

# AI-Assisted Insider Threat Detection

## Correlating Gemini Activity with Drive Exfiltration in Google Workspace

haasonsaas

BSides

October 16, 2025

# Overview

- 1 The Problem
- 2 The Detection
- 3 Implementation
- 4 Results & Insights
- 5 Operational Considerations
- 6 Conclusion

# The Insider Threat Landscape

- Traditional DLP focuses on **content inspection**
- Misses **behavioral patterns** that indicate intent
- New attack surface: **LLM-assisted reconnaissance**
- Insiders now use AI to rapidly understand sensitive documents

## The New TTP

Use Gemini to analyze files → Immediately exfiltrate them

# Why This Matters

## Traditional Exfil:

- 1 Manually read documents
- 2 Identify sensitive content
- 3 Extract/share

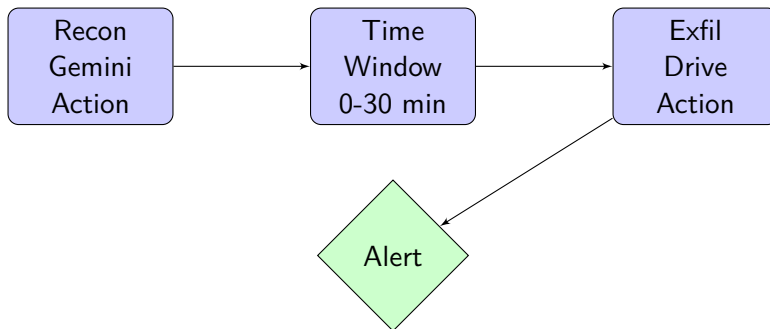
*Time-consuming, leaves traces*

## LLM-Assisted Exfil:

- 1 Ask Gemini "summarize this"
- 2 Instantly understand value
- 3 Immediately exfiltrate

*Fast, efficient, higher-value targets*

# Detection Logic



**Correlation Key:** Actor + Temporal Proximity

# Recon Signals (Gemini Events)

**Data Source:** Admin SDK Reports API

**Application:** `gemini_in_workspace_apps`

**Event:** `feature_utilization`

## High-Signal Actions:

- `ask_about_this_file` - Direct file query
- `summarize_file` - File summarization
- `analyze_documents` - Multi-file analysis
- `catch_me_up` - Bulk triage
- `report_unspecified_files` - Report generation

**Available Since:** 2025-06-20 (180-day retention)

# Exfil Signals (Drive Events)

**Data Source:** Admin SDK Reports API

**Application:** drive

## High-Risk Events:

- `change_visibility` - Made public/external
- `change_acl` - External principal added
- `export` - Export to PDF/DOCX/CSV
- `download` - File download
- `copy` - File duplication
- `add_to_folder` - Move to external folder

**Key Parameters:** `doc_id`, `visibility`, `new_value`, `old_value`

# Detection Algorithm

```
1 for exfil_event in drive_events:
2     for recon_event in gemini_events:
3         if exfil_event.actor == recon_event.actor:
4             delta = exfil_event.time - recon_event.time
5
6             if 0 <= delta <= 30_minutes:
7                 severity = calculate_severity(
8                     delta,
9                     exfil_event.type,
10                    exfil_event.visibility
11                )
12
13                emit_finding(severity, recon_event, exfil_event)
```



# Severity Rubric

Severity	Criteria	Response
High	External share/export $\leq$ 10 min	Page on-call
Medium	External share/export 10-30 min	Next-day review
Low	Any permission change within 30 min	Log for analysis

## Severity Overrides:

- Actor in high-risk OU (Exec, Finance, R&D)  $\rightarrow$  +1 level
- File labeled confidential/restricted  $\rightarrow$  +1 level

- ➊ **Authentication:** Service account with domain-wide delegation
- ➋ **Data Collection:** Fetch Gemini + Drive events via Admin SDK
- ➌ **Correlation Engine:** Temporal matching by actor
- ➍ **Scoring:** Apply severity rules and suppressions
- ➎ **Output:** JSON findings to SIEM/alerting

## Deployment Options:

- Cron job (every 10 minutes)
- Systemd timer
- Cloud Function / Lambda

# Example Finding (Enhanced)

```
1 {
2   "severity": "high",
3   "actor": "john.doe@company.com",
4   "exfil_event": "change_visibility",
5   "doc_title": "Q4 Financial Projections.xlsx",
6   "delta_minutes": 5.55,
7   "reason": "External share within 10min; High recon score (12.5)",
8   "recon_score": 12.5,
9   "file_context": {
10     "sensitivity": "high",
11     "labels": ["confidential", "finance"],
12     "owner": "cfo@company.com"
13   },
14   "intent_analysis": {
15     "intent": "malicious",
16     "confidence": 0.85,
17     "reasons": ["Unknown destination domain",
18               "Sharing someone else's file",
19               "Off-hours activity"],
20     "destination_domain": "competitor.com"
21   }
22 }
```

