xGitGuard: ML-based Secret Scanner for GitHub

Bahman Rashidi
Sr. Security Architect
Comcast Cable
Overview

GOAL
THREAT
xGITGUARD
VALIDATION

PUBLIC & ENTERPRISE GITHUB
DATA BREACH
AI ACCURACY SCALABILITY
CONTEXT SPLUNK
OTHER SOLUTIONS

Existing Approaches

- **TruffleHog**
  - Detect high-entropy text

- **Gitguardian**
  - Detects secrets using Regex classifiers

- **EarlyBird**
  - Detects specific patterns in files

- **NightWatch**
  - Entropy-based detection

Limitations with existing works:

- User decides the repos
- Regex-based detection
- Relying on entropy
- Scalability
What Secrets?

**CREDENTIALS**
- Username & passwords
- Server credentials
- Account credentials
- etc.

**TOKENS / KEYS**
- Service API tokens (AWS, Azure, etc)
- Encryption keys
- etc.
XGITGUARD WORKFLOW

Overall Workflow

- Search GitHub at Scale
- Detect & Extract Secrets
- Validate Secrets
- Filter Results
- Developer Identification
- Submit to Dashboard
Credential Detection

Search Keywords
- Primary Keyword (PKEY) abc.xyztest.test
- Secondary Keyword (SKEY) Password Secret
- Extension (EXT) JSON Config

Query Engine
- Query Generator PKEY + SKEY + EXT
  Example: abc.xyztest.test + password + JSON
- Python processes P1 (PKEY1+SKEY1+EXT1)
  P2 (PKEY1+SKEY1+EXT2)
  P3 (PKEY1+SKEY1+EXT3)
  Pn (PKEYm+SKEYk+EXTl)

URL Hash Table Object
- Candidate Files
- Secret Search Pipeline
- Hash Lookup
- Hash (URL+Keyword)

Text Processing
- Coarse Tokenizing
- De-noising
- Fine Tokenizing
- Stop Word Reduction
- English Programming
- Secondary Key Search
- Secret Extraction

Validation Phase
- Secrets
- Entropy Test
- Dic. Check
- Dashboard
**Token/Key Detection**

DETECTION ENGINES

- **Search Keywords**
  - Primary Keyword (PKEY): abc.xyztest.test
  - Extension (EXT): JSON, Config

- **Query Engineer**
  - Query Generator: PKEY + EXT
  - Example: abc.xyztest.test + JSON
  - Python threads: P1 (PKEY1+EXT1), P2 (PKEY1+EXT2), P3 (PKEY1+EXT3), Pn (PKEYm+EXTn)

- **URL Hash Table Object**
  - Hash Lookup

- **Text Processing**
  - Coarse Tokenizing
  - De-noising
  - Stop Word Reduction
  - English, Programming
  - REGEX

- **Validation Phase**
  - Dashboard

© 2021 SCTE®, CableLabs & NCTA. All rights reserved. | expo.scte.org
Validation Model

Data

Historical detections

Prepare

• Cleaning
• Labeling

Build

• Feature engineering
• Training & testing

Deploy

Validation
Thank You!

Bahman Rashidi
Sr. Security Architect
Comcast Cable
Email: bahman_rashidi@comcast.com