Privacy Implications of Exposing Git Meta Data

Arne Beer
Matriculation number: 6489196

Department of Computer Science
University of Hamburg

July 5, 2018

Table of Contents

- Introduction
  - Topic
  - Motivation
  - Leading Question and Goals
- Aggregation
  - Data Source
- Attack Models
  - Attack Models
  - Attacks
- Research
  - Holiday and Sick Leave
  - Sleep Rhythm and Working Hours
  - Geographic Location
- Conclusion and Outlook
  - Conclusion
  - Outlook

Topic

Main topic of the thesis

Can Git metadata be maliciously used?

Motivation

- Used in nearly every project

Motivation

- Used in nearly every project
- No obvious leak of personal information

Motivation

- Used in nearly every project
- No obvious leak of personal information
- Possible workplace/contributor surveillance

Leading Question and Goals

- Feasibility of scanning repositories on different scales
- Possible extraction of interesting information
- Possible attack vectors

Why Github?

- Largest accumulation of open-source Git repositories

Why Github?

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

Why Github?

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

Exploration

- Repository ownership
- Stars
- Following

Gitalizer

- Crawls Github
- Starts at user or company
- Highly optimized

The Three Attack Models

- Employer

The Three Attack Models

- Employer
- Individual

The Three Attack Models

- Employer
- Individual
- Industrial Spy

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

- Automatic detection

Holiday and Sick Leave: Goals

- Automatic detection
- Accurate detection

Example

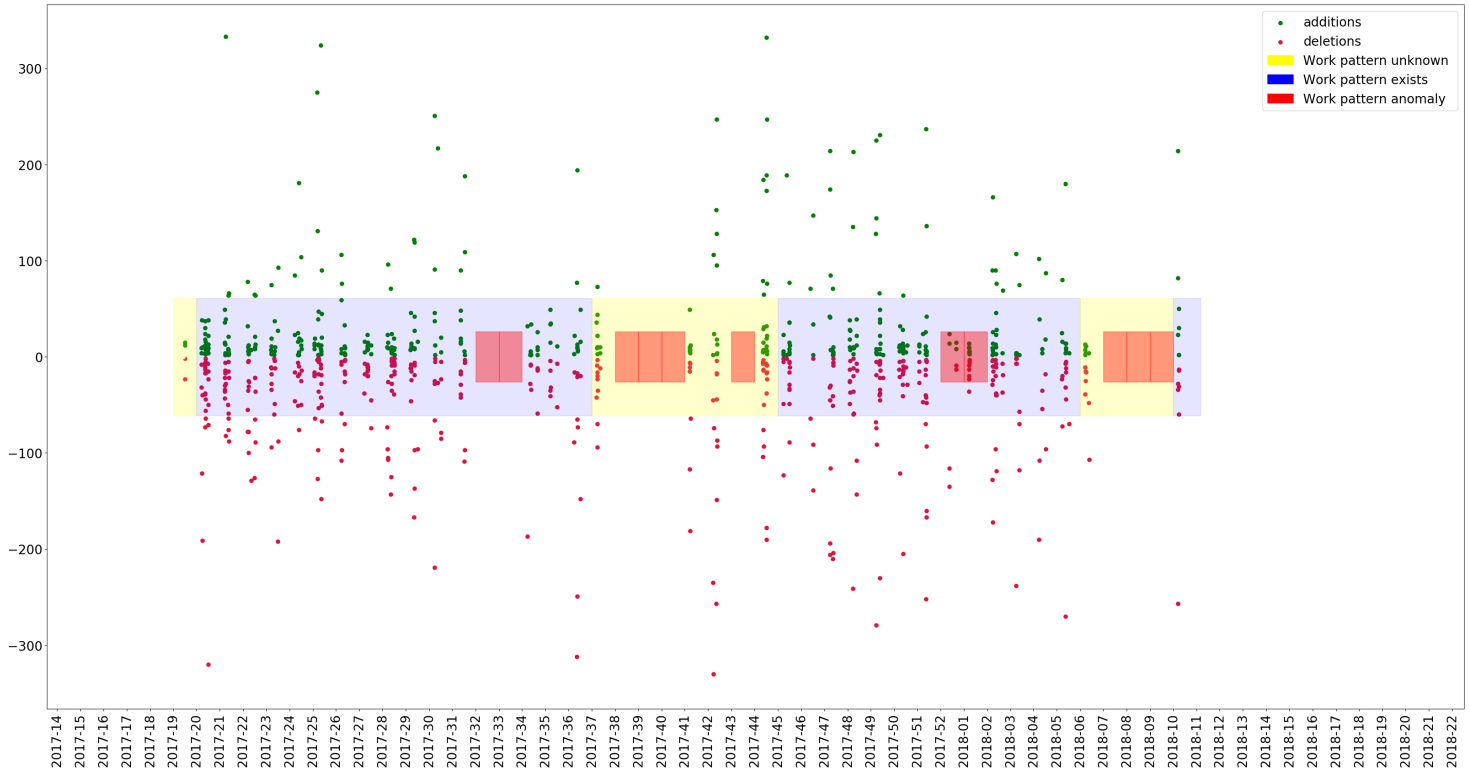

