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

Topic

Main topic of the thesis

Is it possible to extract personal information?

Motivation

- Used in most projects for version control

Motivation

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

Motivation

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

Leading Question and Goals

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

Git Metadata

```
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

Why Github?

GitHub

- Largest accumulation of open-source Git repositories

GitHub

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

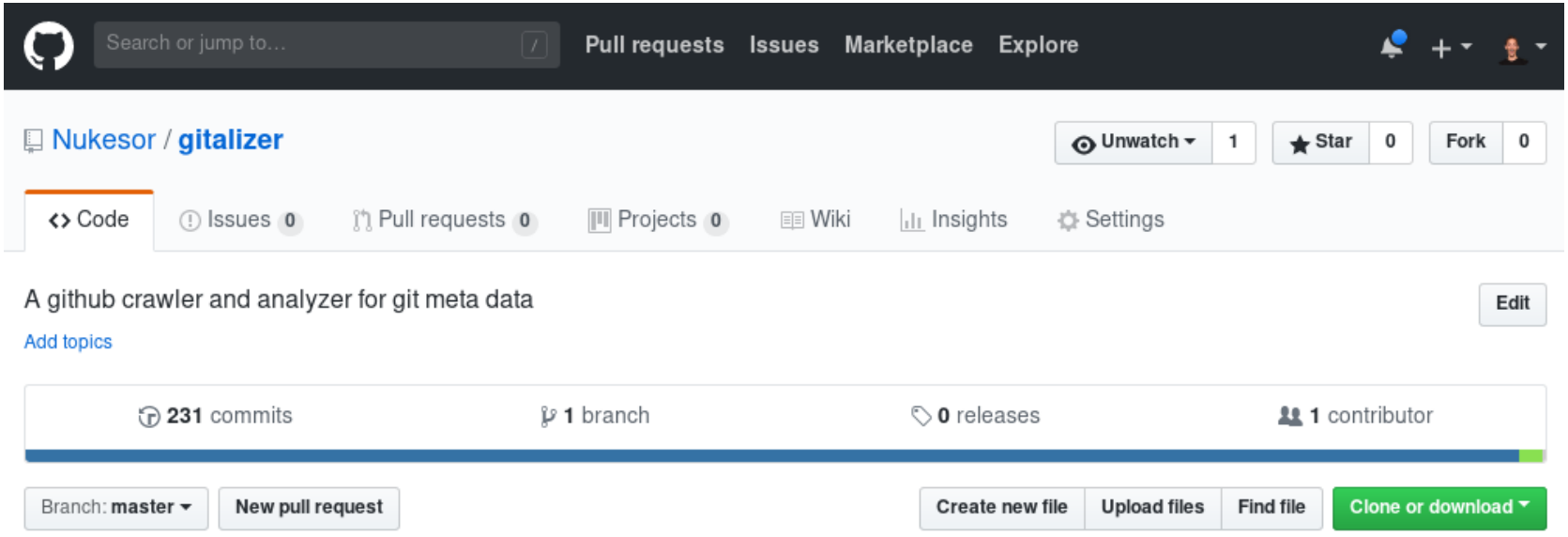
GitHub

- 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

