The Dawn of a New Blue(sky) Era: Structural and Content Analysis of the Growing Social Media

Damiano Orlandi

***Abstract*—This paper presents an analysis of the structural and content characteristics of Bluesky. Two analyses are conducted: (1) a Social Network Analysis (SNA) to examine interaction patterns and network structures, and (2) a topic modeling analysis using BERTopic to identify dominant themes in user posts. The study leverages a static subset of one million posts collected via the Bluesky API. The analysis code is available on GitHub.**

***Index Terms*—Bluesky, SNA, Topic Modeling, BERTopic, CSS, X-odus**

I. Introduction

Since its inception in 2023, Bluesky has emerged as a novel social media platform attracting significant attention. In its two years of life, the platform reached a user base of 30 million [1], registering a net increase in 2024 [1]. Factors contributing to its growth include innovative functionalities (e.g. customized feeds) [9] and also the context of the competitors social platforms. The so-called X-odus phenomenon, during which scientists and journalists departed from other platforms for various reasons (including those related to Donald Trump and Elon Musk [6]), has further accentuated its relevance. In the wake of the US elections and the ongoing mass migration, it is interesting to analyze it by using the same lens applied for other platforms.

This research addresses two central questions:

1) What is the structure of the interaction networks on Bluesky?
2) What are the main topics discussed on this emerging platform?

Social Network Analysis (SNA) and topic modeling are employed to address these questions. The approach combines techniques such as centrality metrics, community detection, and BERTopic for extracting latent topics, as demonstrated in previous studies on social media [2], [7], [10].

II. Related Work

Social media platforms have long been studied from both structural and content perspectives. SNA revealed that social networks typically follow power-law degree distributions, where a small number of users (hubs) attract a majority of interactions [2], [12]. Subsequent studies have applied various centrality measures to quantify influence and information flow [13]. The Louvain method, introduced to optimize modularity in community detection, has been widely used to uncover hidden clusters within large-scale networks [14].

On the content side, topic modeling has evolved from traditional methods such as Latent Dirichlet Allocation (LDA) to newer approaches that integrate contextual embeddings. BERTopic, for example, leverages transformer-based embeddings to achieve more semantically coherent topics [8], [10], [11].

III. Methodology

This section outlines the data sources, preprocessing steps, and analysis techniques applied in the study.

A. Data Collection and Preprocessing

1) Data Collection

The dataset employed in this study was obtained from the Hugging Face repository `alpindale/two-million-bluesky-posts`. It consists of 2 million public posts from Bluesky, collected through the platform's firehose API. Each post contains text content, metadata, and information about media attachments and reply relationships. Given computational constraints, the analysis has been conducted on the first one million posts having more than 10 digits, focusing on the following elements:

- `text`: The main content of the post.
- `author`: The post author.
- `uri`: Unique identifier for the post.
- `reply_to`: URI of the parent post if the post is a reply.

2) Data Preprocessing

To ensure the quality and relevance of the analysis, the data was subjected to a multi-step pipeline:

- **Language Filtering:** Posts are filtered to retain only English content using the `langid` library.
- **Text Cleaning:** The cleaning process includes expanding contractions, removing URLs and non-alphabetic characters, and normalizing whitespace [2]. Precompiled regular expressions and SpaCy's English model are employed for tokenization and lemmatization.

- **Parallel Processing:** Language detection and text preprocessing are parallelized using Python's `Pool` to efficiently process the large dataset.

B. Social Network Analysis

1) Network Construction

A directed network is constructed, where each node represents an author and each directed edge represents a reply. Edges are weighted by the number of replies between two authors. A mapping from post `uri` to `author` is used to generate these edges efficiently.

2) Centrality Measures and Degree Distribution

The following network metrics are computed:

- **In-Degree and Out-Degree:** For a node v, the in-degree $d_{in}(v)$ and out-degree $d_{out}(v)$ are defined as

$$d_{in}(v) = \sum_{u \in V} A_{uv}, \quad d_{out}(v) = \sum_{u \in V} A_{vu}, \tag{1}$$

  where A_{uv} is the element of the adjacency matrix corresponding to an edge from node u to node v.
- **Betweenness Centrality:** Measures the length to which a node lies on the shortest paths between other nodes. The betweenness centrality of a node v is given by:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{2}$$

  where σ_{st} is the total number of shortest paths between nodes s and t, and $\sigma_{st}(v)$ is the number of those paths passing through v.
- **Closeness Centrality:** Measures how close a node is to all other nodes in the network. It is defined as:

$$C_C(v) = \frac{1}{\sum_{u \in V} d(v,u)}, \tag{3}$$

  where $d(v,u)$ is the shortest path distance between nodes v and u.
- **Degree Distribution Analysis:** The `powerlaw` library is used to analyze both the in-degree and out-degree distributions [5] and the Kolmogorov-Smirnov (KS) statistic is used to evaluate how well the data fits the distribution.

