Relatório Final - Características de repositórios populares

INTRODUÇÃO E HIPÓTESES INICIAIS:

A atividade de revisão de código, especialmente por meio de *pull requests* (PRs), é essencial para garantir a qualidade, a colaboração e a manutenção sustentável em projetos open-source. Em repositórios populares no GitHub, o fluxo de revisão pode variar conforme o tamanho da contribuição, a clareza da descrição, o tempo de análise e o número de interações. Com base nesse contexto, este estudo tem como objetivo analisar padrões nas revisões de PRs em 200 projetos com alta visibilidade, avaliando duas dimensões centrais: **feedback final das revisões** (aceitação ou rejeição do PR) e **quantidade de revisões realizadas** antes da decisão.

Utilizando dados extraídos da API GraphQL do GitHub, o estudo investigará como variáveis como o tamanho do PR, o tempo de análise, a qualidade da descrição e as interações influenciam tanto no resultado da revisão quanto na sua complexidade (medida pelo número de ciclos de revisão).

Com base na intuição sobre projetos open-source populares, formulamos as seguintes hipóteses informais:

- **H1:** Pull Requests com menos de 300 linhas modificadas tendem a ser aceitos com uma taxa superior a 80%.

- **H2:** Pull Requests analisados em menos de 2 dias têm maior chance de aceitação, com taxa superior a 75%.

- **H3:** Pull Requests com descrições contendo mais de 100 palavras são mais aceitos, com taxa de aprovação superior a 85%.

- **H4:** Pull Requests com mais de 5 interações (comentários e revisões) apresentam taxa de aceitação superior a 80%.

- **H5:** Pull Requests com mais de 500 linhas modificadas tendem a passar por mais de 2 ciclos de revisão.

- **H6:** Pull Requests cujo tempo de análise ultrapassa 5 dias costumam passar por mais de 3 ciclos de revisão.

- **H7:** Pull Requests com descrições acima de 100 palavras geralmente passam por menos de 2 ciclos de revisão.

- **H8:** Pull Requests com mais de 10 comentários tendem a passar por mais de 3 ciclos de revisão.

Essas hipóteses serão avaliadas por meio da coleta e análise dos dados extraídos da API GraphQL do GitHub, permitindo a identificação de padrões e tendências entre os projetos mais populares da plataforma.

METODOLOGIA

Coleta de Dados: Utilizamos a API GraphQL do GitHub para extrair dados de Pull Requests revisados em repositórios populares. As informações coletadas incluem tamanho das alterações (linhas), tempo de análise, descrição do PR, interações (comentários e revisões), status final (merge ou rejeição) e número de revisões.

Armazenamento: Os dados foram salvos em um arquivo `.csv` estruturado, facilitando a leitura e organização para análise posterior.

Processamento: Para evitar distorções causadas por valores extremos, utilizamos a mediana como principal medida estatística. Agrupamos os dados por faixas (ex.: tamanho pequeno, médio e grande) para facilitar a comparação.

Classificação: Categorias foram definidas para cada métrica (ex.: tempo curto, médio e longo) com base na distribuição dos dados.

Análise Comparativa: Investigamos a relação entre as métricas analisadas (tamanho, tempo, descrição e interações) e dois fatores principais: o feedback final da revisão (merge ou rejeição) e o número de revisões realizadas por PR.

RESULTADOS

RQ 01. Qual a relação entre o tamanho dos PRs e o feedback final das revisões?
Métrica: linhas modificadas (adicionadas + removidas) e taxa de merge

- Média de linhas adicionadas: 176.5
- Mediana de linhas adicionadas: 61.3
- Média de linhas removidas: 120.2
- Mediana de linhas removidas: 22.3
- Média de taxa de merge: 71.0%
- Mediana da taxa de merge: 75.0%

RQ 02. Qual a relação entre o tempo de análise dos PRs e o feedback final das revisões?
Métrica: tempo médio de análise (em horas) e taxa de merge

- Média de tempo de análise: 1002.6 horas
- Mediana de tempo de análise: 152.0 horas
- Observação: projetos com menor tempo de análise (ex: < 48h) têm taxas de merge mais altas

RQ 03. Qual a relação entre a descrição dos PRs e o feedback final das revisões?
Métrica: tamanho médio da descrição (em número de palavras) e taxa de merge

- Média de tamanho da descrição: 2.9 palavras
- Mediana do tamanho da descrição: 2.0 palavras
- Observação: a maioria dos PRs possui descrições curtas e não houve diferença clara na taxa de merge associada ao tamanho da descrição

RQ 04. Qual a relação entre as interações nos PRs e o feedback final das revisões?
Métrica: comentários e participantes por PR vs. taxa de merge

- Média de comentários por PR: 2.9
- Mediana de comentários por PR: 2.2
- Média de participantes por PR: 1.9

