# Step Report: D-1v2 — Enriched Data Analysis

**Projeto:** Clawkeeper **Dia/Etapa:** D-1 revisado — Dados enriquecidos (2000 PRs + additions/deletions)
**Timestamp:** 2026-02-17T10:25:00-03:00 **Duração:** ~20 min

---

## O que fiz

1. Ampliei histórico: 500 → **2000 PRs** (20 páginas API, unique by number)
2. Enriqueci 200 PRs (100 merged + 100 closed) com additions/deletions/changedFiles/reviews
3. Analisei merge rates por: size, component, channel, contributor type, author
4. Busquei score numérico do Greptile (1-5) — não encontrado nos dados atuais
5. Instalei e autentiquei `gh` CLI (tiuitobot)

## O que encontrei

### Merge Rate Real (N=2000): 25.5%

- **Amostra de 500 dava 39%** — viés de recência severo. Os dados mais antigos têm merge rate menor.
- 509 merged, 1491 closed sem merge.

### Size → Merge Rate (confirmado, N maior)

| Size | Merge Rate | N | Delta vs N=500 |
|---|---|---|---|
| XS | 30% | 457 | -3pp |
| S | **46%** | 373 | -4pp |
| M | **43%** | 203 | -12pp |
| L | **47%** | 106 | -8pp |
| XL | **13%** | 162 | -3pp |

Pattern mantém: sweet spot S-L, XL quase auto-rejeição. M caiu mais que esperado (de 55% pra 43%).

### Contributor Type → Merge Rate (CORREÇÃO IMPORTANTE)

| Tipo | Merge Rate | N |
|---|---|---|
| maintainer | **88%** | 110 |
| trusted-contributor | **60%** | 96 |
| experienced-contributor | **50%** | 99 |
| sem label de contributor | ~15-20% | ~1700 |

⚠ **Com N=500, trusted-contributor dava 34% — FALSO.** Com N=2000, dá 60%. A amostra menor estava enviesada. Hierarquia é clara e forte: maintainer > trusted > experienced > unknown. Preditor poderoso pro logit.

### Component → Merge Rate

| Component | Merge Rate | N |
|---|---|---|
| docs | **24%** | 295 |
| agents | 20% | 770 |
| gateway | 20% | 390 |
| cli | 16% | 218 |
| commands | 16% | 280 |
| app: web-ui | **12%** | 228 |

web-ui é a área mais difícil de entrar. docs a mais fácil (faz sentido — menor risco).

### Channel → Merge Rate

| Channel | Merge Rate | N |
|---|---|---|
| telegram | **18%** | 259 |
| discord | 16% | 266 |
| slack | 12% | 185 |
| whatsapp-web | 9% | 175 |
| mattermost | 8% | 89 |
| signal | **6%** | 138 |
| imessage | 6% | 124 |
| matrix | **5%** | 119 |

Telegram/Discord são os canais "core" — mais PRs e mais aceitos. Signal/Matrix/iMessage quase não passam.

### Top Authors (merged)

| Author | Merges |
|---|---|
| **steipete** | **22** |
| 0xRaini | 20 |
| mcaxtr | 20 |
| arosstale | 18 |
| mbelinky | 18 |

**steipete = Peter (criador/maintainer).** Os top contributors externos têm 18-20 merges.

### Additions/Deletions (amostra enriquecida, N=200)

| Métrica | Valor |
|---|---|
| Mean additions | 5,964 (distorcido por outliers) |
| **Median additions** | **89** |
| P90 additions | 1,617 |
| Max additions | 752,543 |
| Mean deletions | 697 |
| Mean changed files | 103 |

**Mediana de 89 linhas adicionadas.** P90 em 1617 — confirma que 1000 LoC rule pega ~15-20% dos PRs.

### Greptile

- Presente em **muitas PRs** (324 reviews na amostra de 200)
- Formato: `X files reviewed, Y comments` — **sem score numérico 1-5**
- Bruno vai verificar e mandar exemplo do score
- Variáveis extraíveis: `greptile_files_reviewed`, `greptile_comment_count`

## O que pensei

1. **O viés de recência era perigoso.** Se tivéssemos estimado o logit com N=500, os coeficientes estariam errados — especialmente trusted-contributor (34% → 60% é uma inversão de sinal no modelo). A decisão de ampliar pra 2000 foi correta.

2. **Contributor type é o preditor mais forte**, mais que size. maintainer=88% é quase determinístico. Isso sugere que o logit vai ter um β_maintainer enorme. Implicação: o Clawkeeper precisa dizer "este PR foi submetido por um maintainer — merge quase certo" de forma transparente.

3. **Channel PRs têm merge rates baixos em geral** (5-18%). Hypothesis: PRs de channel são frequentemente de contribuidores one-shot que não conhecem o codebase profundamente. Se essa hipótese se confirmar, `is_channel_pr` × `author_merge_count` pode ter interação interessante.

4. **Mediana de 89 additions** = a maioria dos PRs são pequenos. O quality gate de 1000 LoC provavelmente rejeita top ~10-15%. Isso é feature, não bug — PRs enormes têm 13% merge rate.

5. **steipete como top contributor** — ele mergeia seus próprios PRs (maintainer privilege). Isso pode inflar o β_maintainer. Considerar excluir self-merges do treino, ou criar variável separada.

## Decisões tomadas

- Ampliei de 500 → 2000 PRs (validou que N=500 era insuficiente)
- Enriqueci amostra de 200 (suficiente pra estatísticas descritivas, logit usa as 2000)
- Mantive busca por Greptile score — se Bruno confirmar, faço scrape direcionado
- Plano DIFF atualizado em `plans/clawkeeper/PLAN-v4-DIFF.md`

## Riscos / Alertas

- **Self-merges do maintainer** podem enviesar o logit. Precisa de variável `is_self_merge` ou exclusão.
- **Additions/deletions só pra 200 PRs** — pro logit completo, preciso enriquecer as 2000. São ~~2000 API calls (~~20-30 min com rate limiting).
- **Greptile score 1-5** não encontrado — se existir, é feature nova ou em PRs específicos. Aguardando Bruno.

## Próximo passo

1. Enriquecer 2000 PRs com additions/deletions (batch com gh CLI)
2. Começar pipeline Step 1-3 (ingest + signal extraction + categorization)
3. Estimar logit baseline com variáveis disponíveis

---

*Report gerado em sessão principal. Próximos via spawn Sonnet.*