POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A computational framework to study online conspiracy theories on Reddit after Epstein's death

Tesi di Laurea Magistrale in
Computer Science Engineering -
Ingegneria Informatica

Author: **Asja Attanasio**

Student ID: 11012024
Advisor: Prof. Francesco Pierri
Co-advisors: Francesco Corso, Gianmarco De Francisci Morales
Academic Year: 2024-25

Abstract

This thesis examines how increased visibility influences the dynamics of conspiracy communities, specifically investigating whether Reddit's homepage exposure following Jeffrey Epstein's death altered the composition of *r/conspiracy*'s user base, their patterns of participation, and their linguistic integration processes. Understanding these dynamics is essential, as platform algorithms determine which millions of users are exposed to conspiracy content, yet it remains unclear whether algorithmic exposure fosters forms of community membership comparable to those emerging from organic discovery.

Through a computational approach integrating analyses of toxicity, user retention, and the lexical and semantic dimensions of language, the research shows that homepage visibility operates as a complex selection mechanism rather than a simple amplifier. Users who discovered the community organically before Epstein's death tended to integrate more rapidly into its linguistic and thematic norms, whereas those drawn through homepage exposure maintained a greater semantic distance from the core discourse. The user retention analysis confirms that organically arriving users displayed more stable engagement, while toxicity analysis reveals an immediate decline among core members. Finally, lexical analysis highlights significant thematic divergence between long-standing members and newcomers.

The findings show that algorithmic visibility alters audience size, cultural cohesion, and ideological evolution. Newcomers who do not join organically arrive with different incentives and social structures; they integrate poorly, leave quickly, and remain separate from existing members, which prevents organic growth. In parallel, in this higher-risk community, typical radicalization pathways described in traditional narratives do not persist; visibility changes the audience without producing durable radicalization.

Keywords: conspiracy community, Reddit, algorithmic exposure, radicalization.

Abstract in lingua italiana

Questa tesi esamina in che modo l'aumento della visibilità influenzi le dinamiche delle comunità complottiste, e in particolare se l'apparizione nella homepage di Reddit dopo la morte di Jeffrey Epstein abbia modificato la composizione degli utenti di *r/conspiracy*, i loro modelli di permanenza e i processi di integrazione linguistica. Comprendere questi fenomeni è essenziale, poiché gli algoritmi delle piattaforme determinano quali milioni di utenti entrano in contatto con contenuti complottisti, ma non è ancora chiaro se l'esposizione algoritmica produca forme di appartenenza simili a quelle derivanti da una scoperta spontanea.

Attraverso un approccio computazionale che integra l'analisi della tossicità, dei tempi di permanenza e delle dimensioni lessicali e semantiche del linguaggio, la ricerca mostra che la visibilità in homepage opera come un meccanismo di selezione complesso, più che come un semplice amplificatore. Gli utenti che avevano scoperto la comunità per interesse spontaneo, prima della morte di Epstein, tendevano a integrarsi rapidamente nel linguaggio e nei temi del gruppo, mentre quelli arrivati grazie all'esposizione in homepage mantenevano una distanza semantica più marcata dal discorso centrale. L'analisi della fidelizzazione conferma che gli utenti giunti in modo organico alla comunità mostrano una partecipazione più costante, mentre quella della tossicità rivela un calo immediato tra i membri più attivi. Infine, l'analisi lessicale evidenzia una divergenza tematica significativa tra i membri storici e i nuovi arrivati.

Nel complesso, i risultati indicano che l'aumento di visibilità modifica la composizione del pubblico, il grado di coesione culturale e le traiettorie ideologiche. I nuovi membri che arrivano in modo non organico hanno incentivi e strutture sociali diverse: partecipano poco, restano per breve tempo e rimangono separati dai membri già presenti, ostacolando una crescita realmente organica della comunità. Parallelamente, in questa comunità rischosa, i percorsi di radicalizzazione descritti dalle narrative tradizionali non producono effetti duraturi: la visibilità attira nuovi arrivi, ma non genera una radicalizzazione duratura.

Parole chiave: visibilità algoritmica, comunità complottiste, Reddit, radicalizzazione.

Contents

Introduction

Online conspiracy communities have become critical environments for analyzing how digital platforms influence ideological discourse and collective meaning-making [31, 35, 44, 46]. Reddit's *r/conspiracy*, hosting over 2.2 million members [37], stands as one of the largest self-identified communities of "free thinkers" online, offering an optimal context for examining how external events and platform mechanisms shape conspiracy engagement, user retention, and linguistic integration patterns.

The death of Jeffrey Epstein on August 10, 2019, inside the Metropolitan Correctional Center in New York, occurred under highly irregular circumstances, including camera malfunctions, well-documented connections to powerful political and business figures [33], and the failure of guards to perform mandatory checks [16, 17], became an example of how online conspiracy discourse can momentarily align with mainstream skepticism.

On Reddit, *r/conspiracy* became a central point for this skepticism, as discussions about Epstein's death surged, temporarily bridging fringe communities and mainstream publics [36], by bringing this subreddit on Reddit's homepage, hence exposing the community to millions of users who would not typically encounter conspiracy content. This platform visibility mechanism constitutes a powerful yet understudied influence on online community dynamics. While previous scholarship has investigated conspiracy engagement pathways [34, 36, 43], linguistic adaptation within online communities [22], and content moderation intervention effects [30, 35, 38, 42], the role of algorithmic visibility in determining conspiracy community membership and linguistic integration remains substantially unexplored.

This thesis examines how Reddit's homepage visibility following Epstein's death operated as a compositional filter altering *r/conspiracy*'s user composition and cultural integration dynamics. Specifically, this research addresses three interconnected questions:

[RQ1] How do sudden visibility events alter the behavioral and linguistic signatures of the established community members?

To investigate this question, we conducted three complementary analyses of core user

discourse patterns surrounding Epstein's death. First, we employed interrupted time series analysis to assess whether the sudden community visibility following August 10, 2019 triggered shifts in offensive language and discourse tone in core users. Second, we applied Empath lexical category analysis to characterize thematic shifts in core users' language before and after the event. Third, we trained a binary classifier using Empath-derived lexical features across temporal folds to quantify discriminability between pre and post-event discourse.

[RQ2] How does homepage visibility affect community composition and user retention patterns?

To address this question, we constructed four mutually exclusive user cohorts based on temporal boundaries and the topic of their first post, respectively joining between arrest and death versus after death, and initial engagement with Epstein-related versus general conspiracy content. We then applied Kaplan-Meier survival analysis to model retention trajectories across these cohorts, treating users still active at the dataset's temporal endpoint as right-censored observations. Then, the Integrated Brier Score was employed to evaluate prediction accuracy across cohorts, enabling systematic comparison of engagement persistence between users discovering the community through homepage exposure versus organic pathways.

[RQ3] Does the timing and context of community entry influence linguistic integration patterns?

To measure linguistic integration we employed Sentence-BERT embeddings to calculate semantic similarity between newcomer posts and established community discourse. This temporal analysis evaluated whether users arriving during heightened visibility demonstrated persistent linguistic divergence from community norms compared to those joining through intentional discovery.

The analysis employed this multi-method computational framework to examine the mechanisms through which platform visibility shapes conspiracy community dynamics. It helps to further understand platform governance and online radicalization dynamics.

Outline

This thesis is organized into six chapters that progressively build from foundational concepts to empirical findings and broader implications for platform governance.

The investigation begins with Chapter 2, which establishes the necessary foundation for understanding Reddit's platform architecture, the specific time period surrounding Epstein's death, and *r/conspiracy*'s role as a research site. This chapter also introduces the computational tools employed throughout the analysis.

The research continues with Chapter 3 that situates this research within existing computational social science scholarship on online conspiracy communities. This chapter also identifies key gaps in current understanding and articulates the novel contributions of this thesis.

Chapter 4 details the comprehensive computational framework employed to address the three research questions. The chapter progresses systematically from data collection and cleaning procedures through multiple complementary analytical approaches, to finish with retention modeling and linguistic integration measurements.

Then, Chapter 5 synthesizes findings across these analytical dimensions.

Finally, Chapter 6 distills the central insight of the research, while acknowledging methodological limitations and proposing future research directions.

1 | Background

1.1. Reddit

Reddit [15] is a social online platform founded by Steve Huffman and Alexis Ohanian in 2005, designed as a space where users submit content, referred to as *submission*, in the form of links, textual posts, images, or videos. Other users can respond to these submissions through *comments*, creating threaded discussions beneath each post. The platform's core mechanism is a voting system: users up-vote or down-vote both submissions and comments, directly influencing their visibility and ordering. This voting system is particularly consequential for submissions appearing on Reddit's homepage, which aggregates highly up-voted posts from across the platform, effectively serving as a global visibility amplifier. Reaching the homepage represents peak exposure within Reddit's ecosystem, as it places content before millions of users who may not subscribe to the originating community.

Content on Reddit is organized into thematic subdivisions called *subreddits*, each operating under its own rules and cultural norms enforced by volunteer moderators from the community's membership. The platform combines forum features with social news aggregation, as content flow is determined by user engagement rather than chronological submission or editorial curation. This hybrid model has driven substantial growth, reaching 52 million daily active users across over 138K active subreddits [37]. However, Reddit's relative anonymity and decentralized moderation have also fostered environments where toxic behavior such as harassment, hate speech, and misinformation can proliferate [19, 27]. This darker side has made Reddit a significant site of study for computational social scientists examining online toxicity, community dynamics, linguistic conformity, and the effectiveness of moderation interventions across diverse digital communities.

From a usage perspective, Reddit attracts substantial global traffic, consistently ranking among the most visited websites in the United States, though its user base remains disproportionately English-speaking [26, 29, 41]. This concentration of activity makes Reddit particularly valuable for researchers studying English-language online discourse, while also highlighting limitations in its representativeness of global digital communities.

1.2. r/conspiracy

The subreddit *r/conspiracy* is one of Reddit's largest and most active conspiracy theory community, founded in January 2008 during a period of growing online disbelief toward mainstream narratives. As of 2025, the subreddit has grown to more than 2.2 million members, making it arguably the largest self-described community of *free thinkers* on the internet [9]. The subreddit's history reflects broader transformations in conspiracy theory culture, evolving from discussions of historical events like JFK's assassination and moon landing skepticism to contemporary political controversies. This evolution positioned the community at the intersection of multiple cultural currents: distrust of mainstream institutions, alternative media consumption, and increasingly polarized political discourse.

