Privacy In Open Source
A Case Study: Wikipedia

Abu Saleh Md Noman
School of Informatics and Computing, Indiana University

What is there in Wikipedia?

- Wikipedia is the largest online Encyclopaedia that is **free and open** for all.
- Large database of contributors (both registered and **anonymous**).
- Every article is evolved through edits. Complete **edit history** is saved. Each page has talk and discussion page.

A typical article

HTTP://WEB.STANFORD.EDU/CLASS/CS345A/SLIDES/06-PROJECTS.PDF

Anonymous user in shadow

https://en.wikipedia.org/wiki/Wikipedia:IPs_are_human_too

Research Question

- What anonymous and registered people are doing (**type of edits**)?
- How people are reacting to those (**edit over edits**)?

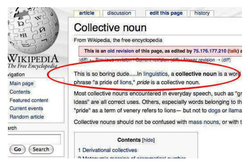
Why Is it Important?

- Data mining the type of contribution associated with the type of editor.
- Prevention of vandalism/foul edits.

What is Vandalism?

***'Vandalism** is the act of editing the project in a malicious manner that is intentionally disruptive'*

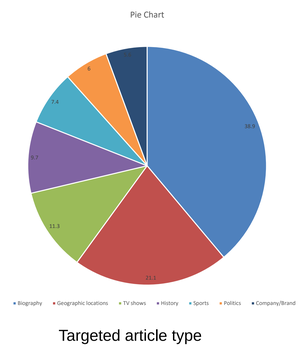
- Pattern in version history Metadata [1], edit category Classification [2], effect of editor Motivation on edit types [3] etc.
- However, this issue has not Been well studied



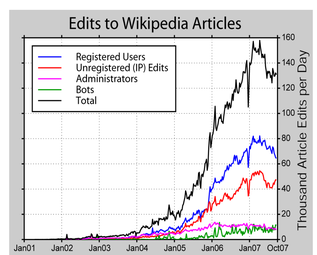
https://www.google.com/search?q=wiki+vandalism+study&rlz=1C1CHWA_enUS629US825&espv=2&tbm=1920&bih=893&site=webhp&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiMy9-ukprMAhVFFg4MKHaWUBCsQ_AUIBig

Results

Type of Edit:



Targeted article type



Number of edits by year*

- Summary of *vandalism* study:

Year	# of Articles	# of Edits	# of Vandalism	# Caused By Anonymous Users
2010	100	670	34	25
2011	80	584	29	18
2012	120	1015	59	45

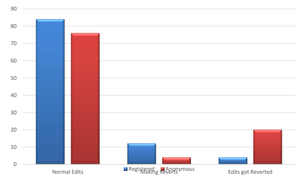
Number of Vandalism Reports

Study on user page, however, show different result. A study on a particular user page revealed 53:47(%) ratio!

90% of vandalism done by unregistered IP users.

Community Reaction:

If *i<j<k* in chronological order and *i=k*; then article j is a **revert.**



Activity (%)

Out of 150 randomly sampled articles, 32 were vandalized. ~25% of them were corrected in less than 90 seconds. The mean response time was **5 minutes.**

Community bias?

Methods

- Complete edit history: data dump available in Wikimedia (XML).
- Random sampling (Random article feature of Wiki)
- *Mwdumper,* a java tool for converting XML to SQL.
- Analysis performed in MySQL, Visualization in 'Listen to Wikipedia'.

Why Wikipedia Data?

- Dataset public and ideal for longitudinal study.
- Myriads of editors from different background
- Harmless dataset, no personal information.

Discussion

- Registered users account for most edits while anonymous users cause most vandalism→(~**80%** of IP editors are not vandals!).
- **Counter-vandalism unit** of Wiki reverts most of these vandalisms.
- Bias in study: restrict scope of analysis. Also, statistical testing are not performed.
- A detailed comprehensive study can reveal further trend and require rigorous data mining.

Future Work:

- Bag-of-words based ML classifier to identify vandals.
- Demographics of vandals, % of dynamic IPs, % of self-reverted vandalism etc..
- Other privacy issues of editors and factors influencing privacy loss.

References

[1] Maass, D. (2013). Data Mining Revision Controlled Document History Metadata for Automatic Classification.
[2] Daxenberger, J., & Gurevych, I. (2013, October). Automatically Classifying Edit Categories in Wikipedia Revisions. In*EMNLP*(pp. 578-589).
[3] Anthony, D., Smith, S. W., & Williamson, T. (2009). Reputation and reliability in collective goods the case of the online encyclopedia wikipedia.*Rationality and Society*,*21*(3), 283-306.

Thanks to Wikimedia project for the data and free tools available for use. * I thank Robert Rohde for his existing work on vandalism which is reused in this study.