



# Security Frontiers

2025

# DARK WATCH

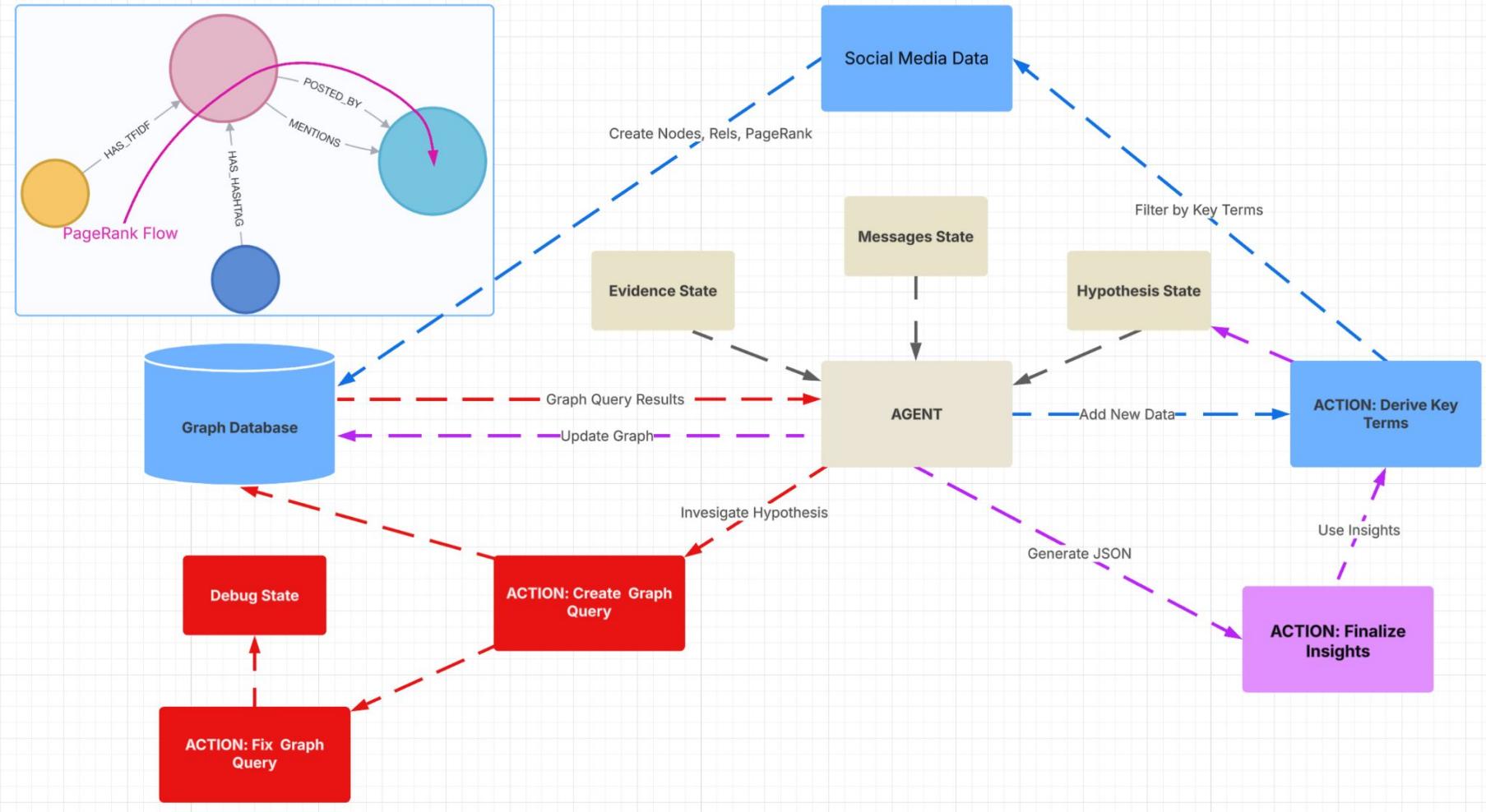
**Simulation Objective:** Hunting  
Social Media Accounts Opposed to  
Trump Policy



We seek top influencers based on PageRank among users who do not already have an existing report. This helps us avoid re-investigating previously covered users while focusing on new, potentially influential individuals.