The subreddit operates under 10 moderator-enforced rules designed to balance open discourse with content moderation, including prohibitions on bigoted slurs, ad hominem attacks, and misleading headlines. Despite these guidelines, the user engagement remain sustained: a 2024 article documented that *r/conspiracy* generated approximately 360 new discussion threads daily, each receiving an average of 20 comments[23].

Researches have documented that dramatic real-world events correspond to significant influxes of new users to *r/conspiracy*, with some users joining Reddit specifically to participate in conspiracy discussions following major incidents [39]. Events involving high-profile figures and suspicious circumstances prove particularly catalytic for *r/conspiracy* activity. The death of financier Jeffrey Epstein in August 2019 exemplifies this phenomenon: mentions of Epstein's name spiked dramatically on *r/conspiracy* immediately following his death, with conspiracy-related content thriving in the subreddit during this period. This event generated sustained engagement that extended beyond the initial news cycle, with discussion frequency fluctuating in response to new revelations and the emergence of memetic phrases that eventually entered mainstream discourse. The Epstein case thus presents a methodologically valuable opportunity to examine how *r/conspiracy* responds to real-world events that align with the community's existing skepticism toward official narratives, offering insights into user behavior, content dynamics, and community evolution during periods of heightened attention surrounding controversial deaths and alleged conspiracies.

1.3. The Epstein case

As previously mentioned, Jeffrey Epstein's death on August 10, 2019, created an extraordinary convergence between mainstream skepticism and conspiracy theory discourse [20].

His death occurred in a high-security facility where multiple violations of standard prison procedures coincided with his apparent suicide [16, 17]. The circumstances proved immediately suspect: Epstein was not under suicide watch despite having been injured in an unexplained incident just days earlier, and his claims to possess compromising information about powerful figures had been well-documented [33]. These anomalies triggered public doubt that transcended typical ideological boundaries, with polls conducted shortly after his death revealing that only 29% of U.S. adults believed he actually died by suicide, while 42% thought he was murdered to prevent testimony against powerful associates [10, 11, 32]. This widespread skepticism solidified into a cultural phenomenon when the phrase *Epstein didn't kill himself* went viral in November 2019, as new details about his death continued to emerge. The phrase evolved from a conspiracy theory assertion into an internet meme, spreading across social media platforms and appearing in unexpected contexts ranging from political commentary to popular culture [12, 24].

This represented an unusual moment when conspiracy-oriented communities found their characteristic distrust of official narratives validated by mainstream commentary, including remarks from prominent political figures across the spectrum, even President Trump retweeted conspiracy theories linking Epstein's death to former President Bill Clinton [21, 45]. This convergence between fringe and mainstream discourse transformed *r/conspiracy* from a marginal community into a space where discussions aligned, however temporarily, with broader public sentiment.

1.4. Employed technologies

In the section below, there is a concise description of all the technologies that have proven useful during the development and implementation of this research work.

1.4.1. Pushshift API

We used the Pushshift API [5] to retrieve the corpus of data used in this research. Pushshift [18] is a comprehensive archive of Reddit submissions and comments that researchers use to study online communities at scale. The system continuously collects data from Reddit's public API and stores it in PostgreSQL databases, with Elasticsearch enabling efficient searches across billions of posts. Pushshift makes this data available through several channels: a public API for real-time queries, monthly compressed files in JSON format, and aggregation tools for summary statistics.

However, it is worth noting that Pushshift's access policies have changed significantly

since 2023 [3], with API access now restricted primarily to moderators and authorized researchers. The availability of these archived monthly dumps enabled complete temporal coverage of *r/conspiracy* activity throughout 2019 while maintaining consistency with established computational social science practices for studying Reddit communities.

1.4.2. Polars

Polars [4] is a high-performance DataFrame library for Python designed for efficient manipulation and analysis of large-scale structured data. Unlike traditional DataFrame libraries such as pandas, Polars leverages a columnar memory format, enabling vectorized query execution and automatic parallelization across multiple CPU cores without requiring additional configuration.

In this research, Polars was employed to handle the complete 2019 corpus of *r/conspiracy* submissions and comments, facilitating memory-efficient filtering, temporal aggregation, and cohort construction operations that would be computationally prohibitive with conventional DataFrame libraries. The library's expressive API, which supports chained operations for filtering, grouping, and joining, enabled streamlined data preprocessing pipelines while maintaining compatibility with other Python scientific computing libraries through zero-copy conversions to NumPy arrays and Arrow data structures.

1.4.3. Detoxify

Detoxify is a Python library [28] for classifying toxic content in text. It uses pretrained transformer models developed for Jigsaw's toxic comment classification challenges to predict categories such as toxicity, obscenity, insult, threat, and identity attack. In this context, for "toxic content" we denotes language that is likely to be perceived as offensive, hostile, or discriminatory toward individuals or groups.

In this research, only the toxicity parameter was used; that is, the score representing the likelihood that a post contains toxic language. This toxicity score was employed to investigate the trend of toxicity in posts within the community under study.

1.4.4. Empath

Empath is a Python library [25] that provides a tool for analyzing a text in terms of lexical categories. It comes with several pre-built lexical categories, such as "violence," "disputes," "government," and "politics," and allows you to calculate how much a text expresses each specified category.

Moreover, it is possible to create new lexical categories using vector-space word embeddings. For the purpose of this project, as discussed in Section 3.5, we customized a new lexical category called "paranoia."

1.4.5. Statistical methods

In this study we used two complementary statistical techniques to analyze temporal patterns: Interrupted Time Series (ITS) and the Kaplan-Meier estimator.

ITS [13] analysis is a quasi-experimental design that evaluates intervention effects by tracking outcomes over extended periods before and after a specific event, where the interruption represents the controlled external influence being studied. This method is particularly valuable when randomized controlled trials are infeasible, as it allows researchers to assess population-level interventions by comparing observed post-intervention trends against counterfactual projections based on pre-intervention patterns. The approach enables simultaneous quantification of two distinct intervention effects: level changes, representing immediate shifts in the outcome at the moment of intervention, and slope changes, capturing sustained alterations in the temporal trajectory. In this research, ITS was applied to evaluate the causal impact of Epstein's death on toxicity levels in *r/conspiracy*.

Additionally, survival analysis methods [2] were applied to model user retention patterns as time-to-event data, where the *event* of interest was user departure from the community and *survival* represented continued activity in *r/conspiracy*. The Kaplan-Meier estimator [14], also known as the product limit estimator, is a non-parametric statistic that estimates the survival function by calculating the cumulative probability that a user remains active beyond a given time point. At each observed event time when users stop posting, the estimator computes the conditional probability of survival as the number of users still active divided by the number at risk, then multiplies these successive probabilities to obtain the cumulative survival estimate. This multiplicative structure makes the Kaplan-Meier method particularly valuable for handling censored observations, specifically users still active at the dataset's temporal endpoint, whose true retention time remains unknown but exceeds the observation period. Treating these users as right-censored observations allows the estimator to incorporate partial information without introducing bias that would result from either excluding them or artificially assigning departure times.

We used Kaplan-Meier to construct survival curves for four mutually exclusive user cohorts defined by the timing of their entry (between arrest and death versus after death) and their initial engagement focus (Epstein-related threads versus general conspiracy content).

1.4.6. RandomForestClassifier

RandomForestClassifier [6] is a supervised machine learning algorithm from the scikit-learn library that implements an ensemble learning approach for classification tasks. The algorithm works by constructing multiple decision trees during training, where each tree is built on a random subset of the training data through bootstrap sampling and considers only a random subset of features at each split point. When making predictions, each tree in the forest independently classifies the input sample, and the final prediction is determined through majority voting across all trees. The algorithm handles high-dimensional feature spaces effectively, making it well-suited for lexical category features derived from Empath analysis. Also, Random Forests naturally resist overfitting through the combination of many diverse trees, which is particularly valuable when working with noisy social media data.

In this research, we employed this algorithm to build binary classifiers distinguishing core community members from newcomers based on temporal folds of Empath-derived lexical feature.

1.4.7. RandomUnderSampler

RandomUnderSampler [7] is a preprocessing technique from the imbalanced-learn library that addresses class imbalance problems by reducing the size of the majority class to match the minority class. It works by randomly selecting a subset of samples from the majority class to create a balanced dataset where both classes have equal representation. While this approach reduces the total amount of training data, it prevents the classifier from being overwhelmed by majority class patterns and ensures that both classes contribute equally to model learning.

In the context of this research, the distinction between core users and newcomers in *r/conspiracy* exhibited class imbalance: core users who were active throughout the study period generated disproportionately more content than newcomers who joined during visibility events. Applying RandomUnderSampler before training the RandomForestClassifier ensured that the binary classification analysis fairly captured linguistic patterns from both user groups rather than simply learning to identify the more prevalent class.

1.4.8. SBERT

Sentence-BERT (SBERT) [8] is a modification of the pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence

embeddings that can be compared using cosine similarity. Unlike traditional BERT, which requires feeding both sentences into the network simultaneously, SBERT generates fixed-size vector representations for individual sentences, enabling efficient semantic similarity computation at scale.

In this research, SBERT embeddings are employed to measure linguistic alignment between new users and the established *r/conspiracy* community. Specifically, we computed sentence-level embeddings for posts from the same user cohorts used in the survival analysis and calculated cosine similarity scores to quantify how semantically aligned newcomers' language is with core community discourse. This approach enabled us to capture the semantic convergence at a granular level, providing a more nuanced measure of linguistic integration than lexical or category-based methods.

2 | State of the Art

Having established Reddit's platform mechanics, *r/conspiracy*'s community characteristics, the Epstein event, and the analytical tools employed in this research, we now turn to the existing body of computational social science research that provides the foundations for our analysis.

2.1. Conspiracy engagement

Klein et al. [34] pioneered the systematic study of precursors to conspiracy engagement by examining Reddit users before their first post in *r/conspiracy*. Their study revealed that individuals who later engaged with conspiracy content already displayed distinctive linguistic tendencies and gravitated toward particular online spaces even before entering those communities. By examining language patterns and online interactions, the study found clear tendencies for users to engage with topics and groups connected to conspiracy-related views. Crucially, Klein et al. demonstrated that individuals inclined toward conspiracy thinking were not merely passive participants in these communities. Rather, they actively sought out platforms and groups that reflected their interests and reinforced their existing beliefs.

The research revealed that the trajectory toward conspiracy engagement stems from the interplay between individual psychological traits, such as an individual's conspiratorial mindset and their pre-existing skepticism, and social factors, including the availability of spaces that validate and amplify such perspectives. This underscores the idea that conspiracy engagement results from both internal motivations and active searching for compatible social environments, making the process more recursive and self-selective than previously assumed.

