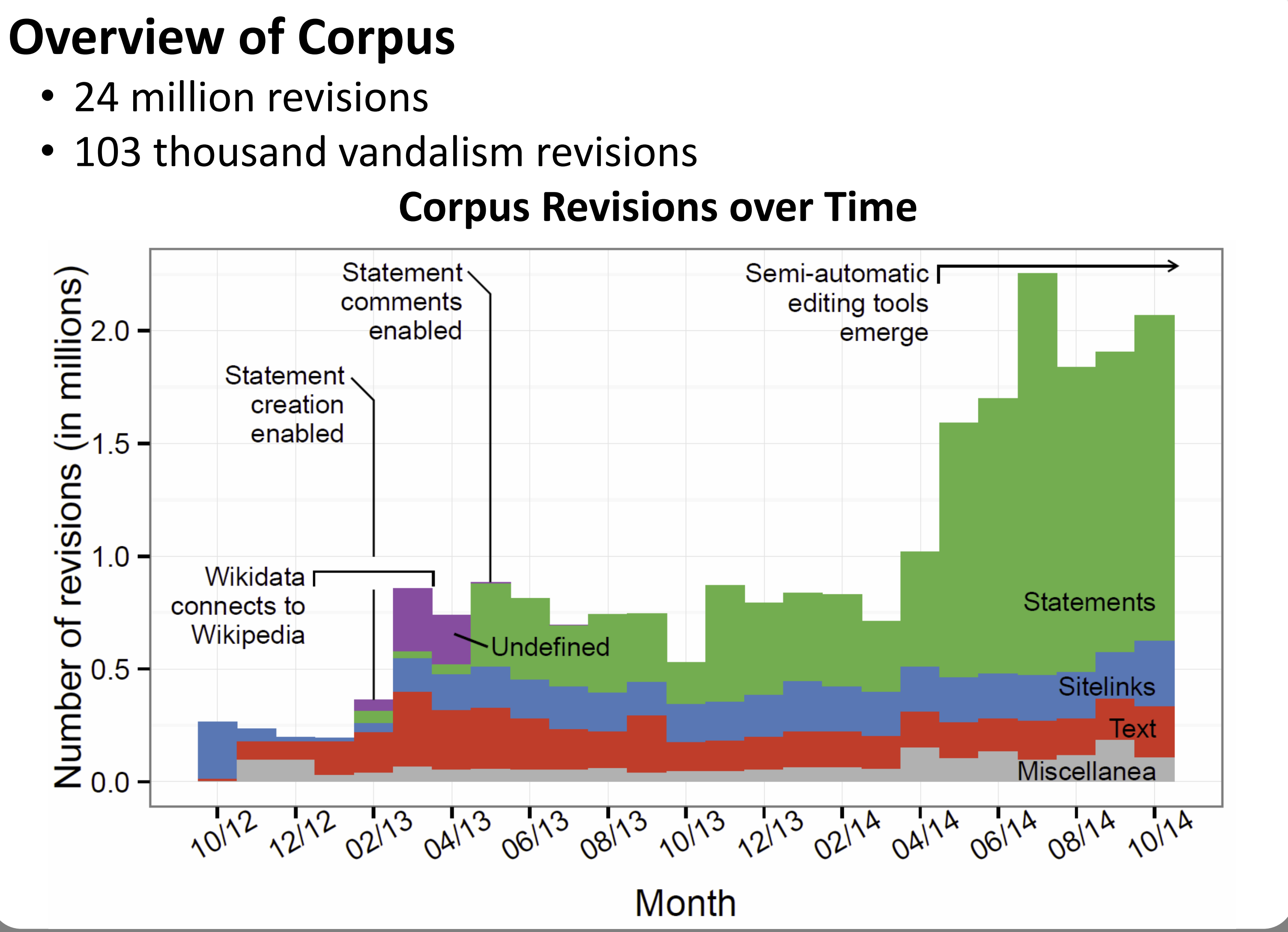
Option 1: Rollback (by administrators)
86 ± 3 %* revisions indeed vandalism

Option 2: Undo/Restore (by all users)
62 ± 3 %* revisions indeed vandalism

For our corpus: **Rollbacked revisions** are considered vandalism

* 95 % confidence level

Corpus Analysis



Top vandalized items

Cases	Item title
47	Cristiano Ronaldo
43	Lionel Messi
43	One Direction
41	Portal:Featured content
34	Justin Bieber
33	Barack Obama
29	English Wikipedia
29	Selena Gomez

Top items by category

Category	Top 1,000 vandalized items (%)	Top 1,000 of all items (%)
Culture	20%	12%
People	20%	21%
Society	16%	9%
Nature	14%	15%
Meta items	13%	8%
Technology	9%	4%
Places	8%	8%
Other	1%	1%

Conclusion & Outlook

- Vandalism can reduce the quality of knowledge bases
- First standardized corpus to study vandalism in *structured* knowledge bases
- Next step: detect vandalism automatically

Download our corpus!