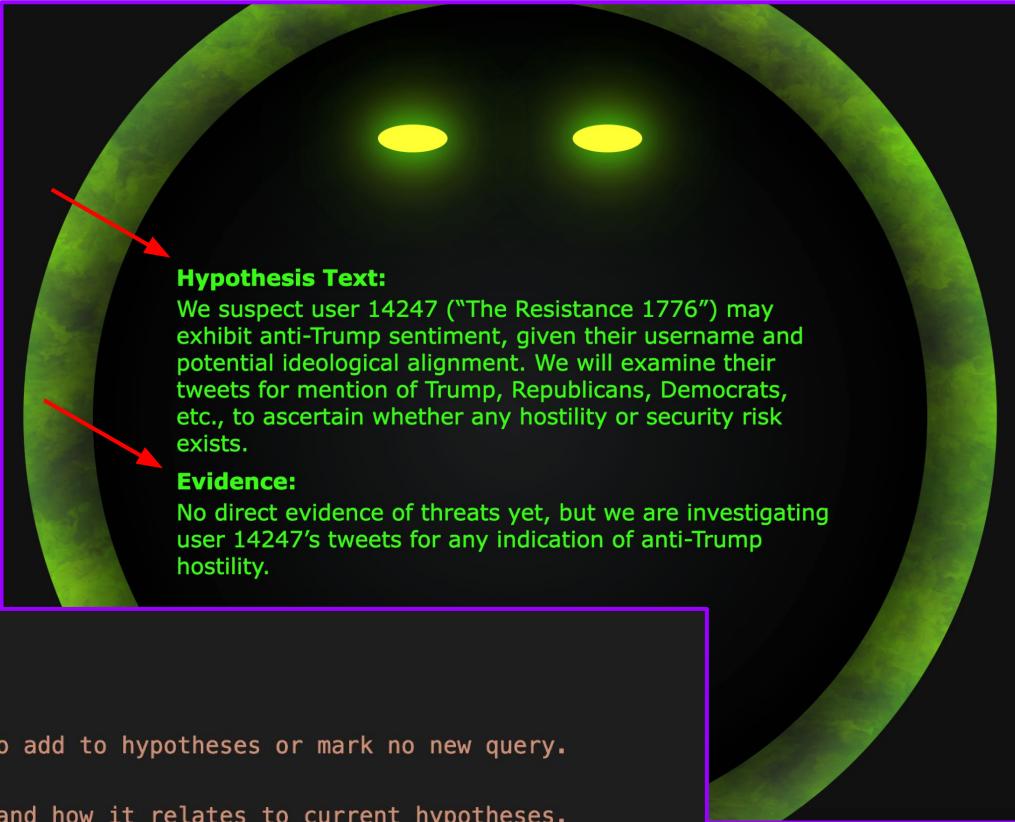
**Graph Query:**

```
MATCH (u:User)
WHERE u.report IS NULL
RETURN id(u) AS user_id, u.user_name AS user_name,
u.user_description AS user_description, u.pagerank AS
pagerank
ORDER BY u.pagerank DESC
LIMIT 10
```



# Adversarial Tension: Hypothesis vs. Evidence

- Agent generates a hypothesis / query
- Agent Analyzes query results to update evidence
- If evidence contradicts hypothesis:
  - it pivots or refines
- If evidence partially confirms:
  - it keeps digging



#### 4. \*\*Structured Reasoning Process (for each iteration):\*\*

- Analyze current hypotheses to decide on next steps.
- Tool Calls: Use the schema above to generate queries to add to hypotheses or mark no new query.
- Observations:
  - Summarize the data retrieved from the graph query and how it relates to current hypotheses.
  - Weigh your hypotheses against the collected evidence.
- Updated Hypotheses (if any): Indicate newly formed or refined hypotheses in the format specified above.
- Note where the data is lacking.
- Conclusion: Decide next steps: continue, or finalize.