Building on this foundation, Phadke et al. [36] conducted a longitudinal analysis of users' behavior after joining *r/conspiracy* to trace the development of conspiracy engagement over time. They distinguished four principal participation trajectories: consistently high, increasing, decreasing, and consistently low involvement. Users showing stable or ris-

ing engagement showed progressive radicalization through adoption of insider language, participation in small-group discussions, and development of a monological conspiracy worldview characterized by engagement with increasingly generalist conspiracy content. In contrast, users following decreasing trajectories limited their participation to specialized conspiracy topics, maintained diverse community connections, and never developed strong lexical conformity with conspiracy communities.

2.2. Linguistic conformity and conspiratorial frameworks

Researchers focus not only on how new members eventually engage with a community, but also on the ongoing evolution of linguistic behavior within these spaces: this represents a fundamental challenge in understanding online dynamics. Danescu-Niculescu-Mizil et al. [22] developed an innovative framework for tracking linguistic change, providing empirical evidence that users undergo a well-defined two-stage lifecycle characterized by early adaptation and subsequent linguistic rigidity. During the initial linguistic adolescence which spans roughly the first third of a user's contributions, members are highly sensitive to existing norms and rapidly integrate community-specific language features, actively mirroring terminology and stylistic conventions as a strategy to gain acceptance and establish credibility within the group. However, as users transition beyond this period, they enter a conservative phase, where further adaptation slows markedly and individuals increasingly rely on the previously acquired linguistic habits and conventions. Their language begins to diverge from the ever-evolving community norm, resulting in a fossilization effect where the user's speech patterns remain fixed even as the broader group continues to innovate linguistically.

Beyond individual linguistic adaptation, Schatto-Eckrodt et al. [40] investigated conspiracy theory formation following Jeffrey Epstein's death, finding that conspiratorial narratives emerged rapidly, even before official news reports, by leveraging pre-existing conspiracy myths, established theories such as Pizzagate, and political tensions within these communities. Their analysis revealed that users who shared alternative news sources were more likely to reference unsubstantiated conspiracy theories, with alternative platforms such as 4chan and Gab exhibiting substantially greater conspiratorial content than mainstream platforms like Twitter and Reddit.

2.3. Platform interventions

Platform interventions targeting conspiracy communities reveal complex dynamics that extend beyond their immediate effects, since conspiracy communities demonstrate remarkable resilience. Monti et al. [35] compared behavioral shifts following the ban of *r/GreatAwakening*, a QAnon conspiracy community, and *r/FatPeopleHate*, a hate speech community, finding that conspiracy community members were substantially more likely to completely migrate from Reddit to the unmoderated platform Voat, maintain higher activity levels on the new platform, and successfully recreate their social network structure. These findings suggest that conspiracy communities have stronger social bonds and a deeper attachment to shared narratives than communities centered on general hate speech. The authors attribute this resilience to the way conspiracy narratives shape participants' identities and encourage their commitment to preserving collective knowledge and shared values.

However, interventions designed to restraint harmful content produce complex and occasionally counterproductive consequences, demanding a deeper understanding of their aftermath. Trujillo and Cresci's [42] comprehensive analysis of *r/The_Donald* revealed that while quarantine and restriction successfully reduced user activity, they generated adversarial side effects: the quarantine initially reduced toxicity within the subreddit, but this effect decayed over approximately six months, with toxicity eventually surpassing pre-intervention levels. Moreover, interventions caused affected users to share progressively more politically polarized and less factual news content, both within *r/The_Donald* and across other online spaces to which they migrated. These spillover dynamics suggest that users subject to restrictive interventions may actively seek alternative spaces for expression, potentially amplifying the visibility and spread of harmful narratives in less regulated environments. Consequently, effective management of online conspiracy ecosystems requires sustained and context-sensitive approaches that anticipate unintended outcomes and the mobility of online conspiracy ecosystems.

2.4. Contributions

This research extends the existing body of work by:

- Introducing a novel methodology for measuring linguistic alignment in online conspiracy communities through sentence-level semantic embeddings using SBERT;
- Examining how rapidly newcomers align linguistically with established conspiracy communities members;

- Determining whether linguistic alignment patterns predict long-term user engagement.

3 | Methods

This chapter discusses in detail the methodological framework undertaken in this research that led to our final results. The methods described below combined quantitative and computational approaches with the aim of answering our research questions.

3.1. Data collection

The data for this study consisted of all Reddit comments and submissions from 2019, obtained from a large-scale aggregator, Pushshift, through compressed archives of Reddit's posts aggregated by month.

To focus the analysis on the *r/conspiracy* subreddit, each monthly archive was decompressed and parsed, retaining only submissions and comments uttered in the target community. Subsequently, the filtered corpus was stored using the Polars DataFrame.

Taken together, the data collection steps produced two optimized Polars datasets representing the complete 2019 history of user comments (2.5M) and submissions (96K) exclusively within the *r/conspiracy* community.

3.2. Data cleaning

Following data collection, a comprehensive cleaning procedure was applied to ensure the integrity and usability of the dataset. In fact, the initial datasets were first filtered to exclude any entries with missing or deleted authors, or null or deleted text content, ensuring that only meaningful contributions were retained. For future reference, the union of the comment dataset and the submission dataset is collectively referred to as posts. The results of the filtering procedure are summarized in Table 3.1, which reports the final counts of cleaned corpus.

Type	Size
Submissions	73 320
Comments	2 207 905
Total posts	2 281 225

Table 3.1: Number of submissions, comments, and total posts after data cleaning.

In addition to filtering rows, only a subset of the original fields provided in the compressed files was retained for further analysis. This selection was made to focus on variables relevant to linguistic, temporal, and user-level analyses, while excluding extraneous metadata. A full list of the retained fields is provided in Appendix A.

To facilitate consistent processing across the combined dataset, several fields were standardized:

- For comments, the field *body* of each comment was renamed *text*;
- For submissions, the *title* was concatenated with the post content to create a single *text* field representing the full content of the submission.
- For both comments and submissions, the *created_utc* timestamp was converted to a datetime object and stored as *date*.

At this point, we have ensured a consistent structure between comments and submissions, enabling their integration as a single dataset for subsequent analyses.

3.3. Data exploration

Since the goal of the research was to characterize how external news events drive community discovery and how different user cohorts exhibit distinct engagement trajectories, in the data exploration, we focused on investigating temporal patterns of new and core users' engagement within *r/conspiracy* before, during, and after the Epstein case gained widespread attention. We performed both a general exploration of the entire community, and a more focused one on the time window between Epstein's arrest on July 6, 2019, and his death on August 10, 2019, as well as the subsequent month when *r/conspiracy* achieved heightened visibility on Reddit's homepage. This focused temporal segmentation allows to differentiate between users who discovered the community during the anticipatory period (between arrest and death) versus those who arrived during the peak visibility phase following Epstein's death.

Temporal trends in users engagement

To extract users engaged in Epstein-related discussions we implemented a systematic filtering approach:

- Firstly, all submissions were screened using a case-insensitive regular expression pattern to identify threads explicitly mentioning Jeffrey Epstein or Ghislaine Maxwell, made explicit in Appendix A;
- Then, the *parent_id* field of all comments in the dataset was used to retrieve every comment posted within these identified Epstein-related submission threads.

From this filtered posts set, all unique user identifiers were extracted, establishing the population of users who actively engaged with Epstein-related content during the study period.

New user cohorts definition

To characterize the behavioral differences between users who discovered *r/conspiracy* spanning the two main events surrounding Epstein in 2019 (arrest and death), four mutually exclusive cohorts were constructed based on two temporal boundaries and initial behavior of participation. The temporal segmentation distinguishes between users who joined between arrest and death versus those who joined after death through one month later. The behavioral dimension differentiates users whose first contribution engaged with Epstein-related threads from those whose initial post addressed other conspiracy topics.

Hence, we highlighted four disjoint sets of users:

1. Users who made their first post in the community between the arrest and death, specifically in an Epstein-related thread.
2. Users who made their first post in the community between the arrest and death, but not in an Epstein-related thread.
3. Users who made their first post in the community during the month following the death, specifically in an Epstein-related thread.
4. Users who made their first post in the community during the month following the death, but not in an Epstein-related thread.

These cohorts are fundamental for examining whether different pathways to community entry also correspond to differences in user retention and linguistic integration, which we explore in subsequent analyses.

3.4. Toxicity trend

In this section, we investigate the potential relationship between the toxicity of core users' posts and the temporal window surrounding Jeffrey Epstein's death, employing automated text analysis methods that have proven effective for measuring antisocial behavior in online communities [42].

Core user definition and data preparation

To establish a baseline understanding of community behavior, we focused our analysis on core users, that are those who were already active participants in *r/conspiracy* prior to Epstein's arrest on July 6, 2019. This temporal constraint ensures that we are examining users with established participation patterns rather than those who joined the community specifically in response to the Epstein case. Filtering the dataset by this criterion yielded 70K core users who collectively contributed to more than 1.8M posts during the study period.

Once the corpus was extracted, we employed Detoxify to associate to each post a toxicity scores ranging from 0 to 1, where higher values indicate content that is rude, disrespectful, or likely to drive users away from a discussion.

A methodological challenge in analyzing daily toxicity trends is the substantial variability in post volume across different days. Days with few posts may produce unstable estimates, while days with many posts may be dominated by outliers or particularly active users. To address this challenge and ensure robust daily estimates, we implemented a bootstrapping procedure that controls for variability in daily post volume.

Specifically, for each day in our observation window, we performed the following procedure:

- Drew 100 random samples with replacement, where each sample consisted of up to 100 posts from that day. For days with fewer than 100 posts, we sampled with replacement from the available posts.
- Calculated the median toxicity score for each of the 100 samples. The use of the median rather than the mean provides robustness against extreme values that might excessively influence daily estimates.
- Computed the mean of these 100 median values to obtain a stable estimate of the typical toxicity level for that day.

This bootstrap aggregation procedure yields a time series of estimated mean median toxicity values, with one value per date, providing a robust daily measure of core users'

toxicity.

To conclude the data preparation, we enhanced the interpretability and visualization of the results by standardizing the daily mean median toxicity using the Standard Scaler transformation. This transformation converts the raw toxicity scores into z-scores using the formula:

$$z = \frac{x - \mu}{\sigma}$$

where x represents the daily mean median toxicity value, μ is the overall mean toxicity across the entire time series, and σ is the standard deviation.

