Description

This dataset consists of all starter packs and all following network data available on Bluesky in January/February 2025. Starter packs can be created by any Bluesky user. They are lists of users and curated feeds with a minimum of 8 and a maximum of 150 entities, curated by the starter pack creator, and users typically name them and provide a description. Other users can use a single click to follow all entities in the starter pack, or they can scroll through a specific starter pack to decide which entities to follow within that starter pack. In our dataset, all DIDs (persistent, unique identifiers) are anonymized via a one-to-one mapping between DIDs and integers in the range [0, num_users); users in the network, as well as users who created starter packs, or appear in starter packs, are identified by a unique, anonymous integer that maps to a single DID. Similarly, starter packs themselves are identified by anonymous unique integer IDs in the range [0, num_starter_packs). All timestamps are rounded to the nearest date.

First, we include the Bluesky following network as it appeared in late January/early February 2025 in three formats: a compressed CSV, a parquet file, and a compressed graph-tool object. This dataset shows all available directed following relationships on Bluesky, along with rounded timestamps indicating when the following event occurred (to the nearest day). We also include a network dataset of starter packs with information on creators and starter pack members; this is intended for users who wish to undertake a computational analysis of the networks created by starter packs or starter packs' influences on networks. Additionally, we include the starter pack network in two file formats that facilitate analysis of the starter pack hypergraph for research on higher-order network interactions.

Filename	**Description**	**Data Structure / Types**
deidentified_starterpack_blob.json.gz	gzipped JSON blob (list) of all starter packs found by scraping all starter packs for all users in the Bluesky PLC directory as of January 2025. Contains pack membership and metadata. All timestamps are rounded to the nearest day to preserve anonymity.	List of dictionaries, each with the following keys and values: **pack-id**: integer; anonymized identifier of the starter pack list **creator-id**: integer; anonymized identifier of the starter pack's creator **date-created:** date; the date on which the starter pack was created **members:** list of members with keys ***id*** (integer; anonymized identifier of the starter pack member) and

		date-created (date; the date on which the member's account was created)
deidentified_starterpack_hif.json.gz	gzipped HIF-compliant JSON file. For details on the HIF standard, see the associated repository and paper.	In the output, **nodes** are anonymized integer identifiers for each user and **edges** are starter packs. Both nodes and edges have an associated **date-created** attribute representing the date the user created an account and the date the starter pack was created, respectively. Each edge also has a **creator-id** attribute listing the anonymized integer ID of the starter pack creator.
deidentified_starterpack_edgelist.csv.gz	A hyperedge list stored as a gzipped CSV file. Each line in file corresponds to a starter pack and each entry in the line is a member of the starter pack.	Each line contains entries separated by commas and corresponds to a starter pack. E.g., "1,2,3,4" indicates that users with anonymized identifiers 1, 2, 3, and 4 belong to the same starter pack. Note that this does not contain any timestamps.
deidentified_follows_edgelist_with_timestamps.csv.zip	Compressed CSV that delineates who follows whom on Bluesky. Includes following timestamps rounded to the nearest day.	An entry of A, B, date_AB indicates that user A (i.e. the integer identifier corresponding to DID A) followed user B (i.e. the integer identifier corresponding to DID B) on the day date_AB
deidentified_follows_graph_with_timestamps.gt.gz	Compressed graph-tool directed graph object with the time-stamped Bluesky following network	The graph has an EdgeProperty called **followed_at** which corresponds to the rounded date the edge was formed (i.e. when the following event took place).
deidentified_follows_edgelist_with_timestamps.parquet	Parquet compressed dataframe.	Has the following columns: **from_id:** the integer ID of the user doing the following **to_id**: the integer ID of the user being followed **followed_at:** the timestamp at which the following event occurred, rounded to the nearest day.

Changelog 2025-08-12

1. Added timestamps (following event timestamps rounded to the nearest date) to following networks; **deidentified_follows_edgelist.zip** is replaced/succeeded by **deidentified_follows_edgelist.csv.zip**, **deidentified_follows_graph.gt.gz**, and **deidentified_follows_edgelist.parquet,** all of which provide the same following network, now with timestamps. The additional formats are meant to make the dataset more usable to a wide variety of practitioners.
2. The starter pack data now includes account creation dates for the users in the starter packs; all other data remains the same. The following files are replaced:
    a. **deidentified_starterpack_hif.json** is replaced by **deidentified_starterpack_hif.json.gz.**
    b. **deidentified_starterpack_edgelist.csv** is replaced by **deidentified_starterpack_edgelist.csv.gz.**
    c. **deidentified_starterpack_blob.json** is replaced by **deidentified_starterpack_blob.json.gz.**
3. This codebook has been updated to reflect the updated data files.