# Generate and Repair Queries

```
def fix_cypher_query(self, reasoning, graph_query, query_results, max_attempts=3):
    attempt_count = 0
    debug_messages = []

    while not query_results or "Statement.SyntaxError" in str(query_results):
        print(f"\n\nFixing Query Result...")
        debug_prompt = debug_cypher_query(graph_query, reasoning)
        debug_messages = self.call_model.create_turn("user", debug_prompt, debug_messages)
        response, _, _ = self.call_model.get_openai_chat(debug_messages)
        self.call_model.create_turn("assistant", response, debug_messages)

        graph_query = self.parse_field(response, "Graph_Query")
        print(f"\n\nGraph Query to execute:\n\n{graph_query}")

        query_results = self.run_cypher_query(graph_query)
        print(f"\n\nQuery Result:\n\n{query_results}")

        attempt_count += 1
        if attempt_count >= max_attempts and not query_results:
            query_results = "There are no results for this query. Please try a new idea and query."
            print(query_results)
            return query_results, "NO RESULTS:\n"+graph_query
    return query_results, graph_query
```

```
MATCH (t:Tweet)-[:POSTED_BY]->(u:User)
WHERE id(u) = 14247
AND (
    toLower(t.text) CONTAINS "trump" OR
    toLower(t.text) CONTAINS "republican" OR
    toLower(t.text) CONTAINS "gop" OR
    toLower(t.text) CONTAINS "democrat" OR
    toLower(t.text) CONTAINS "biden" OR
    toLower(t.text) CONTAINS "progressive" OR
    toLower(t.text) CONTAINS "resist"
)
RETURN
id(t) AS tweet_id
```

- Generate and run queries to validate hypotheses
- Repairs queries for correctness

# Pivoting After Not Finding Objective Alignment

## Hypothesis Text:

Based on User 14247's tweets and description, it appears they support Trump rather than oppose him. We should pivot to searching for anti-Trump hashtags (e.g., #resist, #dumptrump, #notmypresident, etc.) to locate other potential users who might exhibit anti-Trump sentiment and have not yet been reported.

## Evidence:

No direct anti-Trump hostility was found in user 14247's content. Evidence suggests they are critical of Democrats and supportive of Trump.

### 3. \*\*Brainstorm New Hypotheses:\*\*

- Suggest new streams of investigation if new evidence emerges.
  - For example, if you detect important new evidence, you might propose a fresh hypothesis.
- Weigh your hypotheses against the collected evidence.
- Store new hypotheses in a proper list.

```
if not next_step or "finalize" in next_step.lower() or turn_count > 21:  
    key_words = []  
  
summary_prompt = findings(hypothesis_list, user_list, evidence_list)  
messages = self.call_model.create_turn("user", summary_prompt, messages)  
final_report, _, _ = self.call_model.get_openai_chat(messages)  
print(f"\n\nFinal Report:\n{final_report}")  
  
json_log = self.json_entry(json_log,  
                           hypothesis_text,  
                           next_step,  
                           graph_query,  
                           evidence,  
                           users,  
                           reasoning,  
                           final_report)  
  
self.data_work.update_persons_of_interest (final_report)
```

- 
- Hashtag Relationships: Hashtags like #Resistance and #resist appeared in tweets, but no verifiable hostility or threats toward Trump were discovered.
  - TFIDF Relationships: Not found to yield any anti-Trump or threatening patterns in the investigated set.

#### Key Insights:

- No newly identified users with explicit anti-Trump hostility.
- #Resistance and #resist may imply opposition to Trump, but no threatening or hostile content was confirmed.

#### \*\*Expected Output Format:\*\*

Executive Summary: Paragraph  
Users Involved: names and IDs and explain how they are related  
Relationship Maps: Explain who is connected to who.  
Key Insights: bullet points  
Key Node IDs Involved: int values in proper python list  
Key Terms: Search terms to find more related data where you feel the data is currently lacking. str values in proper python list.

---

#### \*\*CRITICAL:\*\*

- Ensure You haven't misunderstood the User's sentiment.
- DO NOT add users and tweets who have already been investigated.
- DO NOT add users and tweets to this report who are NOT aligned with your objective:

{obj}

# Finalizing Report: No Findings

```
def derive_keywords(final_report,
                   avoided_terms: list):
    TEMPLATE = f"""
    **Instruction:**
    1. Analyze the `final report` and the `Key Terms` in the final report to derive 3-5 keywords which will be used as search terms.
        - Reason: These search terms will be used for filtering new raw data to extend the graph and continue your investigation.
    2. **Do not use** keywords {avoided_terms} as they have already been used to filter the initial graph.
    3. Avoid keywords like "politics" as this is too general of a search term and will hamper your focused investigation.
    4. Choose keywords which could extend the ideas in the final report by adding new graph data to explore, and potentially adding new insights.
    5. Do not add hashtags or other symbols to your chosen keywords.
    6. Choose **ONLY single word** keywords.

    ---

    **Final Report:**
    {final_report}
    --
```