Standardization centers the data around zero and scales it by the standard deviation, making it easier to identify deviations from the baseline toxicity level and to compare effect sizes across different time windows. This approach is commonly employed in time series analysis of online behavior, as it allows for intuitive interpretation of how many standard deviations a given day's toxicity deviates from the average [42].

ITS analysis

To evaluate the causal impact of Epstein's death on community discourse, we employed an Interrupted Time Series (ITS) analysis, a quasi-experimental design widely used to assess the effects of interventions on temporal data. ITS analysis is particularly well-suited for this context because it allows us to distinguish between pre-existing temporal trends and changes attributable to the heightened media attention and visibility of *r/conspiracy* following Epstein's death on August 10, 2019.

We specified a segmented linear regression model using ordinary least squares (OLS) estimation, treating August 10, 2019 as the intervention point t_0. As done in [42] the model takes the following form:

$$y_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 P_t$$

where:

- y_t represents the standardized daily mean median toxicity score at time t;
- D_t is a binary indicator variable equal to 0 before t_0 and 1 afterward (capturing any immediate level change in toxicity);
- P_t denotes the number of days elapsed since Epstein's death (capturing any change in the post-event trend);

- The baseline intercept β_0 represents the initial level of toxicity;
- β_1 captures the pre-existing linear trend in toxicity prior to Epstein's death;
- The coefficient β_2 quantifies the immediate change in toxicity on the day of the event;
- β_3 measures any shift in the slope of the toxicity trend following the intervention.

3.5. Lexical shift analysis

To characterize how community discourse evolved in response to major developments in the Epstein case, we conducted a lexical shift analysis examining changes in thematic language use among *r/conspiracy* users. This analysis focused on two critical events: Epstein's arrest on July 6, 2019, and his death on August 10, 2019.

For each event, we defined a two-month observation window comprising the month immediately preceding and the month following the event. This timeframes allow for adequate capture of baseline discourse patterns while minimizing misleading effects from temporally distant events. All posts (comments and submissions) authored within these windows were included in the analysis, providing sufficient linguistic data for robust statistical comparison.

We employed Empath, a validated Python library for analyzing semantic categories in text, to quantify usage of seventeen theoretically motivated thematic categories. These categories were selected are further discussed in Appendix A.

Two analytical approaches

We conducted two complementary analyses to capture different aspects of community response. First, we examined lexical shifts within the core user population to assess how established community members, as chosen in Section 3.4, adjusted their language use. Second, we compared lexical patterns between core users and new users to identify distinctive linguistic characteristics of users joining the community during or after the events. This comparison enables assessment of whether external events attracted users with fundamentally different discursive orientations.

- **Core user longitudinal analysis**

  For the analysis of core users across time, we identified all users who posted at least once both before and after each focal event, ensuring that shift scores represent within-individual change rather than compositional differences in the posting pop-

ulation. For each user, we computed Empath category scores for all their posts in the pre-event month and separately for all posts in the post-event month. We then calculated user-specific shift scores by subtracting the mean pre-event score from the mean post-event score for each category. Finally, we aggregated these individual shift scores by computing the mean shift across all qualifying users, yielding a single shift value per category that represents the average within-user change in thematic emphasis. This approach controls for individual-level baseline differences in language use and isolates the effect of the temporal event.

- **Cross-group comparison**

  For comparing new users with core users, we could not employ within-individual comparisons because new users had no pre-event posting history by definition. Instead, we computed the mean Empath category score for all core user posts in the month preceding the focal event, establishing a baseline representation of community discourse. For new users, we computed mean category scores across all posts authored from the event date through one month afterward. The shift score for each category was then calculated as the difference between the new user mean and the core user baseline mean. Positive shift values indicate that new users exhibited higher emphasis on that category compared to the pre-event baseline, while negative values indicate reduced emphasis.

In both of the two analysis described above, we had to test seventeen thematic categories, raising important statistical concerns regarding Type I error inflation from multiple comparisons. We addressed this by applying the Benjamini-Hochberg correction to control the false discovery rate across all tests. This statistical approach offers an appropriate compromise: it limits spurious findings while preserving sufficient power to detect genuine effects, which is especially valuable when examining correlated linguistic variables through exploratory analysis.

3.6. Binary user classification via Empath features

Beyond examining lexical shifts between user groups, we wanted to test whether these linguistic patterns could actually distinguish core users from newcomers when looking at individual posts. We framed this test as a supervised binary classification task, in which the target variable *is_core* indicated whether a post came from a core user (defined in Section 3.4), while features consisted of the seventeen Empath category scores we used in Section 3.5.

However, simply using raw Empath scores presented a problem: they only capture a snapshot of language use at a single moment, missing how discourse evolves over time. To address this, we created rolling window statistics tracking how each user's linguistic profile changed across different temporal spans. For every Empath category, we calculated rolling means and variances using window sizes of 3, 5, 7, and 14 time periods. These multiple windows let us capture both short-term fluctuations and longer-term trends in language use. We also computed percentage changes and mean differences between windows, which helped the model identify genuine shifts in thematic emphasis rather than just static patterns. This temporal engineering proved essential because, as our lexical shift analyses showed, discourse patterns changed substantially throughout the observation period.

Our dataset presented another challenge: core users vastly outnumbered new users, reflecting *r/conspiracy*'s established community structure. This class imbalance creates problems for classification models, which often achieve deceptively high accuracy simply by predicting the majority class every time. To mitigate this, we adopted a dual approach. First, we used *RandomUnderSampler* to reduce the number of majority class examples in our training data to balance class distributions. Second, we configured our *RandomForestClassifier* to penalize misclassification of minority class examples more heavily during the tree-building process.

For what concernes the validation, standard k-fold cross-validation is inappropriate for time series because it violates the inherit temporal order and leaks future information, yielding overly optimistic performance estimates. We instead implemented a time-based cross-validation scheme with 10 sequential folds. This non-contiguous splitting approach avoids information leakage while maximizing training data utilization, providing realistic estimates of how the model would perform on unseen temporal segments.

We evaluated model performance using Average Precision (AP) because it condenses the entire precision–recall curve into a single value and is well suited to imbalanced data. AP is computed as the weighted mean of precision at each recall level, where the weight is the increase in recall from the previous threshold; thus, higher AP reflects better ranking of true core posts near the top and stronger precision at corresponding recall levels.

3.7. Survival analysis

In this phase of the research, we aimed to assess whether user retention varies depending on the triggering event (arrest vs. death) and whether initial engagement with the Epstein topic correlates with longer or shorter user lifespans on the platform. The four cohorts compared in this analysis were already defined in Section 3.3. To address this

objective, a survival analysis [2] was conducted using the Kaplan–Meier estimator, which is appropriate for handling right-censored survival times.

Before computing the survival curve, the data had to be prepared. Specifically, for each author in the target cohort, we computed the following variables:

- The first observed activity date after the reference event.
- The last observed activity date after the reference event.
- The duration of participation in days, defined as the number of days between the first and last interactions for users who disengaged, or between the first interaction and the dataset's end date for censored users.
- An event indicator, set to 1 if the user became inactive for at least 30 days before the dataset end (interpreted as *churn*), and 0 otherwise (interpreted as *censored*).

This process yielded one observation per author, containing their respective duration and event flag variables, which served as the required inputs for the Kaplan–Meier estimation. After the data preparation, for each users cohort the estimator computed the conditional probability of a user remaining active as a function of time since their first post-event activity.

To evaluate the predictive accuracy of the Kaplan–Meier models, the Integrated Brier Score (IBS) [1] was computed for each cohort. The IBS measures the average squared difference between observed survival status and predicted survival probability, and return a value between 0 and 1, with lower values indicating better model calibration.

3.8. Linguistic analysis

In this section, we describe our approach to quantifying the semantic alignment between posts authored by new users and the prevailing linguistic patterns produced by core community members over time, to understand if timing and context of community entry influence linguistic integration patterns.

3.8.1. Building weekly language references

The first step in our analysis involved transforming all posts authored by core users into fixed-length vector representations that capture their semantic content. To accomplish this, we employed a state-of-the-art sentence embedding method: Sentence-BERT (SBERT).

Having obtained embeddings for all core user posts, we next sought to characterize the typical linguistic patterns exhibited by the core community during each week of the observation period. To this end, we computed a weekly centroid vector for each week by taking the arithmetic mean of all core user post embeddings that fell within that week's time window. Formally, the weekly centroid LM_w was defined as:

$$LM_w = \frac{1}{N_w} \sum_{i}^{N_w} SBERT(p_i)$$

where $SBERT(p_i)$ denotes the SBERT embedding of the i_{th} core user post occurring in week w, and N_w represents the total number of core user posts published during that week. This centroid serves as a reference point in the semantic space, effectively summarizing the collective semantic mean of the week's core user discourse. By computing a distinct centroid for each week, we were able to capture the evolving nature of community language over time.

3.8.2. Semantic alignment of new users

Our analysis focused on four distinct cohorts of new users, which were defined and described in detail in a previous section of this thesis (3.3). As already seen, these cohorts were constructed based on the timing of user entry into the community, specifically, whether they joined after a particular arrest event or after a subsequent death event, as well as the topical content of their first post, distinguishing between users who posted about Epstein-related topics and those whose did not. This cohort design allowed us to investigate whether the context in which new users enter the community, as well as their topical engagement patterns, systematically influence their linguistic alignment with core community members.

For each post authored by a new user, we computed a semantic alignment score that quantified the degree to which the post's language resembled the typical discourse of core users during the week in which the post was uttered. The first step in this process involved identifying the week which each new user's post belonged to, based on the post's timestamp. Next, we embedded the post using the same SBERT model that had been applied to core user posts, yielding a vector representation e_p in the same semantic space as the weekly centroids. To measure the semantic alignment between the new user post and the corresponding weekly centroid LM_w, we computed the cosine similarity between the two vectors, which captures the directional alignment of the vectors regardless of their magnitude. Cosine similarity is formally defined as:

$$sim(p, w(p)) = \frac{e_p \cdot LM_{w(p)}}{\|e_p\| \|LM_{w(p)}\|}$$

where the numerator denotes the dot product of the two vectors, and denominator represent their respective Euclidean norms. Cosine similarity ranges from -1 to 1, with values closer to 1 indicating greater directional alignment in the semantic space.