## False Positive Reduction:

- Allowlist trusted external domains (partners)
- Suppress security/IT OUs investigating files
- Exclude service accounts
- Adjust time windows based on your environment

## Calibration Process:

- 1 Week 1: Observation mode (no alerts)
- 2 Week 2: Tune suppressions, enable high severity
- 3 Week 3-4: Refinement
- 4 Month 2+: Ongoing review

# Why This Works

## Behavioral Sequence Detection:

- **Intent:** Gemini logs reveal *what* the user wanted to understand
- **Action:** Drive logs reveal *what* they did with it
- **Correlation:** Temporal proximity reveals insider TTP

## No Content Inspection Required:

- Privacy-preserving detection
- No prompt/response content visible
- Metadata-only telemetry

## High-Fidelity Signal:

- Precise timestamps
- Actor email
- File IDs
- Action types

## Google-Recommended:

- Official audit logs
- Designed for security telemetry
- 180-day retention
- No extra cost

# Common Patterns Detected

- 1 **Rapid Reconnaissance:** User asks Gemini about 10+ files, then downloads 3 with highest value
- 2 **Pre-Resignation Exfil:** Employee uses Gemini to triage entire project folder, then exports key documents
- 3 **Competitive Intelligence:** Sales rep summarizes customer contracts, immediately shares with personal email
- 4 **IP Theft:** Engineer asks Gemini to analyze codebase documentation, then copies to external Git repo

# Deployment Checklist

- ❶ Create service account with domain-wide delegation
- ❷ Grant scopes:
  - `admin.reports.audit.readonly`
  - `drive.readonly` (for file context)
- ❸ Configure Redis for stateful recon tracking (optional)
- ❹ Configure partner domains & suppressions
- ❺ Test with known benign patterns
- ❻ Start with observation mode (no alerts)
- ❼ Gradually enable alerting by severity
- ❽ Integrate with SIEM/SOAR

## Security Best Practices:

- Store service account key in secrets manager
- Rotate keys every 90 days
- Monitor the monitor (alert if detector stops)



## Detection Quality:

- True positive rate
- False positive rate
- Time to detection
- Time to response

## Operational:

- Alert volume (findings/day)
- API success rate
- Coverage (Gemini users / total users)
- Alert fatigue score

# Advanced Features Implemented

## Multi-Stage Attack Detection:

- Cumulative recon scoring with 48hr decay half-life
- Detects delayed exfil (Day 1: recon, Day 3: exfil)
- Redis-backed stateful tracking (with in-memory fallback)

## File Context Enrichment:

- File sensitivity classification (labels, ownership, sharing history)
- Automatic severity elevation for confidential files
- Workaround for missing doc\_id in Gemini events

## Intent Classification:

- Destination domain reputation (trusted/partner/unknown)
- User behavioral baselines (typical sharing patterns)
- File ownership checks + off-hours detection
- Auto-suppression for legitimate workflows

# Limitations & Future Work

## Current Limitations:

- Gemini events only since 2025-06-20
- 180-day retention window
- Gemini API doesn't expose doc\_id (requires heuristics)
- Requires Google Workspace Enterprise

## Future Extensions:

- Automated response (revoke links, notify owner)
- Integration with HR data (resignations, PIPs)
- Cross-correlation with physical security (badge logs)
- Bulk recon + mass exfil detection

# Key Takeaways

- 1 **New Attack Surface:** LLMs enable rapid, efficient insider reconnaissance
- 2 **Behavioral Detection:** Focus on sequences, not content
- 3 **High-Signal:** Temporal correlation of recon + exfil is highly indicative
- 4 **Privacy-Preserving:** No content inspection required
- 5 **Actionable:** Deploy today with Google Admin SDK

## The Bottom Line

AI-assisted insider threats require AI-aware detection. This technique provides high-fidelity telemetry without compromising user privacy.

## GitHub Repository:

<https://github.com/haasonsaas/gemini-exfil-detector>

## Google Documentation:

- Admin SDK Reports API
- Gemini in Workspace Apps Events
- Drive Audit Events

## Contact:

GitHub: @haasonsaas

# Questions?