# Extending the Graph

(Would be paired with web research agent)

Extending graph with filter words: ['fascist', 'impeach', 'orangemanbad', 'biden', 'progressive']

Using: ['fascist', 'impeach', 'orangemanbad', 'biden', 'progressive']

Building graph: 100% |

2000/2000

```
final_report = None
hypothesis_list = []

key_words, key_words_reasoning = run_agent.run_keywords_loop()
hypothesis_list.append(f"**USE THESE** keywords and ideas to start your investigation: {key_words_reasoning}")
run_agent.load_target_data(key_words=key_words, limit_data=2000)
```

#### Reasoning:

1. We hypothesize that tweets with these specific words could indicate negative or hostile sentiment toward Trump.
2. By filtering on these TFIDF nodes, we focus on highly relevant text and reduce noise.
3. The returned set will help us identify possible influential users (via PageRank) who are involved in sharing or discussing anti-Trump content.

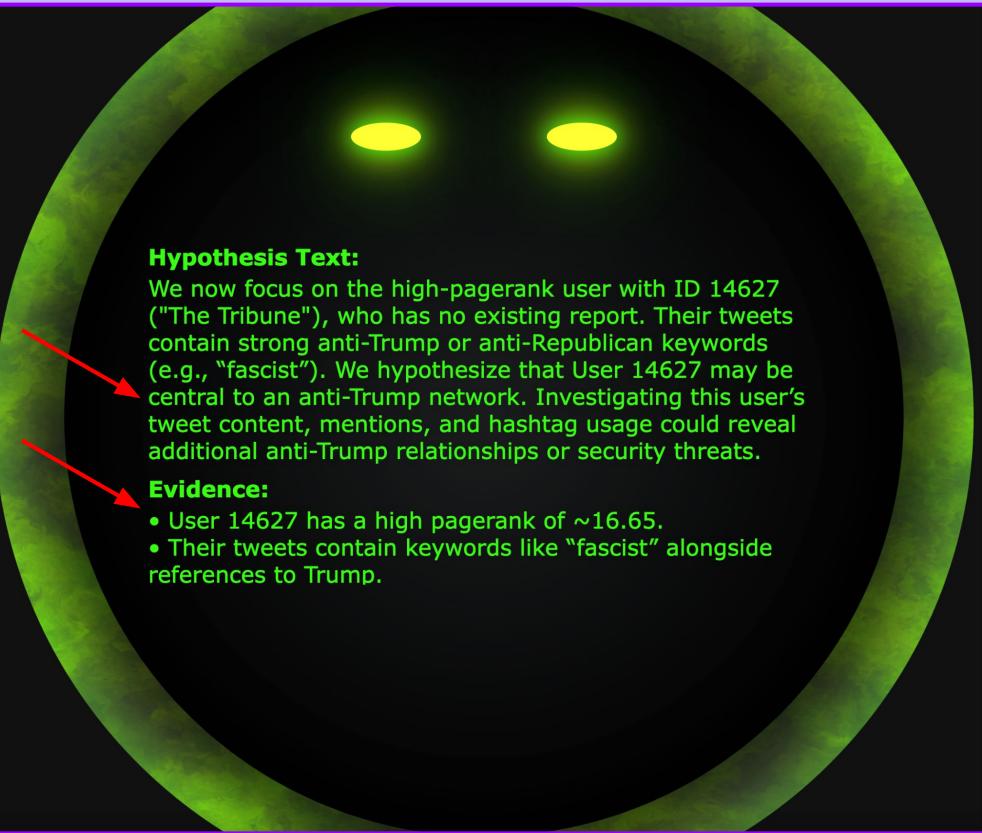
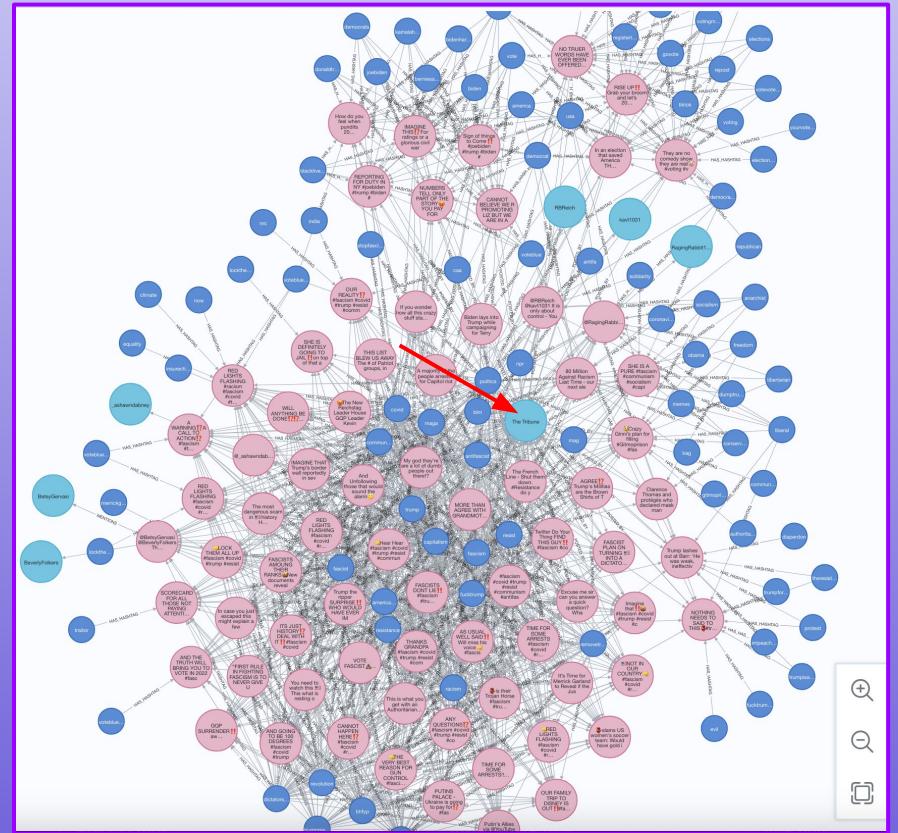
#### Graph Query:

```
MATCH (k:TFIDF)-[:HAS_TFIDF]->(tweet:Tweet)-[:POSTED_BY]->(u:User)
WHERE k.text IN
```