To facilitate interpretation, we transformed cosine similarity into cosine distance by subtracting the similarity from 1, yielding the semantic alignment score:

$$dis(p, w(p)) = 1 - sim(p, w(p))$$

This transformation ensures that higher values correspond to greater semantic distance, meaning that posts with higher alignment scores are less typical of the core community's language during that week. In other words, a new user post with a high cosine distance is semantically divergent from the prevailing discourse patterns of core users, while a post with a low cosine distance closely mirrors the linguistic norms of the core community at that point in time.

To examine temporal trends in semantic alignment at the cohort level, we aggregated the individual post-level alignment scores into weekly summaries for each of the four new user cohorts. Specifically, for each week and each cohort, we computed the mean cosine distance across all new user posts from that cohort that were published during that week. This yielded a time series of weekly mean semantic alignment scores for each cohort, capturing how the average linguistic divergence from core users evolved over the observation period. Because week-to-week fluctuations in these mean scores can be noisy and may obscure underlying trends, we applied a three-sample rolling mean to smooth the time series and make temporal patterns more easily interpretable. The rolling mean was computed using a standard moving window estimator, in which each smoothed value at time t is the average of the observed values at times $t-1$, t , and $t+1$. This smoothing procedure helps to reduce the influence of random week-to-week variation while preserving the general trajectory of semantic alignment over time.

4 | Results and limitations

This section presents the empirical findings from a multi-method analysis examining how Reddit homepage visibility following Jeffrey Epstein's death altered *r/conspiracy* community dynamics.

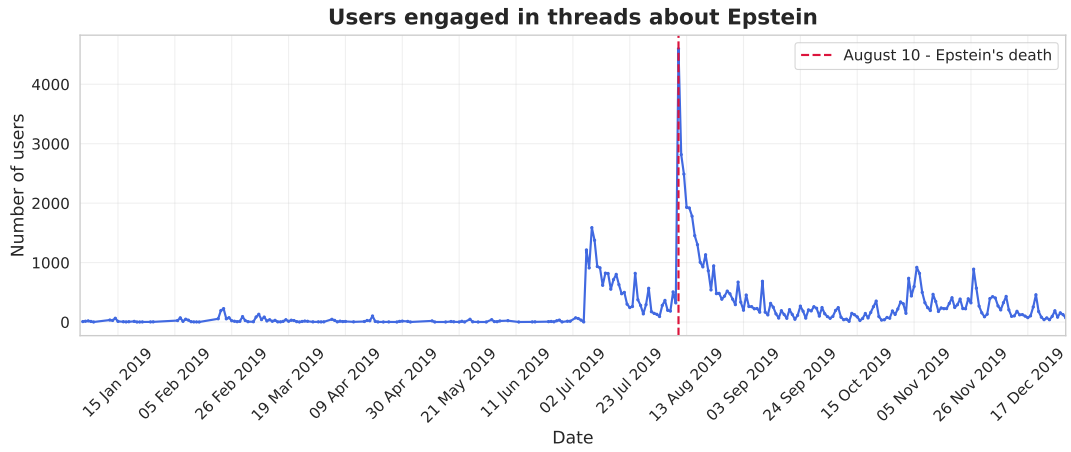
4.1. Exploration of temporal engagement

Before examining specific behavioral and linguistic metrics, we comment on temporal patterns in user engagement (Figure 4.1), that characterize how the Epstein case shaped *r/conspiracy*'s compositional dynamics.

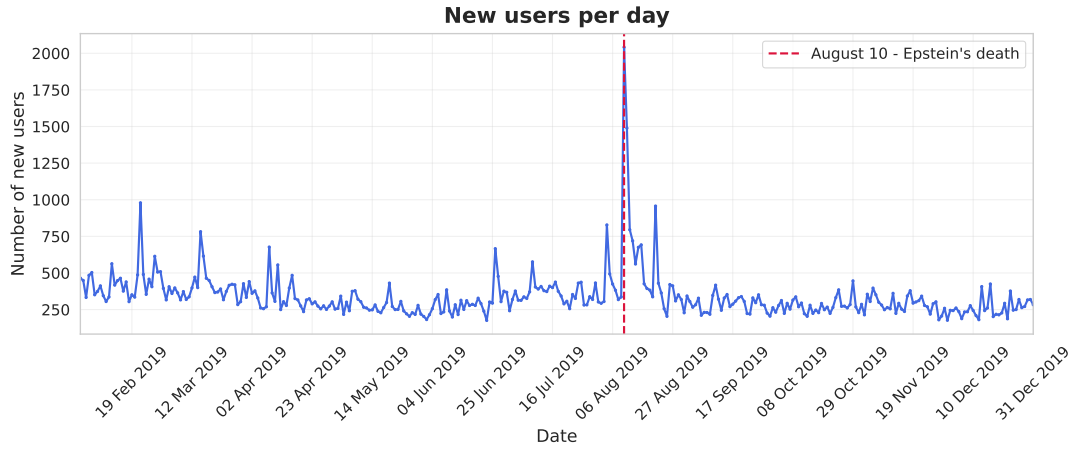
Both visualizations exhibit similar temporal patterns: during the first half of 2019, user engagement with Epstein-related content and new user arrivals to r/conspiracy remained relatively low, with modest baseline activity and usual user acquisition rates. A moderate increase in the posting activity corresponds to Epstein's arrest in early July, suggesting that the arrest generated initial attention but had not yet triggered widespread community discovery.

The most dramatic and synchronized pattern occurs the day of Epstein's death, which is clearly marked in both figures: Figure 4.1.a shows engagement spiking to more that 4K users, while Figure 4.1.b reveals a concurrent surge in new user registrations reaching approximately 2K users. This simultaneous spike in both engagement and new user acquisition demonstrates that Epstein's death intensified interest among existing community members and served as a primary driver of community discovery, attracting substantial numbers of users who had not previously participated in *r/conspiracy*. The magnitude of these peaks exceeds baseline activity observed in preceding months, confirming that this event represented a turning point for the subreddit's visibility and growth. Following the spike in the day of Epstein's death, both visualizations show elevated but gradually declining levels through the remainder of 2019.

This pattern suggests sustained interest beyond the immediate news cycle, with the subreddit continuing to attract new users and maintain heightened



(a)



(b)

Figure 4.1: (a) Time series of unique users engaged in threads about Epstein and his wife during 2019. (b) Time series of new users joining *r/conspiracy* in the same period.

engagement in Epstein-related discussions for weeks after the initial event.

4.1.1. Posting behavior across new user cohorts

By segmenting the user base into four distinct cohorts, we examined how the timing of users' arrival and their initial focus on Epstein-related content relate to later patterns of participation. As shown in Figure 4.2, the stacked area visualization reveals the relative contribution of each cohort to overall community activity over time.

The cumulative posting volume peaks dramatically in late August 2019, reaching more than 20K posts per week across all cohorts combined, before declining steadily through the end of the year.

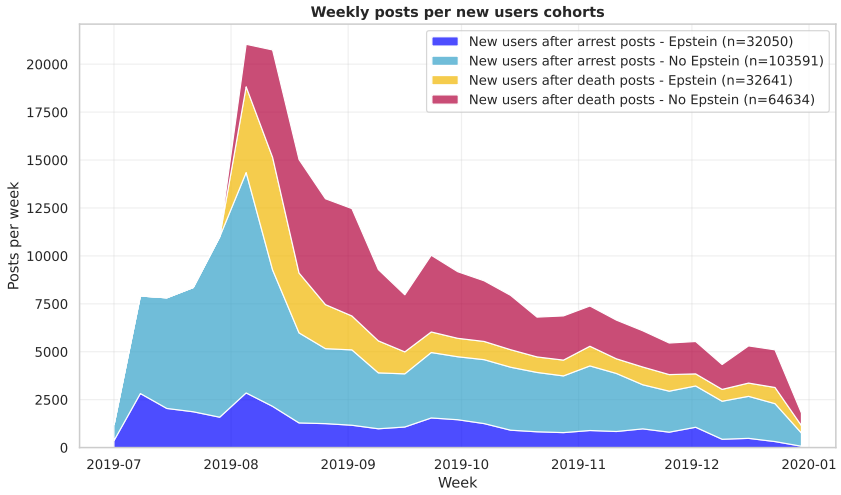


Figure 4.2: Longitudinal cohort analysis showing weekly posting activity for each of the four user cohorts.

The users who joined following the arrest and did not initially interact with Epstein content (shown in light blue) represent the largest single contributor to overall activity, particularly during the peak period in August-September. This cohort maintains the most substantial share of weekly posts throughout the observation window, occupying the largest area in the stacked plot. This dominant contribution is largely attributable to cohort size: the light blue group comprises approximately 104K users, substantially more than any other cohort. The second-largest contributor comprises users who entered after the death and avoided Epstein topics (red, around 65K users), showing particularly strong engagement during and after the August peak. Despite having roughly 40% fewer members than the light blue cohort, this group maintains a considerable share of overall activity, suggesting relatively high per-user engagement levels. The two smaller cohorts (orange and blue) contribute the smallest proportions of activity, with remarkably similar patterns. These cohorts are roughly equal in both size, approximately 32K users each, and posting behavior throughout the observation period.

Across all cohorts, posting activity declines steadily from the collective peak in late August toward the end of 2019, mirroring common retention patterns in online forums [42]. Despite this overall decrease, the relative ordering of cohort contributions remains stable, with the light blue and red cohorts maintaining their positions as primary contributors.

4.2. Preliminary behavioral and thematic trends

Before examining retention and linguistic integration directly, three preliminary analyses established the surface-level manifestations of homepage visibility effects.

4.2.1. Toxicity trends

The segmented regression model used to perform the toxicity analysis (specified in Section 3.4) revealed three key findings:

- Pre-intervention trend: The baseline toxicity trajectory showed a marginally significant upward drift ($p = 0.053$), indicating slight toxicity increases before August 10.
- Immediate impact: Toxicity dropped substantially at the intervention point, with a decrease of approximately 1.14 standard deviations ($p < 0.001$). This represents the largest and most statistically robust effect detected in the model.
- Post-intervention trajectory: No significant change in the toxicity growth rate was observed following the initial drop ($p = 0.624$), indicating the post-event slope paralleled the pre-event trend.

	Coefficient	P-value
const	0.0717	0.567
t	0.0019	0.053
D_t	**-1.1377**	**0.000**
P_t	0.0010	0.624

Table 4.1: Regression estimates. The underlined result represent the sudden and significative drop in the toxicity on the day of Jeffrey Epstein's death.

