SecHack365

GitHub 上で怪しい行動を早期発見するスコアリング

研究駆動コース 4R 臼井 悠人

背景

- GitHub 上の協力は、**不特定多数 × 長期間 × 非同期**で進むため相手の信頼性を前提にしないと成り立たない
- しかし現実には、参加者が多すぎて全員を**人手で精査するのは不可能**
- スパムや侵入などの**怪しい行動を早期に兆候を拾い**、精査する対象を絞りたい

→ 信頼度を自動的に評価する仕組み (= スコアリング) が必要

既存手法の限界

- コードレビュー・静的解析
  - 悪意のあるコード自体は、最終段階まで現れない
  - XZUtils への攻撃といった**長期的信頼獲得型の攻撃**には無力
- 身元確認・人的信頼
  - OSS の開放性と根本的に相反
  - スケールしない／属人化する
- ルールベース検知
  - 攻撃手法が変わると簡単に破られる

→「内容」や「属性」ではなく、**「行動」**を見る必要

ユースケース

- XZ Utils への攻撃※のような、**長期的信頼獲得型の攻撃**では、攻撃者は正規貢献者として**徐々に浸透**
- 本手法は公開行動ログから 協調性・返報性・継続性 を算出し、行動の不自然さを根拠付きで可視化
- コードが**混入する前段階**で危険なユーザーを絞り込む

→ 全件精査が不可能なインターネット上の協力でも、早期の兆候検出として運用可能

※CVE-2024-3094

提案手法

- GitHub 上の観測可能な行動ログから信頼度をユーザー間のグラフを生成する（行動のコストに応じて重み付け）
- 信頼を 1 指標で判定せず、行動を **協調性・返報性・継続性**（＋役割の自然さ）に分解して評価する
- 3 **指標を統合して**「総合信頼スコア」を算出し（1 つでも不自然だと下がる設計）、根拠付きで提示する

→ 管理者は「何が不自然か」を理解し、警戒すべきユーザーを絞り、攻撃を防ぐことが可能

仕組み

1.観測：GitHub の公開行動ログ（PR/Issue/Review/ コメント等）を収集

2.行動指標化

A. 継続性：活動が急激に変化しないか、一定の頻度で続くか

B. 協調性：関わる相手・関わり方が多様か（特定クラスタへの偏りが少ないか）

C. 返報性：やり取りが往復しているか（相互に反応・承認が発生しているか）

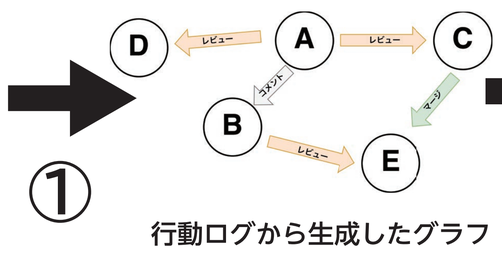
3.統合：複数指標を統合して「総合信頼スコア」を算出（1 つでも不自然だと下がる設計）

4.提示：スコアと根拠（どの指標が低いか）を表示し、要レビュー対象を早期に抽出

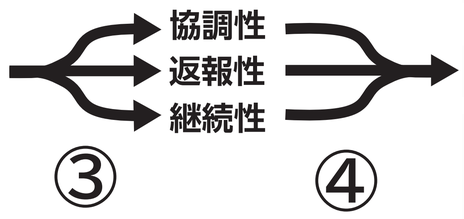
デモ画面

created_at	event_type	source_user	target_user	action
2021-09-30 13:48:39+00:00	PushEvent	JiaT75	None	None
2021-10-04 14:07:28+00:00	PullRequestEvent	JiaT75	JiaT75	opened
2021-10-07 14:43:20+00:00	PushEvent	JiaT75	None	None
2021-10-12 14:08:15+00:00	PushEvent	JiaT75	None	None
2021-11-02 14:55:27+00:00	PullRequestEvent	JiaT75	JiaT75	opened
...	...	...	...	...
2024-04-06 10:49:28+00:00	IssueCommentEvent	ninjamar	JiaT75	created
2024-04-06 16:41:16+00:00	IssueCommentEvent	xealits	JiaT75	created
2024-04-07 07:07:12+00:00	IssueCommentEvent	molhamalnasr	JiaT75	created
2024-04-16 00:58:41+00:00	IssuesEvent	MaskRay	JiaT75	closed
2024-04-16 16:31:04+00:00	IssueCommentEvent	furiousdroid	JiaT75	created

GitHub 上の行動データ



行動ログから生成したグラフ



評価

検知率と汎用性を評価する

実験設定

- 既存のリポジトリにおける活動履歴に人工的にボットを注入
  - 攻撃シナリオ 4 種類
  - 各シナリオ 3 個ずつ

検知率（注入ボットを上位に浮上させられるか）

- 総合信信頼スコアが低いユーザ上位 20 件にいくつボットが入るか

→ 注入ボット 12 体のうち 11 体（92%）が、20 件のうちに浮上

汎用性（特定パターンに依存せず効くか）

- 信頼スコアが低いユーザ上位 20 件に 11 個のボットが入った

→ 4 種類すべてのシナリオで一貫して上位に浮上

注意（誤検知の可能性）

- 上位に現れた非注入アカウントの一部は一般的な Bot だった（例:CLAassistant）
- → 人工ボット以外も拾い得るため、人の判断支援として運用するが必要

	period	user	final_score
20178	2024-04-15 00:00:00	CLAassistant	0.008531
20432	2024-04-15 00:00:00	bot_random_pull_request_01	0.011486
20431	2024-04-15 00:00:00	bot_random_pull_request_00	0.011719
20433	2024-04-15 00:00:00	bot_random_pull_request_02	0.011752
20434	2024-04-15 00:00:00	bot_selfish_pull_request_00	0.013806
20436	2024-04-15 00:00:00	bot_selfish_pull_request_02	0.013918
20435	2024-04-15 00:00:00	bot_selfish_pull_request_01	0.014751
20501	2024-04-15 00:00:00	ddevault	0.021526
20353	2024-04-15 00:00:00	XtremeOwnageDotCom	0.025811
20791	2024-04-15 00:00:00	mgravell	0.025957

今後の展望

- より現実に即したデータを利用した検証
  - 実際にボット以外の悪意のあるユーザーを見分けられるか
- 他のサービス・分野への応用の検討
  - Wikipedia への利用
  - etc...
- 実運用を見据えた UI / ワークフローの検討
  - 誤検知の防止などを行うため
- 攻撃者のこのシステムへの対策についての耐性
  - いたちごっこにならないか

NICT National Institute of Information and Communications Technology