Gitalizer



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

Three Chosen Attacks

- Holiday and sick leave detection

Three Chosen Attacks

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

Three Chosen Attacks

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

Holiday and Sick Leave: Goals

- Detect holiday or sick leave

Holiday and Sick Leave: Goals

- Detect holiday or sick leave
- Detect other anomalies

Holiday and Sick Leave: Goals

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

Methodology

- All commits of one contributor
- Find regular work pattern
- Detect unregular work pattern
- Detect anomalies

Example

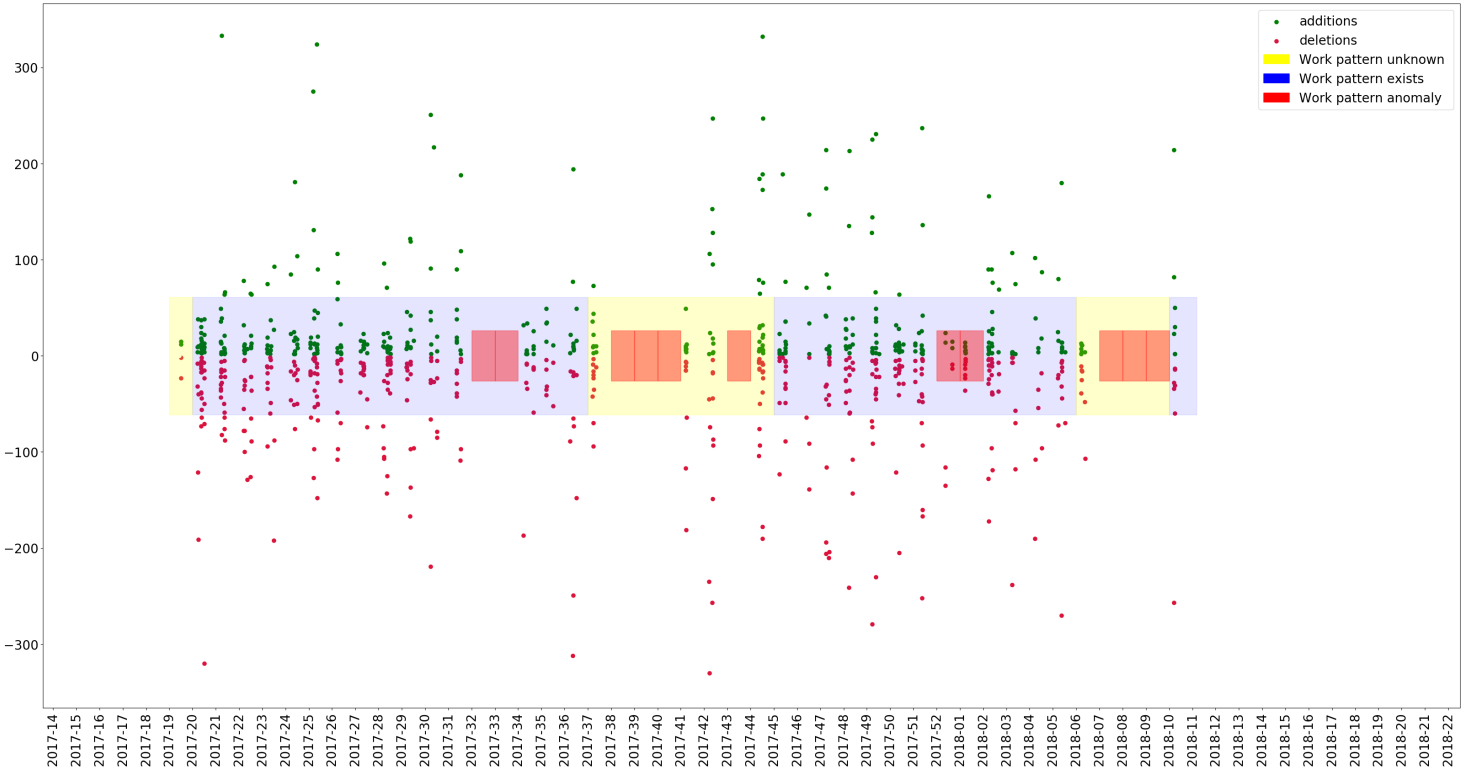


