# Animated social map around Donald Trump

21 Feb 2026

## Data source and processing

### *Data source*

This paper uses the **Common Crawl News Dataset (CC-NEWS)** as its text source, restricted to the period **October 2024 to February 2026**.

- Data provider: **Common Crawl Foundation**
- Dataset page: https://data.commoncrawl.org/crawl-data/CC-NEWS/index.html
- Access method: month-by-month CC-NEWS `warc.paths.gz` listings + direct parsing of archived WARC response records

We sampled CNN politics coverage using URL-level filters on `cnn.com`, politics paths, and a multi-keyword Trump/presidential set in URL text (for example: `trump`, `donald-trump`, `biden-trump`, `vance-trump`, `jd-vance`, `kamala-harris`, `joe-biden`, `maga`, `republican`, `democrats`, `gop`, `white-house`, `inauguration`, `president`, `presidential`, `election`, `campaign`, `swing-state`, `electoral-college`, `senate`, `house`, `congress`, plus issue terms such as `epstein`, `tariffs`, `doge`, `ice`, `immigration`, `deportations`, and `border`). We then extracted archived HTML text snippets from selected records.

For the current run, we switch to a looser retrieval mode: no keyword gate, keeping politics + article-like URL constraints, and sampling across the full October 2024 to February 2026 month window. (A strict CNN-only domain gate is temporarily disabled because recent CC-NEWS samples showed near-zero `cnn.com` coverage.)

Current practical collection route is RSS-based headline retrieval (Google News RSS search, month-windowed, source-site filtered), storing headline + short snippet + link, then month-balancing to target sample size.

In practice, this RSS route can under-fill historical months (especially older windows), so achieved sample size may fall below target even with broad query terms.

### *Processing pipeline*

1. Enumerate monthly CC-NEWS WARC file lists for the full date window.
2. Parse WARC response records directly and keep CNN URLs only.
3. Filter URLs for politics-oriented paths and require at least one Trump/presidential keyword in URL text.
4. Exclude obvious non-article paths (for example video/live/gallery URLs).
5. Deduplicate by canonical URL (scheme + host + path; query/fragment removed).
6. Extract title and plain-text excerpt from archived HTML in each matching record.
7. Draw an approximately month-balanced random sample across the full target month range.

Operationally, the crawl checks newer months first but still processes the full configured window before sampling, so month balancing is done across the full October 2024 to February 2026 range. Requests use conservative pacing with retry/backoff and jitter to reduce transient endpoint failures.

### *Link identification for social map*

From sampled text, links are coded using the same social-link logic described in Social Network Analysis — SNA:

- Nodes represent social actors (people, groups, institutions).
- Directed links represent actor-to-actor relationships (not abstract causal mechanisms).
- Relationship labels/tags are attached to each link.
- Sentiment and time labels are added where possible.

The resulting table is converted to app-compatible link columns (`Cause`, `Effect`, `Tags`, plus metadata) for mapping.

### *Animation method*

Animation follows the Causal Map app's **Animate filter** behavior documented in `causal-map-extension/webapp/README.md`:

- Select one link field to define frames (for example month or another time field).
- Optional cumulative mode includes all earlier frames up to the current frame.
- Frame construction respects pipeline order, so filter ordering affects what appears in each animation step.

### *Attribution note*

Common Crawl provides archived web crawl data, not editorial ranking signals. Therefore, terms such as "top" in this workflow refer to **selection rules in processing** (filters/sampling), not publisher-defined headline prominence.

The Trump focus in this workflow is a **URL-proxy filter**, not a full-text entity extraction step. This improves targeting but can still include some non-Trump presidential politics pages and miss some relevant pages whose URLs do not contain these keywords.

## Related

- chapter intro