- Mediana de participantes por PR: 1.6

- Observação: PRs com mais interações tendem a apresentar taxas de merge mais elevadas

RQ 05. Qual a relação entre o tamanho dos PRs e o número de revisões realizadas?
Métrica: tamanho total do PR vs. revisões médias

- Média de revisões por PR: 2.4

- Mediana de revisões por PR: 2.3

- Observação: PRs maiores (mais de 500 linhas modificadas) tendem a passar por mais revisões

RQ 06. Qual a relação entre o tempo de análise dos PRs e o número de revisões realizadas?
Métrica: tempo médio de análise vs. número de revisões

- Repositórios com maior tempo de análise (> 5 dias) costumam ter mais de 3 revisões em média

- Exemplo: immich (683h de análise, 2.8 revisões); strapi (217h, 4.8 revisões)

RQ 07. Qual a relação entre a descrição dos PRs e o número de revisões realizadas?
Métrica: tamanho da descrição vs. revisões

- A maioria dos PRs possui descrições curtas (menos de 50 palavras)

- Não houve relação clara entre descrições longas e menos revisões

RQ 08. Qual a relação entre as interações nos PRs e o número de revisões realizadas?
Métrica: número de comentários e participantes vs. revisões

- PRs com mais de 5 comentários geralmente têm mais de 3 revisões

- Exemplo: gpt_academic (5.4 comentários, 5.4 revisões); sway (4.5 comentários, 4.5 revisões)

DISCUSSÃO (HIPÓTESES x VALORES OBTIDOS)

H1: Pull Requests com menos de 300 linhas modificadas tendem a ser aceitos com uma taxa superior a 80%.

- Resultado: Confirmado parcialmente. A maioria dos PRs com menos de 300 linhas (somando linhas adicionadas e removidas) teve taxas de merge superiores a 70%, com alguns chegando a 100%. No entanto, há exceções, como projetos com PRs pequenos e taxas abaixo de 60%, sugerindo que o tamanho reduzido não garante aprovação automática.

H2: Pull Requests analisados em menos de 2 dias têm maior chance de aceitação, com taxa superior a 75%.

- Resultado: Confirmado. Repositórios com tempo médio de análise inferior a 48 horas, como system-design-primer (24.14h) e 30-seconds-of-code (8.18h), apresentaram taxa de merge de 100% e 88.9% respectivamente. Em contraste, projetos com tempos muito longos, como build-your-own-x (5688h), tiveram taxas mais baixas (55.6%).

H3: Pull Requests com descrições contendo mais de 100 palavras são mais aceitos, com taxa de aprovação superior a 85%.

- Resultado: Não confirmado. A maioria dos repositórios possui descrições médias entre 0 e 20 palavras. PRs com descrições longas são raros e não apresentaram vantagem clara quanto à taxa de aceitação. Isso sugere que o conteúdo e clareza da descrição podem ser mais relevantes do que o tamanho em si.

H4: Pull Requests com mais de 5 interações (comentários e revisões) apresentam taxa de aceitação superior a 80%.

- Resultado: Confirmado. Repositórios com média de interações acima de 5, como immich (12.0 interações, 75% merge), cpython (10.1 interações, 87.5%) e TheAlgorithm/Java (8.9 interações, 85.7%), mostraram taxas altas de aceitação. Interações parecem estar positivamente associadas a decisões de merge.

H5: Pull Requests com mais de 500 linhas modificadas tendem a passar por mais de 2 ciclos de revisão.

- Resultado: Confirmado. PRs grandes como os dos repositórios gpt_academic (858.7 linhas, 5.4 revisões), supabase (1605.4 linhas, 2.4 revisões), e immich (1215.8 linhas,

2.8 revisões) tiveram média de revisões acima de 2. Repositórios com PRs pequenos geralmente apresentaram 1 ou 2 revisões.

H6: Pull Requests cujo tempo de análise ultrapassa 5 dias costumam passar por mais de 3 ciclos de revisão.

- Resultado: Confirmado. Projetos como gpt4all (152h, 3.0 revisões), strapi (217h, 4.8 revisões), e immich (683h, 2.8 revisões) demonstram uma correlação entre análise prolongada e maior quantidade de revisões.

H7: Pull Requests com descrições acima de 100 palavras geralmente passam por menos de 2 ciclos de revisão.

- Resultado: Não confirmado. A maioria das descrições é curta. Mesmo PRs com muitas revisões costumam ter descrições abaixo de 50 palavras, indicando que a descrição longa não implica necessariamente menor número de revisões.

H8: Pull Requests com mais de 10 comentários tendem a passar por mais de 3 ciclos de revisão.

- Resultado: Confirmado. Projetos com maior número de comentários, como immich (12 comentários, 2.8 revisões), gpt_academic (5.4 comentários, 5.4 revisões) e sway (4.5 comentários, 4.5 revisões), geralmente passaram por mais de 3 revisões.