3) Community Detection

The Louvain algorithm is employed to detect communities, given its efficiency on large datasets and short texts. The graph is converted from directed to undirected for this purpose, and modularity is used to evaluate the quality of the communities. The modularity Q is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{4}$$

where m is the total number of edges, k_i and k_j are the degrees of nodes i and j, and $\delta(c_i, c_j)$ equals 1 if nodes i and j belong to the same community, and 0 otherwise.

C. Topic Modeling

BERTopic is utilized for topic modeling as follows:

- Preprocessed texts are transformed into embeddings using the `MiniLM-L6-v2` model, a lightweight transformer-based sentence embedding model [11].
- Dimensionality reduction is performed using Uniform Manifold Approximation and Projection (UMAP) to project high-dimensional embeddings into a lower-dimensional space while preserving their structure.
- Clustering is conducted using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to identify coherent topic groups. HDBSCAN dynamically determines the optimal number of clusters, allowing for more adaptive topic discovery.
- The class-Term Frequency-Inverse Document Frequency (c-TF-IDF) weighting technique is applied to the extracted clusters, refining topic representation by emphasizing unique words of each topic.
- The resulting topics are labeled based on their most representative words, extracted from the c-TF-IDF model.

A heatmap is used to assess the semantic similarity between topics. The trained BERTopic model is stored for subsequent analysis, allowing for further exploration of the topics.

IV. Preliminary Results

This section contains key outputs from the analysis.

A. Preprocessing

A brief overview of the processed dataset is provided, where each line represents a post with the following elements: text, author ID, post ID, and a label generated by BERTopic, like:

- **Text:** Interesting polling datum national public attitudes...
- **Author:** did:plc:5ug6fzthlj6yyvftj3alekpj
- **URI:** `at://did:plc:5ugkpj/app.bsky.feed.post...`
- **Label:** 4_twitter_tweet_delete_deactivate

Before preprocessing, the number of posts was 1,000,000; after preprocessing, it decreased to 439,377, indicating a reduction of 56.06%.

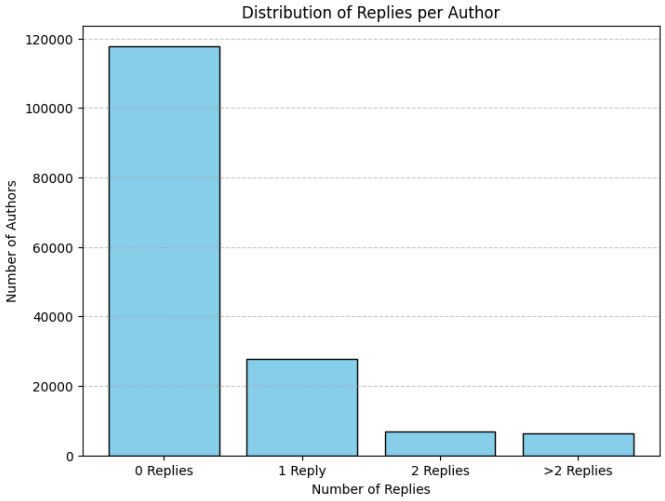
B. Social Network Analysis Results

a) Dataset and Graph Statistics

Directed Graph: 158,742 nodes, 79,403 edges. The following table represents the most influential authors with the respective measures.

TABLE I
NETWORK METRICS FOR MOST POPULAR AUTHORS

Author	In-Degree	Out-Degree	Betweenness	Closeness
did:plc:bybhkdeow67zttcythrzu4p6	312	0	0.0	0.11346
did:plc:cqlpijusxuy4u3viikknppyv	260	0	0.0	0.10293
did:plc:vovinwhtulbsx4mwfw26r5ni	259	0	0.0	0.11894



b) Power-law

The power-law exponents for in-degree and out-degree are, respectively:

- in-degree: 2.95 (KS: 0.0337)
- out-degree: 4.50 (KS: 0.0453)

These results confirm that the Bluesky network structure follows a heavy-tailed degree distribution, typical of social networks.

c) Community Detection

After converting the graph to undirected, 158,742 nodes and 65,291 edges were registered, and 114,825 communities were detected (modularity: 0.93). The SNA analysis took 2 minutes and 30 seconds (155.99 seconds).

C. BERTopic Outcome

The total number of topics identified is 4,770. Posts labeled as -1 (representing outliers) were removed from the analysis as in [2]. The entire BERTopic execution and preprocessing procedure took 2 hours and 2 minutes (7330.80 seconds).

