Epstein Files Analysis So Far: A Data Story

Moses Boudourides

1 The Data Sources

This analysis draws from three primary sources of publicly released documents related to Jeffrey Epstein:

1.1 1. Unredacted Flight Logs (Internet Archive)

Source URL: https://ia801606.us.archive.org/30/items/epstein-flight-logs-unredacted_202304/EPSTEIN%20FLIGHT%20LOGS%20UNREDACTED.pdf

Released: April 4, 2023 (uploaded by Democracy Today)

Date range: Flight logs spanning multiple years of aircraft operations

The most substantial dataset comes from the Internet Archive's collection of unredacted flight logs from Epstein's private aircraft. These logs document flights spanning multiple years and contain structured information including dates, aircraft identification numbers, departure and arrival locations, and passenger manifests. The logs provide a detailed record of movement patterns and associations, capturing who traveled with whom, when, and to which destinations. This dataset is particularly valuable because it represents contemporaneous documentation created during the actual events, rather than retrospective accounts.

1.2 2. Flight Logs from U.S. v. Maxwell Case (DOJ)

Source URL: https://www.justice.gov/ag/media/1391276/dl?inline

Released: Made public during the trial of United States v. Ghislaine Maxwell (November-December 2021)

Date range: Additional flight log records from Epstein's aviation activities

The Department of Justice released additional flight logs as evidence in the criminal case United States v. Ghislaine Maxwell. These documents, officially entered into federal court proceedings, provide a complementary perspective to the Internet Archive collection. While there is overlap between these sources, the Maxwell case logs include official government authentication and may contain annotations or redactions reflecting their use in legal proceedings. The logs detail aircraft movements, passenger lists, and travel itineraries associated with Epstein's aviation activities.

1.3 3. Evidence List (DOJ Records via House Oversight Committee)

Source URL: https://www.justice.gov/ag/media/1391271/dl?inline

Released: September 2, 2025 (33,295 pages); additional documents November 12, 2025

Released by: House Committee on Oversight and Government Reform

Source: Records provided by the U.S. Department of Justice in response to congressional subpoena

On September 2, 2025, the House Committee on Oversight and Government Reform released 33,295 pages of Epstein-related records provided by the U.S. Department of Justice. Among these materials is a comprehensive evidence list cataloging items seized or documented during the federal investigation. This list provides insight into the scope and nature of materials collected by law enforcement, including references to documents, digital media, physical evidence, and other items relevant to the investigation. While the evidence list itself is metadata rather than primary source material, it reveals the breadth of the investigative effort and the types of materials authorities deemed significant. Additional documents from the Epstein estate were released by the committee on November 12, 2025.

2 What the Analysis Does

2.1 Text Extraction and Entity Recognition

This computational analysis begins by downloading these PDF documents and extracting their textual content page by page. Using spaCy's Named Entity Recognition (NER) model, it identifies and classifies entities mentioned in the documents into five categories:

- **PERSON**: Individual names (e.g., Jeffrey Epstein, Ghislaine Maxwell, various passengers and associates)
- **ORG**: Organizations, companies, foundations, and institutions
- **GPE**: Geopolitical entities including countries, cities, states, and other administrative regions
- **LOC**: Geographic locations that don't fall into the GPE category, such as specific addresses or landmarks
- **FAC**: Facilities, including buildings, airports, properties, and other physical structures

This categorization is crucial because it allows the analysis to distinguish between different types of actors and locations in the network. A person appearing frequently is different from a location appearing frequently, and treating them separately provides clearer insights.

2.2 Building Category-Specific Networks

For each entity category, the notebook constructs two complementary graph structures:

Hypergraphs map the document-entity relationship. Each document page becomes a hyperedge connecting all entities mentioned on that page. This structure preserves

the original context: if three people and two locations are mentioned on a single flight log page, they're all connected through that document. The hypergraph reveals which entities appear together in the same primary sources, providing a document-centric view of the network.

Knowledge graphs extract entity-to-entity relationships based on co-occurrence patterns. If two people appear together across multiple documents (with a configurable threshold, defaulting to three or more co-occurrences), a relationship edge is created between them. This transforms the document-centric hypergraph into an entity-centric network, making it easier to identify which individuals, organizations, or locations are most frequently associated with each other across the entire corpus.

2.3 Frequency Analysis and Network Centralities

This computational work calculates how frequently each entity appears across all documents (node degree in the hypergraph) and how many connections each entity has to other entities (node degree in the knowledge graph). Entities appearing in many documents or connected to many other entities emerge as central figures in the network. This quantitative approach identifies key actors, locations, and organizations without requiring manual review of every document page.

3 What the Results Reveal

The category-specific approach yields distinct insights:

PERSON networks show who traveled together, who appears in the same documents, and who forms the core of the social network documented in these files.

ORG networks reveal which companies, foundations, or institutions are mentioned in connection with travel, evidence, or other documented activities.

GPE networks map geographic patterns: which cities, countries, or regions appear together in travel logs or other documents, suggesting routes, destinations, or areas of activity.

LOC and FAC networks provide finer geographic detail, identifying specific properties, addresses, airports, or facilities that appear in the documents.

By separating these categories, the analysis avoids conflating fundamentally different types of entities. A person who traveled frequently is analytically distinct from a location that was visited frequently, even if both have high node degrees in a combined graph. The category-specific approach maintains these important distinctions while still capturing the relationships within each domain.

4 Methodological Considerations

This analysis is fundamentally descriptive and exploratory. It identifies patterns in the documents but does not make causal claims or legal judgments. Co-occurrence in documents indicates that entities are mentioned together in the same source material, but the nature of relationships must be interpreted with reference to the original documents.

The NER model may misclassify some entities or miss others entirely, particularly with abbreviated names, nicknames, or ambiguous terms. The analysis is limited to entities that appear in the specific documents downloaded; it does not incorporate information from other sources or investigations.

The co-occurrence threshold (default: 3 or more documents) is a parameter that affects which relationships appear in the knowledge graphs. Lower thresholds include more relationships but may introduce noise; higher thresholds focus on more robust patterns but may miss significant connections that appear in fewer documents.

Finally, this analysis examines publicly released documents that have been filtered through various disclosure processes—legal proceedings, FOIA requests, declassification reviews. The documents available for analysis represent a subset of all materials related to Epstein, shaped by what authorities chose to release and what has been preserved in public archives.