Figure: Holiday and Sick leave visualization

Results

- Tested in a small company

Results

- Tested in a small company
- Quite accurate

Results

- Tested in a small company
- Quite accurate
- Needs interpretation

Sleep Rhythm and Working Hours: Goals

- Detection

Sleep Rhythm and Working Hours: Goals

- Detection
- Good visualization

Sleep Rhythm and Working Hours: Goals

- Detection
- Good visualization
- Further implications of rhythm

Example

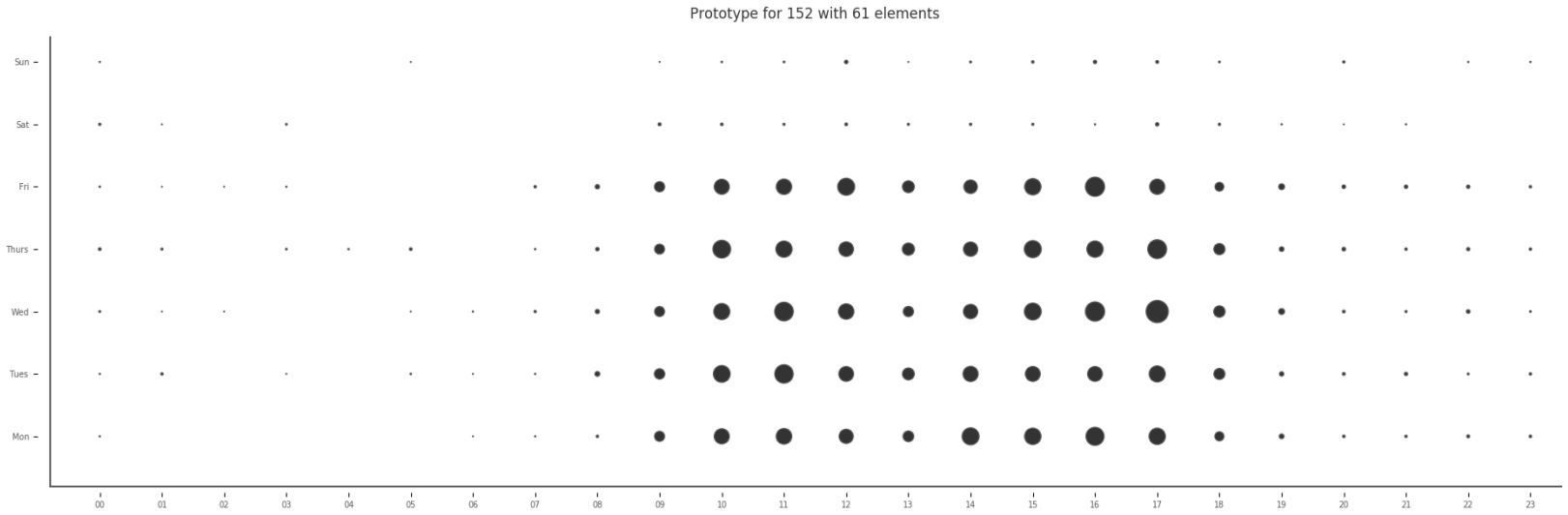

Figure: Regular working hour cluster

Example

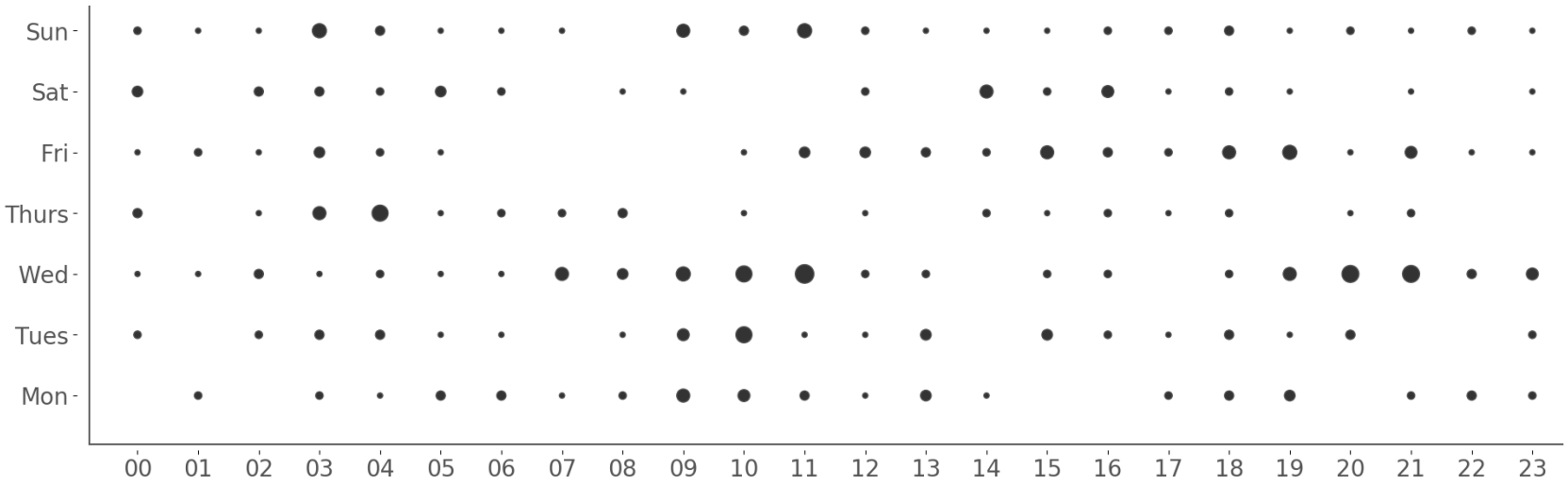

Figure: Person without a sleep rhythm

Results

- Shows general tendency

Results

- Shows general tendency
- Unsuitable for direct personal mapping

Results

- Shows general tendency
- Unsuitable for direct personal mapping
- Allows to guess further information

