Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Privacy Implications of Exposing Git Metadata

Arne Beer
Matriculation number: 6489196

July 10, 2018

Table of Contents

Main topic of the thesis

Is it possible to extract personal information?

- Used in most projects for version control

- Used in most projects for version control
- No obvious leak of personal information

- Used in most projects for version control
- No obvious leak of personal information
- Leaked information could be used maliciously

- Feasibility of scanning repositories
- Possible extraction of interesting information
- Analyse possible attack goals

```
tree      cd7d001b696db430b898b75c633686067e6f0b76
parent    c19b969705e5eae0ccca2cde1d8a98be1a1eab4d
author    Arne Beer <contact@arne.beer> 1513434723 +0100
committer Arne Beer <contact@arne.beer> 1513434723 +0100

Chapter 2, acronyms
```

- Parent commit file reference
- Tree file reference
- Name and email
- Commit timestamp with UTC offset

GitHub

- Largest accumulation of open-source Git repositories

GitHub

- Largest accumulation of open-source Git repositories
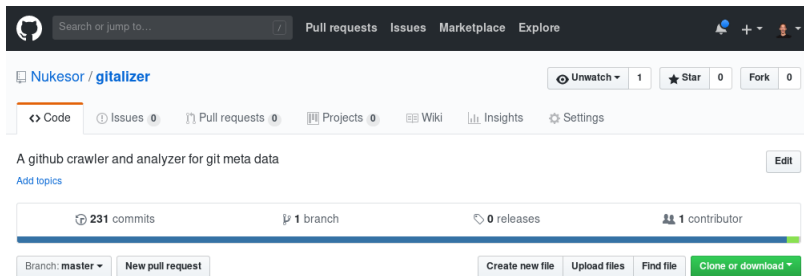- Great API

GitHub

- Largest accumulation of open-source Git repositories
- Great API
- Organizations

GitHub

- Largest accumulation of open-source Git repositories
- Great API
- Organizations
- Allows exploration for collecting user repositories.
    - Stars
    - Following

- Gathers data from Github
- Uses user or organization as entry point
- Highly optimized

- Holiday and sick leave detection

- Holiday and sick leave detection
- Sleep rhythm and working hours

- Holiday and sick leave detection
- Sleep rhythm and working hours
- Geographic location

- Detect holiday or sick leave

- Detect holiday or sick leave
- Detect other anomalies

- Detect holiday or sick leave
- Detect other anomalies
- Good visualization

Example
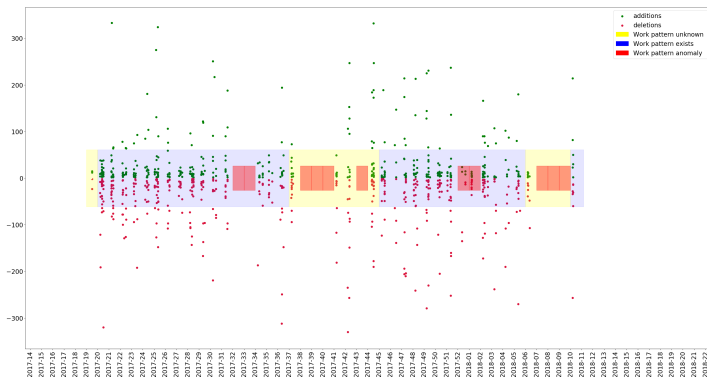

Figure: Holiday and Sick leave visualization

- Tested in a small company

- Tested in a small company
- Accurate holiday detection

- Tested in a small company
- Accurate holiday detection
- Some false positives

- Tested in a small company
- Accurate holiday detection
- Some false positives
- Other anomaly detection needs interpretation

- Good visualization

- Good visualization
- Check visibility of sleep rhythm

- Good visualization
- Check visibility of sleep rhythm
- Check link to working behavior
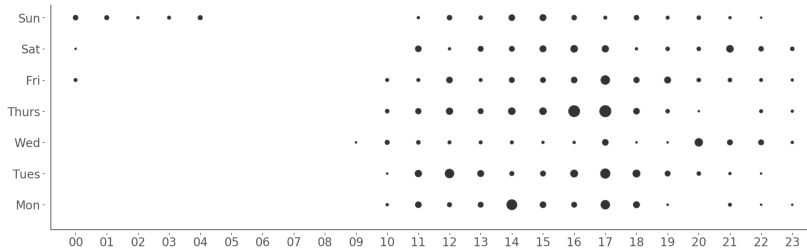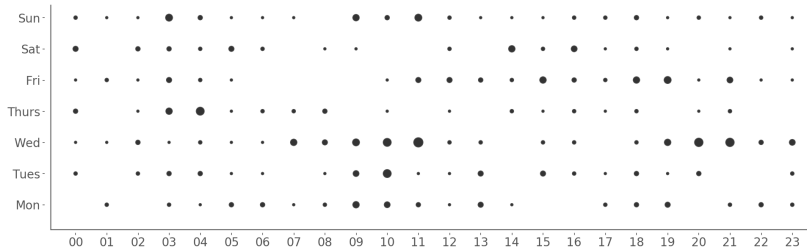
Figure: Regular working hour cluster

Figure: Person without a sleep rhythm

- Tested for correctness in small test group

- Tested for correctness in small test group
- Sleep rhythm rather accurate

- Tested for correctness in small test group
- Sleep rhythm rather accurate
- Allows to detect working behaviour

- Detect home country

- Detect home country
- Detect holiday destinations
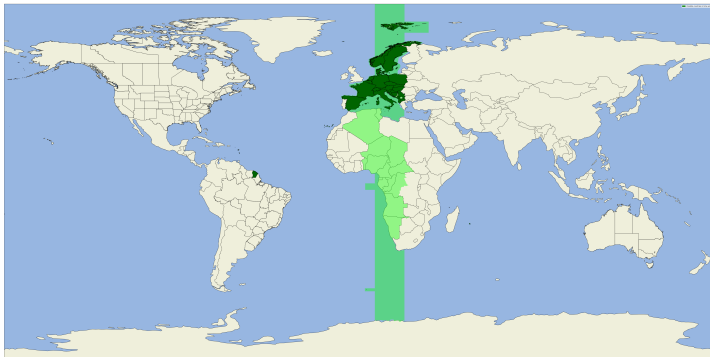
Figure: Home location analysis

- Good detection of home country

- Good detection of home country
- Holiday not checked

- Good detection of home country
- Holiday not checked
- Needs better libraries

- Recall the goal: Is it possible to extract personal information

- Recall the goal: Is it possible to extract personal information
- It is possible to extract further personal information

- Recall the goal: Is it possible to extract personal information
- It is possible to extract further personal information
- Scanning on small to middle scale

- Many more complex and promising attacks

- Many more complex and promising attacks
- It could become a problem

- Many more complex and promising attacks
- It could become a problem
- Methodologies to obfuscate data

Thank you for your attention.