AI-Assisted Insider Threat Detection

Correlating Gemini Activity with Drive Exfiltration in Google Workspace

haasonsaas

BSides

October 16, 2025

Overview

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 $\rightarrow$ Immediately exfiltrate them

Why This Matters

Traditional Exfil:

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

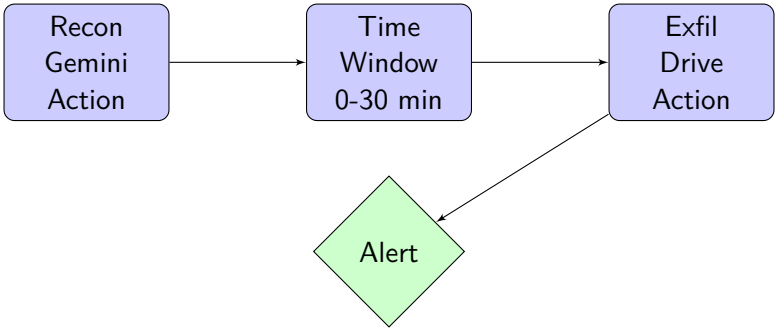
Time-consuming, leaves traces

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

```
for exfil_event in drive_events:
    for recon_event in gemini_events:
        if exfil_event.actor == recon_event.actor:
            delta = exfil_event.time - recon_event.time

            if 0 <= delta <= 30_minutes:
                severity = calculate_severity(
                    delta,
                    exfil_event.type,
                    exfil_event.visibility
                )

                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

Architecture

1. **Authentication:** Service account with domain-wide delegation
2. **Data Collection:** Fetch Gemini + Drive events via Admin SDK
3. **Correlation Engine:** Temporal matching by actor
4. **Scoring:** Apply severity rules and suppressions
5. **Output:** JSON findings to SIEM/alerting

Deployment Options:

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

Example Finding (Enhanced)

```
{
  "severity": "high",
  "actor": "john.doe@company.com",
  "exfil_event": "change_visibility",
  "doc_title": "Q4 Financial Projections.xlsx",
  "delta_minutes": 5.55,
  "reason": "External toggle with rapid revert; High recon score (12.5)",
  "reason_codes": ["external_toggle_revert", "high_recon_score"],
  "recon_score": 12.5,
  "burstiness_score": 8.3,
  "ip_address": "203.0.113.42",
  "file_context": {
    "sensitivity": "high",
    "labels": ["confidential", "finance"],
    "owner": "cfo@company.com"
  },
  "intent_analysis": {
    "intent": "malicious",
    "confidence": 0.85,
    "reasons": ["Unknown destination domain",
                "Sharing someone else's file",
                "Off-hours activity",
                "New ASN for actor"],
    "destination_domain": "competitor.com"
  }
}
```

Tuning & Suppression

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

Detection Advantages

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

1. Create service account with domain-wide delegation
2. Grant scopes:
   - `admin.reports.audit.readonly`
   - `drive.readonly` (for file context)
3. Configure Redis for stateful recon tracking (optional)
4. Configure partner domains & suppressions
5. Test with known benign patterns
6. Start with observation mode (no alerts)
7. Gradually enable alerting by severity
8. 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)

Metrics to Track

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

Attacker Evasion Techniques

How would sophisticated insiders try to bypass this?

1. **Time-delayed exfil:** Recon on Day 1, wait 72 hours, exfil on Day 3
2. **Quick-toggle:** Share externally download immediately revert to private
3. **Shortcut evasion:** Create shortcut to external shared drive instead of direct share
4. **Ownership transfer:** Transfer ownership to external account
5. **Slow-drip recon:** Spread Gemini queries over days to stay under threshold

Detection Challenge

Traditional 30-minute correlation windows miss delayed and obfuscated patterns.

Advanced Detection Features

Multi-Stage Attack Detection:

- Cumulative recon scoring with 48hr decay half-life
- Detects delayed exfil (Day 1: recon, Day 3: exfil)
- Burstiness analysis (rapid-fire queries = high risk)
- Redis-backed stateful tracking (persistent across runs)

Evasion Hardening:

- **Revert-to-clean detection:** External share + rapid revert within 10min HIGH
- **Expanded exfil signals:** Shortcuts, ownership transfer, publish_to_web, move
- **IP/Geo anomaly:** New IP/ASN during exfil phase
- **Structured reason codes:** external_toggle_revert, burst_recon, canary_doc_access

Honeypot/Canary Support:

- Tag decoy high-value docs in config
- **Any** Gemini recon on canary immediate HIGH severity alert
- No content inspection neededpure metadata signal

File Context & Intent Classification

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

Live Demo Scenarios

Red Team vs. Blue Team:

1. **Naive Attack:** Recon → immediate exfil (CAUGHT: immediate correlation)
2. **Quick-Toggle Evasion:** External share → download → revert in 5 min (CAUGHT: revert-to-clean pattern)
3. **Delayed Exfil:** Day 1 recon → Day 3 export (CAUGHT: cumulative recon score)
4. **Shortcut Evasion:** Create shortcut to external shared drive (CAUGHT: expanded exfil signals)
5. **Canary Trap:** Any Gemini query on honeypot doc (CAUGHT: instant HIGH alert)

Result: 5/5 evasion attempts detected with enhanced logic

Key Takeaways

1. **New Attack Surface:** LLMs enable rapid, efficient insider reconnaissance
2. **Behavioral Detection:** Focus on sequences, not content
3. **Evasion-Aware:** Detect revert-to-clean, delayed exfil, obfuscation tactics
4. **Privacy-Preserving:** No content inspectionpure metadata telemetry
5. **Canary-Enhanced:** Honeypot docs provide ground-truth high-signal alerts
6. **Actionable:** Deploy today with Google Admin SDK + open-source code

The Bottom Line

AI-assisted insider threats require AI-aware detection with evasion hardening. This technique provides high-fidelity behavioral telemetry that catches sophisticated attackers.

Resources

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?