Geographic Location: Goals

- Detect holiday destinations

Geographic Location: Goals

- Detect holiday destinations
- Detect home country

Geographic Location: Goals

- Detect holiday destinations
- Detect home country
- Detect time periods

Methodology

- Periodically check commits

Methodology

- Periodically check commits
- Daylight Savings Time

Example

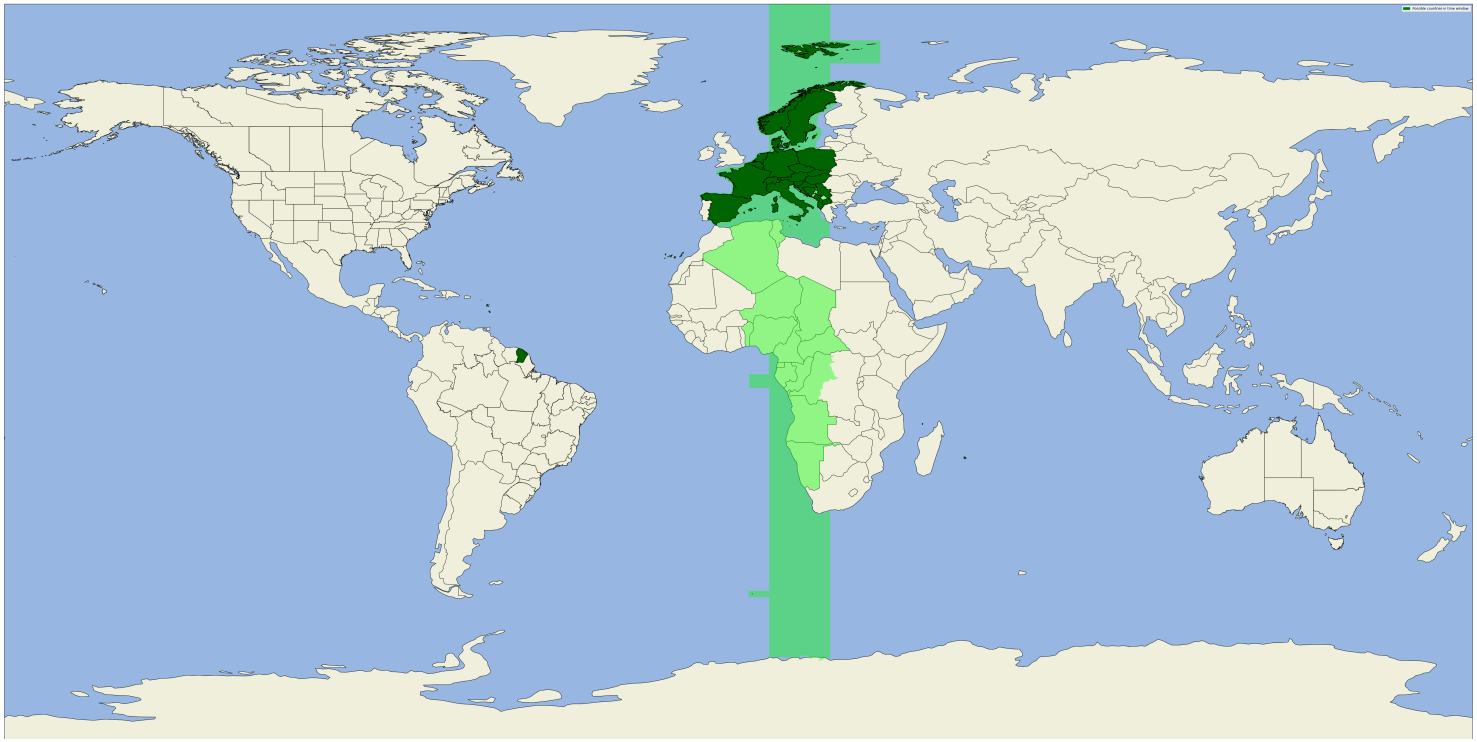

Figure: Home location analysis

Results

- Good detection of home country

Results

- Good detection of home country
- Holiday not checked

Results

- Good detection of home country
- 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
- Scanning on small to middle scale

Outlook

- It can become a problem

Outlook

- It can become a problem
- Many more complex and promising attack vectors

Outlook

- It can become a problem
- Many more complex and promising attack vectors
- Methodologies to obfuscate data

Fin

Thank you for your attention.