Visual inspection of Figure 4.3 confirms the statistical results, that **high-profile events produced temporary reductions in toxic language among core users, followed by rapid return to baseline levels through the remainder of 2019, indicating behavioral adaptation without transformative change in underlying engagement patterns.**

4.2.2. Lexical shift results

Figure 4.4.a reveals that both events triggered notable changes in core users' language, since the magnitude and direction of shifts varied by category and event. The most

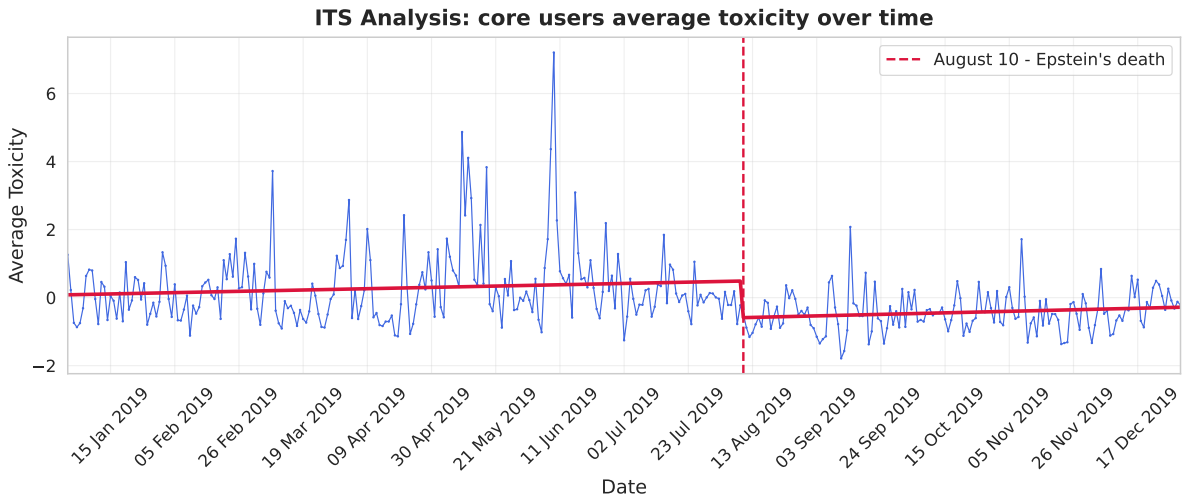


Figure 4.3: ITS analysis of core users' posts toxicity computed by using Detoxify library. The blue line shows the observed mean median toxicity per day, and the red line shows the fixed trend given by an OLS regression. The vertical line at t_0 marks Epstein's death, with the model clearly depicting the immediate drop in toxicity followed by a resumption of the gradual upward trend

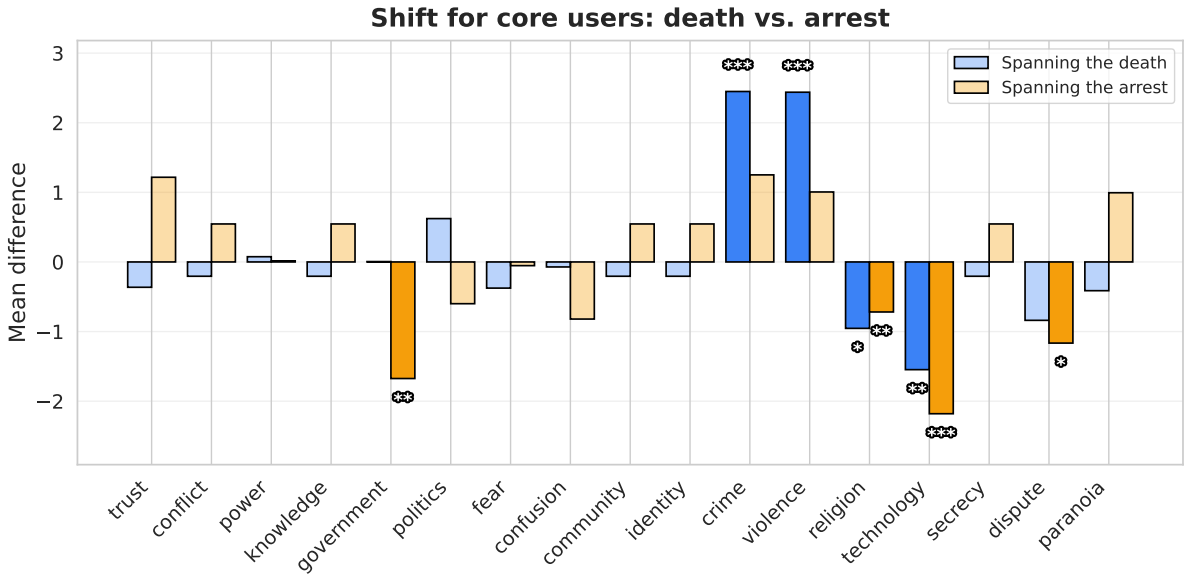
substantial increases associated with the death include crime and violence, suggesting that core users reframed their discourse toward themes of criminality and conflict following this event, reducing the emphasis on topics such as religion in the post-death period.

The arrest period (orange bars) reveals statistically significant negative shifts in government, religion, and technology categories, suggesting a narrowing of thematic focus following the event.

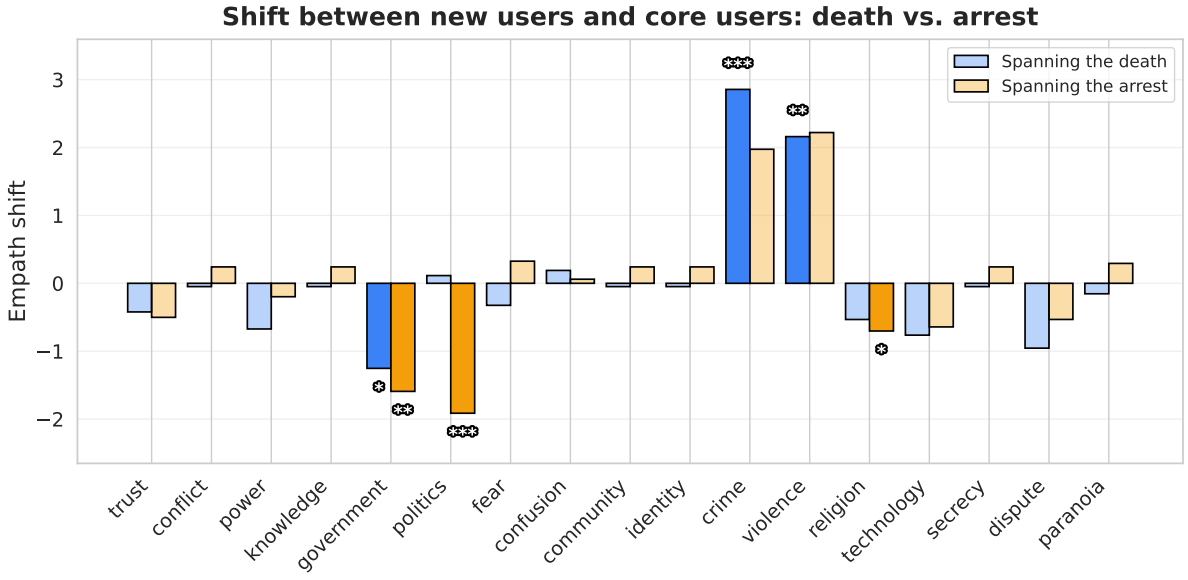
Regarding the comparison between new users and core users, we can see in Figure 4.4.b that in the arrest period (orange bars), new users demonstrated significative divergence from core user baselines in government and politics categories, with a more modest negative difference in religion, suggesting that newcomers arriving during the arrest were less focused on institutional and legal dimensions of the case and more on alternative interpretative frameworks.

The heightened focus on crime and violence aligns with the general suspicion that there was around on Epstein's death [10–12, 32]. So, it is unsurprising that users joining *r/conspiracy* after Epstein's death emphasized criminological and violent dimensions over institutional analysis.

Overall, these patterns indicate that both events prompted distinct linguistic adaptations among users: core users displayed internal reorientation toward



(a)



(b)

Figure 4.4: (a) Shift in Empath category of core users' posts spanning the arrest and the death of Jeffrey Epstein. (b) Shift in Empath category between core and new users' posts in the same period. Asterisks denote significance levels (***: pval $<$ 0.001; **: pval $<$ 0.01; *: pval $<$ 0.05). *r/conspiracy* in the same period.

conflictual narratives, while newcomers introduced alternative framings that further diversified the discourse.

4.2.3. Binary classification

Binary classifiers distinguishing core users from newcomers (methodology detailed in Section 3.6) were evaluated across ten sequential temporal folds from mid-August through December 2019.

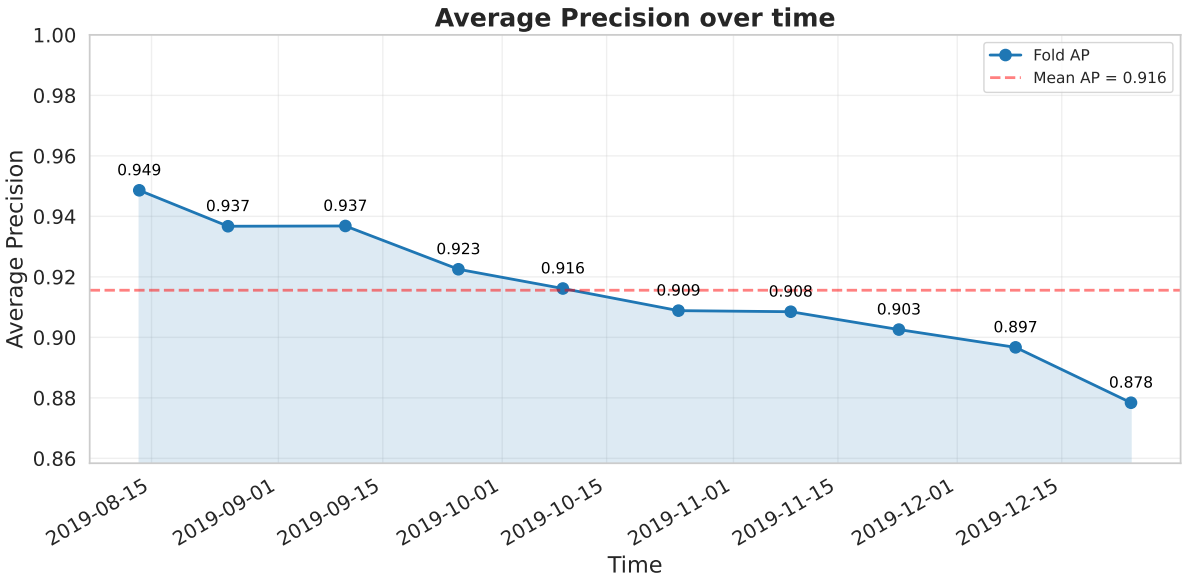


Figure 4.5: Average Precision scores across the ten temporal folds over time of the binary classification of core users by using the Empath scores of their posts.