Figure: Holiday and Sick leave visualization

Results

- Tested in a small company

Results

- Tested in a small company
- Accurate holiday detection

Results

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

Results

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

Sleep Rhythm and Working Hours: Goals

- Good visualization

Sleep Rhythm and Working Hours: Goals

- Good visualization
- Check visibility of sleep rhythm

Sleep Rhythm and Working Hours: Goals

- Good visualization
- Check visibility of sleep rhythm
- Check link to working behavior

Example

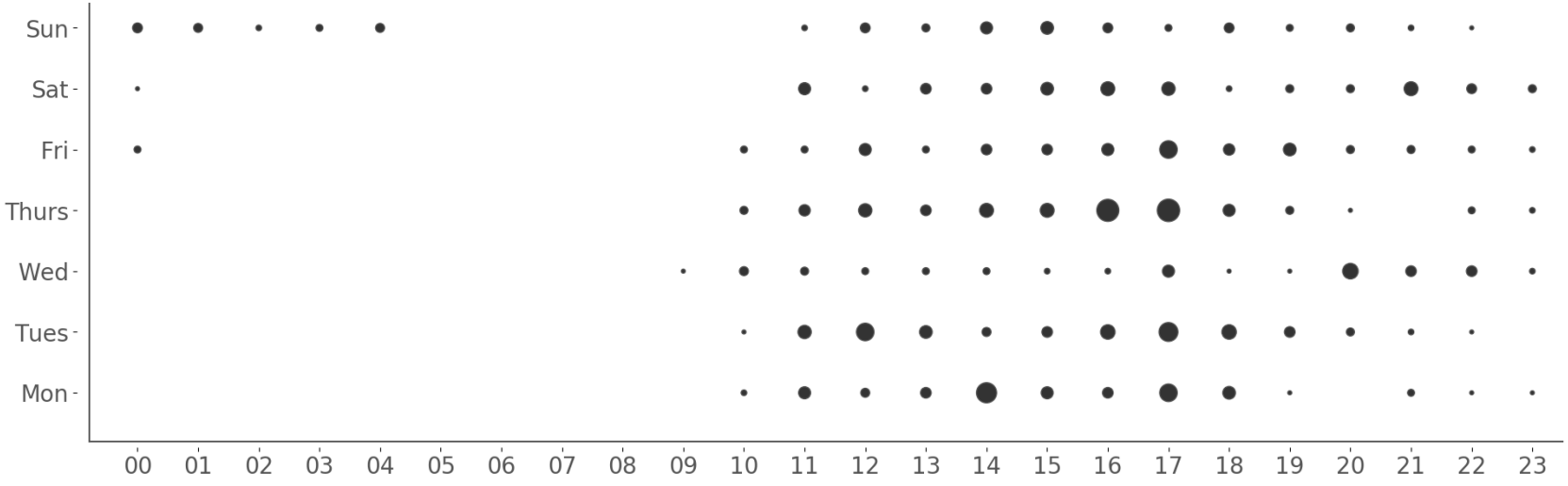


Figure: Regular sleep rhythm

Example

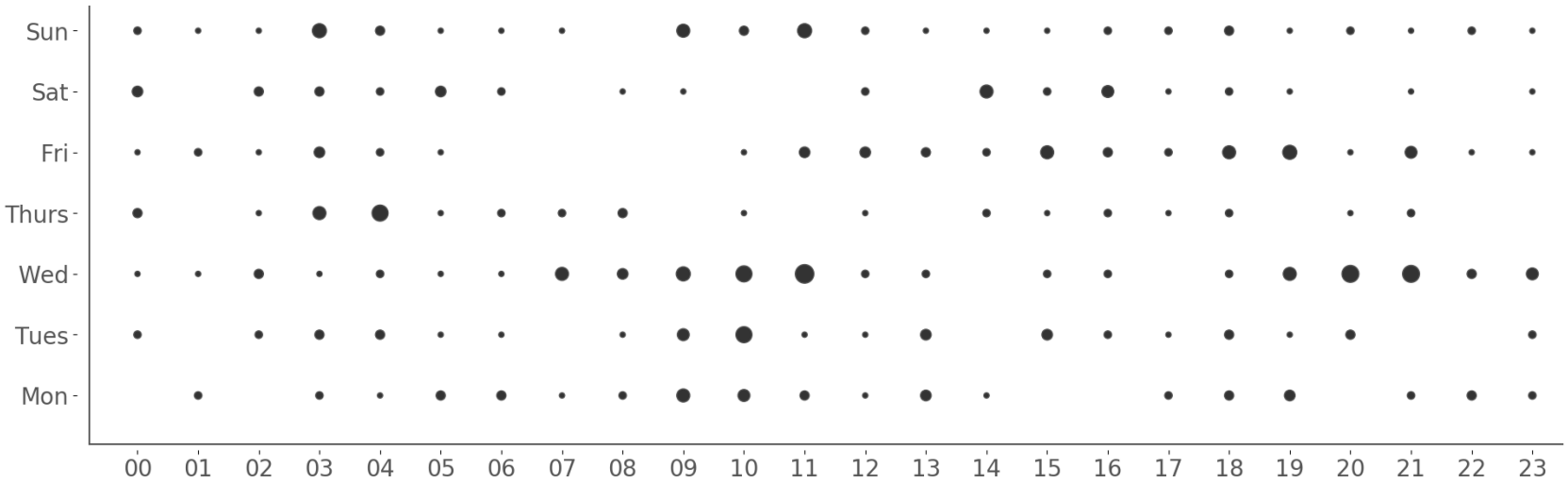