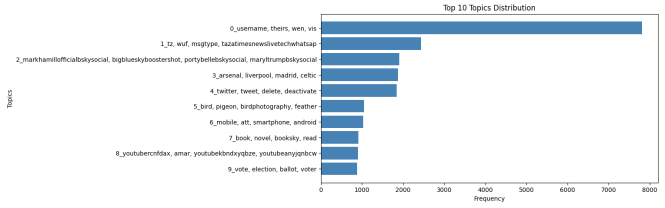


Fig. 1. Top 10 Topics Distribution

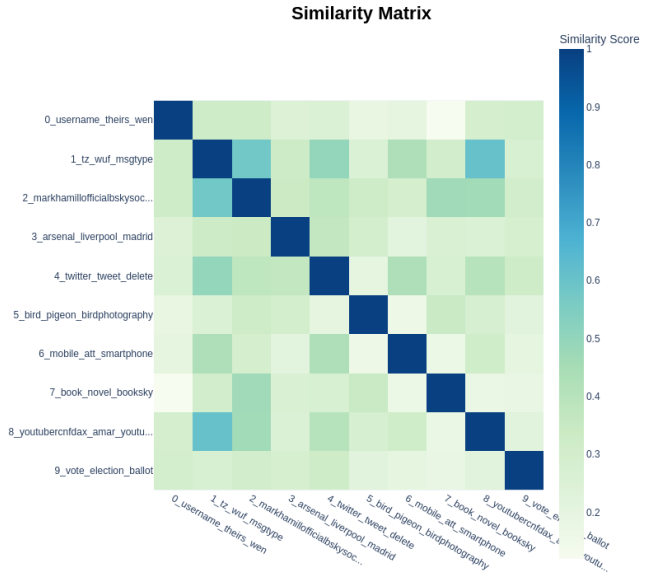


Fig. 2. BERTopic Heatmap Visualization

V. INTERPRETATION OF RESULTS

The analysis shows that the directed graph composed of 158,742 nodes and 79,403 edges is very sparse, with an edge-to-node ratio of approximately 0.5 and a density of about 3×10^{-6}. Approximately 74.24% of authors receive no replies, while only 4% receive more than two. Table I shows that the three most influential users (did:plc:bybhkdeow67zttcythrzu4p6, did:plc:cqlpijusxuy4u3viikknppyv, and did:plc:vovinwhtulbsx4mwfw26r5ni) exhibit a similar interaction pattern: they receive a high number of replies (312, 260, and 259, respectively) but do not reply themselves (out-degree = 0); their betweenness centrality is 0; their closeness centrality ranges from 0.1029 to 0.1189. This suggests that they function as broadcast nodes, likely posting content that generates discussion without engaging in conversations, that are relatively well-connected within the network but not the most central figures, and do not act as bridges. The power-law exponent and the Kolmogorov-Smirnov statistic confirm the fit of the data into a heavy-tailed degree distribution for both the in-degree (2.95, KS: 0.0337) and out-degree (4.50, KS: 0.0453). The Louvain algorithm uncovered 114,825 communities, nearly as many as the nodes in the undirected network (158,742), with a modularity of 0.93. This implies that most users are grouped into isolated clusters with extremely limited cross-group interactions. The BERTopic algorithm identified 4,770 distinct topics, from which the top 10 topics distribution chart has been derived. The plot shows that discussions are dominated by user mentions and usernames (e.g., Topic 0: "username, theirs, wen, vis") with 7,808 posts, and include themes such

as sports, social media discourse, literature, and politics. The similarity matrix heatmap of the ten most trending topics reveals varying degrees of topic overlap, with some topics sharing common discussion patterns (e.g., 1-2, 1-8), while others remain highly distinct (e.g., 5-6, 6-7). These findings indicate that Bluesky hosts a diverse range of niche conversations within a fragmented discussion landscape.

VI. Conclusion

The SNA confirms that Bluesky follows a power-law degree distribution, where a few highly influential users dominate interactions while the majority remain silent. The high modularity score from community detection suggests that the platform fosters isolated groups with minimal cross-community interaction. Topic modeling with BERTopic uncovered a diverse yet fragmented conversation space, where discussions span literature, sports, social media, and politics. However, the analysis also underscores the challenges of extracting coherent topics from short text, as evidenced by ambiguous clusters with low interpretability.

Despite these insights, the research has limitations. First, the static snapshot approach means that the temporal evolution of network structure and discourse dynamics remains unexplored. Future studies should incorporate longitudinal analyses to assess how interactions and topics shift over time. Second, as the user arena grows, it is statistically and computationally challenging to extract data that are truly significant. Third, the high number of detected communities and topics suggests that additional filtering techniques may be required to refine results and improve interpretability.

In conclusion, the findings suggest that Bluesky retains many structural characteristics of mainstream platforms while offering a segmented discussion environment. As the platform grows, further research will be necessary to assess how network topology, user behavior, and discourse evolution continue to shape its role in the broader social media ecosystem.

References

[1] https://vqv.app/stats/
[2] https://arxiv.org/pdf/2404.18984
[3] https://explodingtopics.com/blog/bluesky-users#bluesky-user-stats
[4] https://www.science.org/content/article/old-twitter-scientific-community-finds-new-home-bluesky
[5] https://arxiv.org/pdf/1305.0215
[6] https://www.nature.com/articles/d41586-024-03784-6
[7] https://arxiv.org/pdf/2308.04124
[8] https://easychair.org/publications/paper/HD6L/open
[9] https://arxiv.org/pdf/2404.18984
[10] https://arxiv.org/pdf/2203.05794
[11] https://doi.org/10.3389/frai.2024.1329185
[12] Barabási, A.-L. (2014). Network Science. Chapter 4: Scale-Free Networks
[13] Newman, M. E. J. (2010). Networks: An Introduction. Oxford University Press
[14] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008