Figure 4.5 reveals several notable patterns in classifier performance over time. During August and September 2019, the period immediately following Epstein's death when *r/conspiracy* experienced peak visibility and user influx, the model achieved its highest performance, suggesting that core and new users exhibited particularly distinct linguistic profiles during the period of heightened community attention.

However, performance declined progressively from October through December 2019, likely reflecting the gradual linguistic convergence between core and new users as newcomers acclimated to community norms and discourse patterns. As the initial spike of attention subsided and new users became integrated into the community, their linguistic distinctiveness from core users diminished, making classification increasingly difficult. Despite this, the overall mean AP of 0.916 demonstrates that Empath-derived thematic features, when augmented with temporal dynamics, provide substantial discriminative power for distinguishing core from new users.

This finding validates our lexical shift analyses: the thematic differences we identified between groups are not merely statistically significant but also prac-

tically meaningful enough to support accurate automated classification.

4.3. User retention patterns

Kaplan-Meier survival analysis (detailed in Section 3.7) compared retention trajectories across the four user cohorts defined in Section 3.3.

The resulting survival curves, show that both arrest-era cohorts consistently maintained higher survival probabilities than both death-era cohorts across all measured time points. Within each era, users whose first post addressed general conspiracy topics showed marginally better retention than those whose first post engaged Epstein content.

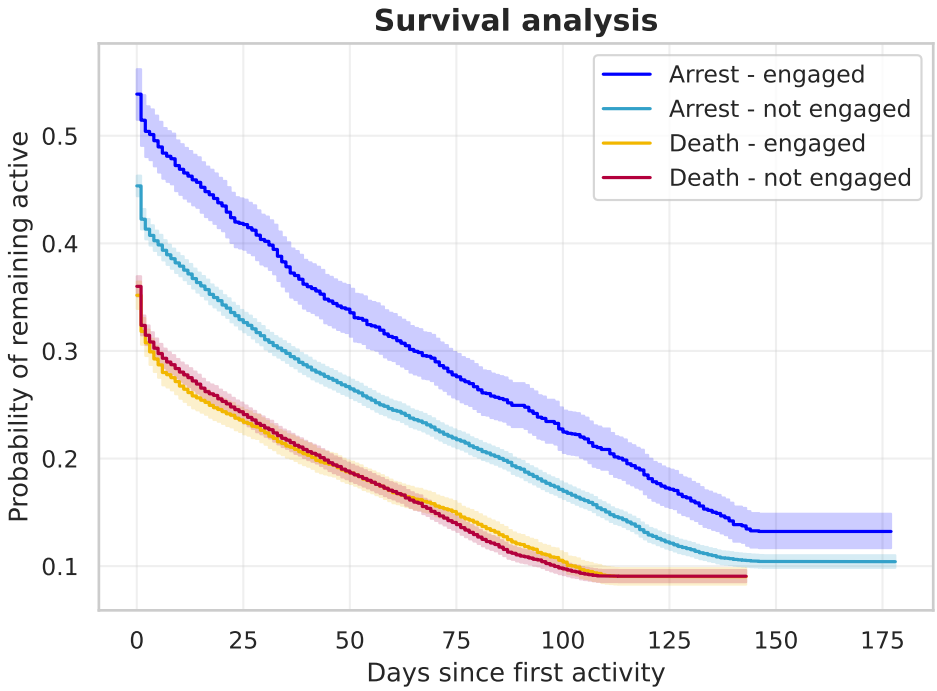


Figure 4.6: Survival analysis performed on 4 cohorts of new users by using a Kaplan-Meier estimator.

The Integrated Brier Score analysis (4.2) suggest that the survival models perform reasonably well, particularly for the post-death cohorts, which exhibit lower scores and hence more stable survival estimates. The slightly higher scores for the post-arrest cohorts may reflect greater variability in user behavior following these events.

Cohort	Integrated Brier score
Arrest – Engaged	0.18252
Arrest – Not engaged	0.15542
Death – Engaged	0.13160
Death – Not engaged	0.13101

Table 4.2: IBS results for the four users cohort estimations.

4.4. Semantic distance trajectories

SBERT embedding analysis (Section 3.8.2) measured semantic alignment between newcomer posts and core community discourse over time. Weekly semantic distance values were computed for each cohort and smoothed using a three-sample rolling mean to show the temporal evolution of semantic distance from July through December 2019, for all four new user groups categorized by two dimensions: the timing of their entry (after Epstein's arrest versus after his death) and the topical focus of their first post (Epstein-related versus non-Epstein content).

To provide transparency regarding the smoothing procedure and to allow readers to assess the degree of variability in the raw data, we also overlay the actual mean for each cohort at each week as a discrete point on the plot. This dual representation allows readers to simultaneously appreciate both the overall trend (via the smoothed curve) and the week-to-week volatility (via the raw points and deviation lines) in semantic alignment.

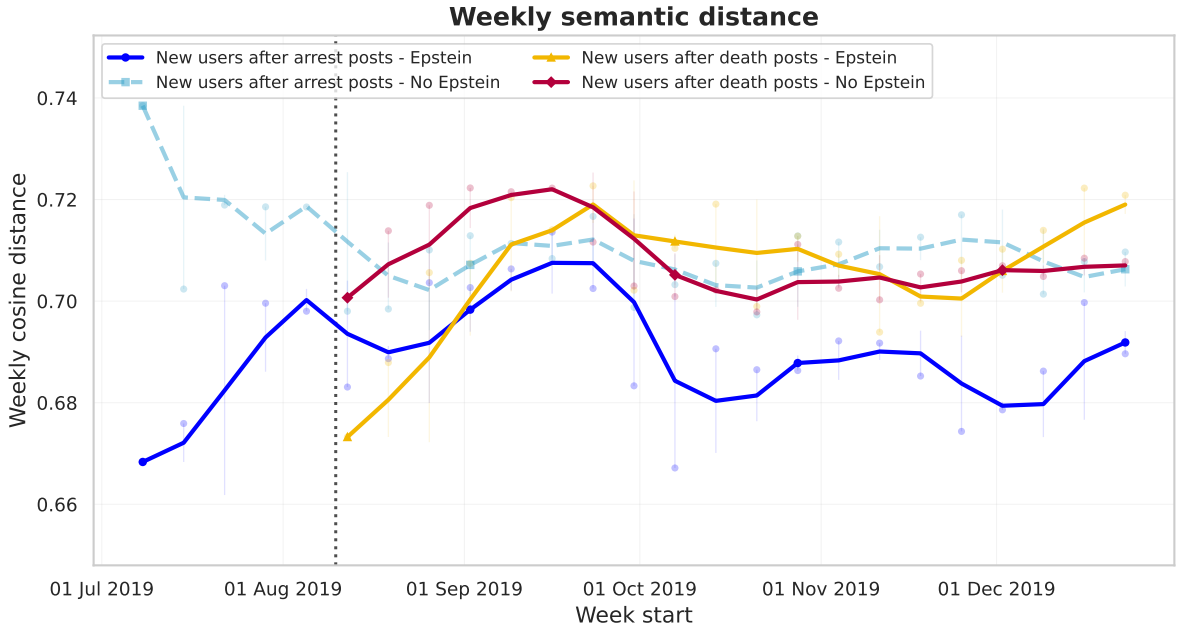


Figure 4.7: Temporal evolution of semantic alignment for new user cohorts, measured as cosine distance from weekly core user language centroids. Solid curves represent three-sample rolling means of weekly average cosine distances; dots indicate unsmoothed weekly means, with vertical lines showing absolute deviation from the smoothed trajectory. The light blue curve is dashed to indicate methodological sensitivity; this cohort's trajectory was inconsistent across alternative analytical approaches. The dotted line marks Epstein's death on August 10.

Distinct and persistent trajectories emerged based on entry timing:

- Arrest-era users (Epstein-focused first posts): This cohort demonstrated rapid alignment, achieving consistently low semantic distance throughout the observation period.
- Arrest-era users (general conspiracy first posts): This cohort exhibited methodological sensitivity across different analytical approaches (detailed in Section Appendix A) and is shown as a dashed line in Figure 4.7.
- Death-era users (Epstein-focused first posts):

  Initial semantic distance of approximately 0.67 increased to around 0.72 by mid-September, then stabilized at $0.70 - 0.71$. This cohort demonstrated limited linguistic convergence with core users after the initial weeks.
- Death-era users (general conspiracy first posts): Initial semantic distance of approx-

imately 0.70 rose to around 0.72 by mid-September and plateaued at $0.70 - 0.71$.

The gap between arrest-era and death-era cohorts remained stable throughout the observation period rather than converging over time. Both death-era cohorts maintained semantic higher distances than the arrest-era Epstein-focused cohort across the final three months of observation (October-December). Unlike arrest-era users who achieved rapid linguistic alignment, death-era users stabilized at persistently higher distances despite months of potential exposure to community discourse.

In sum, these patterns reveal that users' linguistic adaptation depended strongly on when and how they entered the community: early entrants integrated swiftly into the prevailing discourse, while later ones retained distinct linguistic profiles.

5 | Conclusions

Drawing together the results of the preceding sections, this chapter concludes by synthesizing the behavioral, thematic, and linguistic evidence presented above.

5.1. Preliminary behavioral and thematic trends

To address the research question **[RQ1]**, three complementary analyses examined core user discourse patterns surrounding Epstein's death: interrupted time series analysis of toxicity levels, Empath lexical category analysis of thematic shifts, and binary classification to quantify temporal changes in linguistic distinguishability between core users and newcomers.

Interrupted time series analysis revealed that Epstein's death triggered significant immediate toxicity reduction among core users, followed by gradual return to baseline trajectories with no significant post-event slope change. The temporary toxicity reduction during peak attention periods suggests core users strategically managed the community's image during high-visibility periods rather than experiencing genuine normative change, creating deceptively welcoming environments that may facilitate newcomer recruitment while lowering guards against potentially harmful content.