#### Graph Query:

```
MATCH (k:TFIDF)-[:HAS_TFIDF]->(tweet:Tweet)-[:POSTED_BY]->(u:User)
WHERE k.text IN
["fascist", "impeach", "orangemanbad", "biden", "progressive"
RETURN k.text AS keyword, id(k) AS TFIDF_id, tweet.text
AS tweet_text, id(tweet) AS tweet_id,
u.user_name AS user_name, id(u) AS user_id,
k.pagerank AS tfidf_pagerank,
tweet.pagerank AS tweet_pagerank, u.pagerank AS
user_pagerank
ORDER BY k.pagerank DESC, tweet.pagerank DESC,
```

## Realign Investigation Using Residual Ideas & New Data PageRank + TFIDF



# Multi-Perspective Hypothesis Validation

# Finalize Report: Objective Complete

## Key Insights:

- Several users repeatedly label Trump, Republicans, or the GOP as "fascist," "liars," or "authoritarian."
- The cluster around user 14627 ("The Tribune") exhibits strong anti-Trump sentiments, with repeated references to "#f[red]trump" and similar hashtags.
- Additional users 82941 ("CJ7Jeepman") and 82942 ("Hallstrom291") are newly surfaced, displaying anti-Republican critiques.
- Hashtag co-occurrence (e.g., #fascist, #resist, #antifascist) suggests a cohesive sub-network unified around anti-Trump or anti-GOP hostility.

## Relationship Maps:

- User 14627 ("The Tribune") often tweets mentioning threads involving 15854 ("RBReich"), 50383 ("\_ashawndabney"), 10883 ("RagingRabbit101"), and 10757 ("kavt1031").
- User 15854 ("RBReich") is co-mentioned with 82941 ("CJ7Jeepman") and 82942 ("Hallstrom291") in critiques of the GOP.
- Hashtag usage (e.g., #fascist, #f[red]trump) is concentrated among these accounts, reinforcing an active anti-Trump or anti-GOP sentiment network.



# Thank you!