Figure: Person without a sleep rhythm

Example

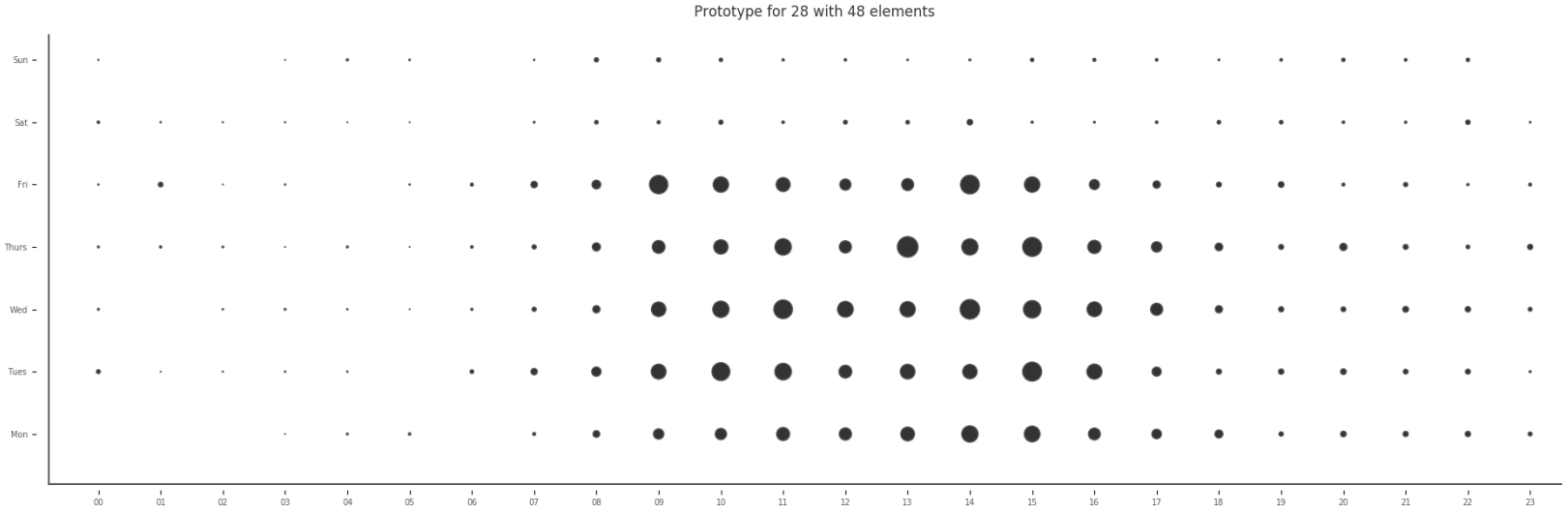


Figure: Normal working hour punchcard

Results

- Tested for correctness in small test group

Results

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

Results

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

Geographic Location: Goals

- Detect home country

Geographic Location: Goals

- Detect home country
- Detect holiday destinations

Example

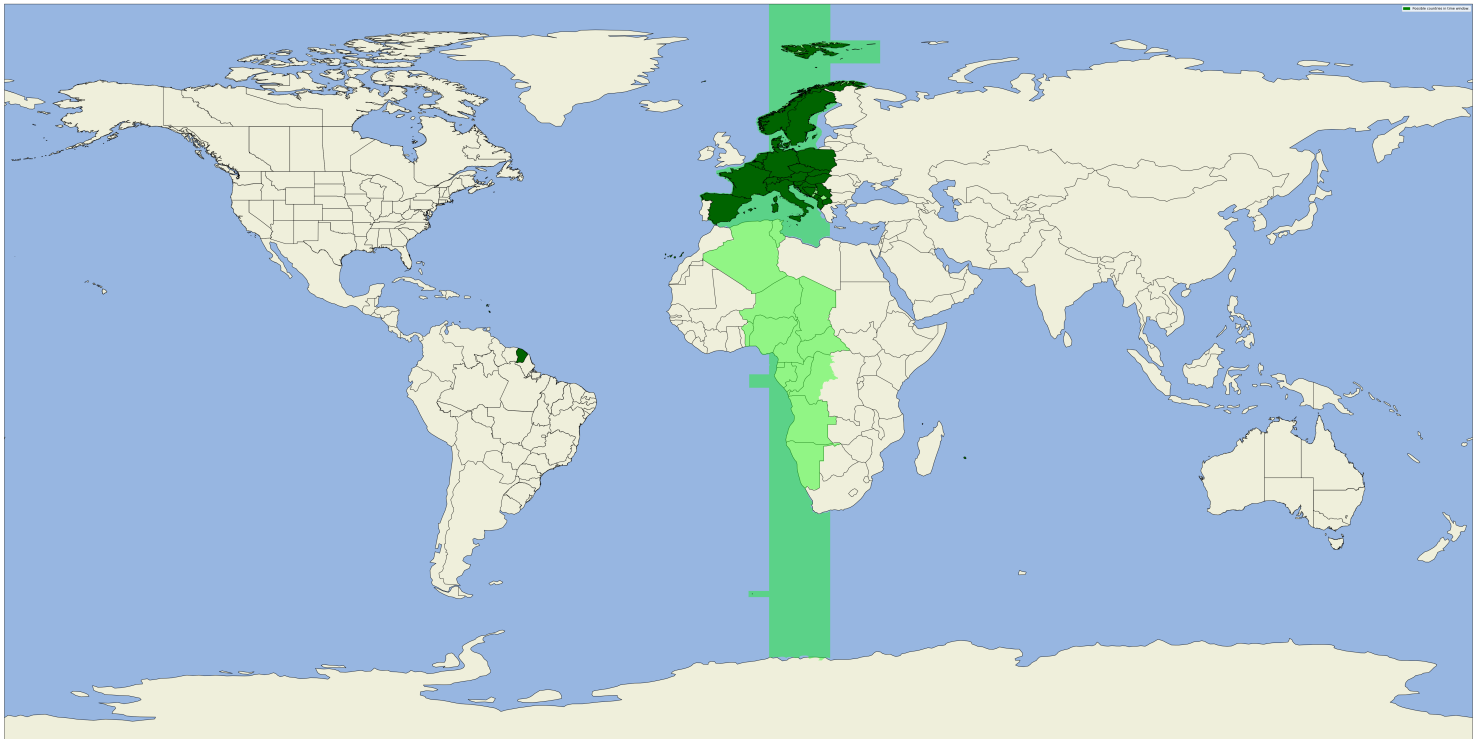


Figure: Home location analysis

Methodology

- All commits of one contributor
- Possible timezones/countries for offset
- Daylight Savings Time
- Detect timezone switches

Results

- Good detection of home country (82%)

Results

- Good detection of home country (82%)
- Holiday not checked

Results

- Good detection of home country (82%)
- Holiday not checked
- Needs better libraries

Conclusion

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

Conclusion

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

Conclusion

- Recall the goal: Is it possible to extract personal information
- It is possible to extract further personal information
- Simple goals, but already offers possibly sensitive information

Outlook

- Many more complex and promising attacks

Outlook

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

Outlook

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

Fin

Thank you for your attention.