Fortgeschrittene Softwaretechnik
Wintersemester 2024/25
Name: Simon Szulik
Matr.Nr: 1474315

Datum: 7. Januar 2025
Reading Group Sessions
Inhaltsskizze zu Block: 05

Titel des Papiers

The Promises and Perils of Mining GitHub

Kurzzusammenfassung

- **Hintergrund und Motivation**
  - GitHub, mit über 10,6 Millionen Repositories (Stand Januar 2014), ist die größte Plattform für kollaborative Softwareentwicklung und eine zentrale Quelle für Forschungsdaten in der Softwaretechnik.
  - Ziel der Studie: Untersuchung der **Forschungsfrage**, welche *Versprechen* und *Gefahren* mit der Nutzung von GitHub-Daten für die Forschung verbunden sind.
- **Methodik und Datengrundlage**
  - Datenbasis: GHTorrent-Datensatz, manuelle Analyse von 434 zufällig ausgewählten Repositories und eine Umfrage mit 240 GitHub-Nutzern.
  - Analyse: Kombination aus quantitativer Analyse der Metadaten (z.B. Commits, Pull-Requests) und qualitativer manuelle Überprüfung.
- **Gefahren**
  - **Unterschiedliche Zwecke:** Viele Repositories sind keine Softwareprojekte, sondern dienen z.B. als Speicher (8,3 %) oder für experimentelle Zwecke (12,2 %).
  - **Inaktivität:** 46 % der Projekte waren in den letzten 6 Monaten inaktiv, 32 % hatten nur einen einzigen Tag Aktivität.
  - **Pull-Requests:** Nur 10 % der Projekte nutzen Pull-Requests, und deren Verwendung ist oft inkonsistent (z.B. alternative Merge-Strategien).
  - **Externe Infrastruktur:** Viele aktive Projekte führen Teile ihrer Entwicklungsarbeit außerhalb von GitHub durch (z.B. auf Mailinglisten oder externen Issue-Trackern).
  - **Verzerrungen:** Persönliche Repositories (71,6 %) und inaktive Projekte dominieren GitHub und beeinflussen die Aussagekraft der Daten.
- **Versprechen**
  - **Reichhaltige Datenquelle:** GitHub bietet umfassende Daten zu Commits, Pull-Requests, Issues und sozialer Interaktion zwischen Entwicklern.
  - **Code-Review-Analysen:** Pull-Requests und Diskussionen erlauben detaillierte Untersuchungen zu Code-Review-Prozessen.
  - **Integration von Daten:** Verknüpfungen zwischen Issues, Pull-Requests und Commits ermöglichen eine ganzheitliche Analyse von Entwicklungsaktivitäten.

Eigene (offene) Diskussionspunkte

- **Datenrepräsentation:** Die Studie nutzt hauptsächlich quantitative Daten, jedoch könnten qualitative Analysen der Nutzerinteraktionen vertieft werden, um die Bedeutung sozialer Funktionen besser zu verstehen.
- **GHTorrent-Abhängigkeit:** Die Ergebnisse basieren stark auf der GHTorrent-Datenbank, deren Vollständigkeit und Aktualität nicht garantiert ist. Alternativen könnten in Betracht gezogen werden.
- **Externe Projekte:** Projekte, die teilweise außerhalb von GitHub arbeiten, werden nicht detailliert untersucht. Wie könnte dies die Aussagekraft der Ergebnisse beeinflussen?
- **Pull-Requests-Bias:** Der Fokus auf Pull-Requests könnte andere Kollaborationsmechanismen auf GitHub, wie Issues oder direkte Commits, unterrepräsentieren.
- **Langzeitwirkung:** Wie könnte sich die steigende Popularität von GitHub und die Veränderung von Nutzungsmustern langfristig auf die Forschungsgrundlage auswirken?