Meanwhile, the lexical shift analysis revealed divergent thematic reorientations between core and new users. Among core users, the arrest triggered significant negative shifts in government, religion, and technology discourse, revealing a focusing effect where established members narrowed thematic scope to concentrate on case-specific criminological details rather than broader conspiratorial frameworks. In contrast, new users arriving after the death emphasized crime and violence language while decreasing government-related discourse, suggesting attraction to sensational elements over institutional analysis. This heightened focus on crime and violence among post-death cohorts reveals how mainstream visibility redirects community attention from institutional analysis toward individualized, sensationalized narratives. The divergence between core users' criminological focus and newcomers' sensationalist orientation provided the first indication that

homepage visibility altered not merely audience size but compositional character.

Finally, the binary classification model that distinguished core users from newcomers revealed interesting patterns over time. The classifier achieved its highest performance in August 2019, maintaining strong accuracy through September, when the linguistic differences between the two groups were sharpest. As the months progressed through December, accuracy gradually declined as some new users began adopting the community's language patterns. This accuracy curve essentially tracked the integration process: highest distinction right after the death event, then slowly converging as newcomers assimilated. However, the declining accuracy masked crucial differences in integration patterns that became clear only through survival and semantic distance analyses.

5.2. Homepage visibility as a compositional filter

The research question **[RQ2]** has been extensively addressed and resolved through survival analysis. In fact, this analysis examined user retention across arrest and death cohorts, revealing that event timing shaped not only the duration of attention but also the fundamental composition of users. The mechanism underlying this divergence centers on differential platform visibility: after Epstein's death, *r/conspiracy* appeared on Reddit's homepage, while after the arrest it did not. This heightened exposure following the death prompted a short-lived, news-driven influx prone to episodic engagement from casual users attracted by mainstream visibility rather than intrinsic interest in conspiracy discourse. These visibility-driven users exhibited faster early drop-off, consistent with shallow commitment and curiosity-driven participation.

Conversely, the arrest attracted relatively more sustained participants who found the subreddit through intentional searches rather than homepage exposure. These users demonstrated steadier retention, reflecting pre-existing interest and higher engagement thresholds characteristic of organic community discovery. The differential retention patterns thus establish that platform algorithmic visibility mechanisms function as powerful selection filters determining not merely audience size but audience composition and commitment levels. Homepage visibility attracts transient, low-commitment users, while organic discovery filters for users with deeper engagement potential.

The survival findings provide crucial context for interpreting all subsequent linguistic analyses: death-era cohorts were fundamentally different populations from arrest-era cohorts before they even posted their first comment. Homepage visibility did not simply expose *r/conspiracy* to more people; it exposed the community to different types of people with distinct motivations, backgrounds, and participation expectations. This compositional

shift established the foundation for the persistent linguistic divergence that semantic distance analysis would subsequently document.

5.3. Semantic integration in visibility-driven vs. interest-driven users

To address **[RQ3]**, SBERT semantic distance analysis measuring linguistic alignment between new users and core community members tested whether homepage visibility affected cultural integration beyond simple retention duration. The findings confirmed fundamentally different integration trajectories based on event timing, validating the compositional filter hypothesis established by survival analysis.

New users joining after Epstein's arrest quickly integrated linguistically with the core community, maintaining consistently low semantic distance throughout the observation period. This rapid linguistic assimilation indicates that arrest-era users effectively adopted the community's discourse conventions and thematic frameworks from the outset, speaking the same language as established members. The sustained low distance suggests not merely superficial imitation, but genuine alignment with community norms and shared interpretive frameworks.

Conversely, users joining after Epstein's death exhibited limited integration. Both death-era cohorts, those making Epstein-related first posts and those making non-Epstein first posts, initially demonstrated lower semantic distance but subsequently stabilized at persistently higher levels, maintaining a consistent gap from core user language through December 2019. This plateau indicates that death-era cohorts developed or retained distinct linguistic patterns rather than converging with established community discourse, suggesting incomplete linguistic assimilation. Unlike arrest-era users, death-era users never achieved full linguistic integration despite months of potential exposure to community norms.

The divergence in linguistic integration supports the survival analysis findings and help explain how homepage visibility shaped who joined the community. Arrest-era cohorts, arriving through intentional community discovery rather than homepage exposure, brought pre-existing familiarity with conspiracy discourse conventions, facilitating rapid linguistic integration. Research on *r/conspiracy* users demonstrates that individuals who seek out conspiracy communities often exhibit distinct psycholinguistic patterns even before their first engagement [34], suggesting self-selection based on pre-existing orientations. Arrest-era users likely possessed this conspiratorial mindset prior to arrival, enabling immediate

alignment with community discourse.

In contrast, the death brought *r/conspiracy* to Reddit's homepage, drawing mainstream users unfamiliar with specialized conspiracy discourse conventions and lacking the prior psycholinguistic patterns characteristic of conspiracy-engaged individuals. These visibility-driven users maintained their distinct communication patterns rather than fully assimilating, explaining the persistent semantic gap observed through the final observation period. The linguistic analysis thus reveals that homepage visibility does not merely affect retention duration but fundamentally alters the cultural integration process itself.

5.4. Implications

The survival and linguistic analyses converge on a unified interpretation: Reddit's homepage visibility functioned as a critical compositional filter that determined both who joined *r/conspiracy* and how they participated. The arrest period, characterized by organic discovery without homepage exposure, attracted users with pre-existing affinity for conspiracy discourse who exhibited both sustained retention and rapid linguistic integration. The death period, characterized by homepage visibility, attracted mainstream users with shallow commitment and persistent linguistic distinctiveness. This divergence was not incidental but structural, reflecting how platform algorithmic visibility mechanisms select for fundamentally different user populations with distinct engagement trajectories and cultural compatibility.

The complementary findings from toxicity, lexical shift, and classification analyses illuminate specific dimensions of this visibility-driven transformation: temporary toxicity reductions during high-attention periods demonstrate core users' strategic behavioral adaptation to external scrutiny, creating deceptively welcoming environments that facilitate newcomer recruitment; lexical divergence between core users' criminological focus and death-era users' sensationalist orientation reflects the thematic heterogeneity introduced by mainstream visibility; classification performance captures the aggregate assimilation trajectory while masking persistent integration failures among visibility-driven cohorts.

These findings indicate that platform-driven visibility reshapes who arrives and how communities develop. When visibility brings in newcomers who did not join through ordinary community pathways, their commitment tends to be shallow and their participation brief; they remain apart from existing members, which limits organic growth. At the same time, in higher-risk settings, increased visibility also does not translate into lasting radicalization; exposure changes the audience mix without producing lasting radicalization trajectories. Platform designers and content moderators should recognize that visibility

interventions produce not merely quantitative changes in audience size but qualitative transformations in community composition and cultural dynamics. Understanding these compositional filter effects becomes increasingly critical as social media platforms continue evolving their recommendation algorithms and visibility mechanisms.

5.5. Limitations

While this study provides convergent evidence for homepage visibility effects on community composition and integration, several methodological and contextual limitations deserve consideration.

5.5.1. Temporal and event-specific constraints

The analysis focuses mainly on a five-month window surrounding two specific Epstein-related events in 2019. This narrow temporal scope limits generalizability in two important ways. First, the findings may reflect unique characteristics of the Epstein case rather than universal patterns of how conspiracy communities respond to mainstream visibility. The case's combination of elite criminality, political polarization, and preexisting conspiracy narratives created conditions that may not apply to other visibility events. Second, the limited observation period prevents assessment of longer-term integration trajectories. Users classified as linguistically distinct in December 2019 might have eventually converged with core user patterns in subsequent months or years, suggesting that the limited integration observed may represent delayed rather than failed assimilation.

5.5.2. Platform and community specificity

This study looks at just one subreddit on one platform, which naturally raises questions about how well these findings apply elsewhere. Reddit has unique features, such as voting systems, nested comment threads, subreddit-specific moderation, that shape how people behave in ways that might not translate to Twitter, Facebook, or dedicated conspiracy forums. Beyond platform differences, *r/conspiracy* itself is somewhat unusual. The community has established norms, tolerates diverse ideological positions, and generally allows heterodox views. This relatively open environment may have allowed death-era users to maintain their distinct linguistic patterns longer. In more tightly controlled or ideologically uniform conspiracy spaces, newcomers might either assimilate faster or get pushed out more quickly.

5.5.3. Operationalization of user categories

Another limitation can be the way we defined *core users* and *new users*. Core users were identified based on their posting history before the arrest, but this approach can't tell us whether someone was a deeply committed believer or just a casual long-term participant. The analysis also can not account for users who maintained multiple accounts or lurked extensively before posting, potentially misclassifying some experienced community observers as naive newcomers. This limitation may introduce noise into cohort comparisons, though the consistency of findings across multiple analytical methods suggests the primary patterns remain robust.

5.5.4. Language model limitations

Our semantic distance measures depend on sentence embeddings from pre-trained language models, which may not fully capture the specialized discourse of conspiracy communities. These models learned from general corpora, so they might miss specific terminology, coded language, or intertextual references that matter within conspiracy circles. We also focused exclusively on linguistic patterns without looking at behavioral signals like voting, comment timing, or network connections. Someone might use language that seems distinct from the community while still being deeply engaged through other ways that text analysis alone will not catch.

5.5.5. Visibility pathway inference

In addition, we identify homepage visibility as the key difference between arrest and death cohorts, but we can not directly confirm how individual users actually found the community. Our inference that death-era users arrived via the homepage while arrest-era users found it through search or navigation makes theoretical sense, but we can not verify it empirically at the individual level. The progression from arrest to death might have attracted different types of people regardless of visibility mechanisms. Media coverage patterns may have shifted between events in a way that influenced who got interested. Reddit's recommendation algorithms and cross-posting practices also evolved during this period, potentially creating unmeasured influences on user recruitment beyond simple homepage featuring.

5.5.6. Visibility event type

Finally, Epstein's death represents a specific type of visibility event: sudden, dramatic, news-driven. Other paths to homepage visibility might produce different effects. Organic upvoting of quality content or algorithmic promotion based on user interests could all shape community composition differently. The findings probably apply most directly to crisis-driven visibility spikes rather than gradual growth or visibility earned through community-generated content quality.

5.6. Future work

The limitations described above suggest several productive directions for future work.

Extended observation periods tracking integration over years rather than months would provide critical insight into whether the linguistic distinctiveness observed among death-era cohorts represents delayed assimilation or permanent separation. Following these users into 2020 and beyond would reveal whether the persistent semantic gap observed through December 2019 eventually closed or remained stable, distinguishing between slow integration processes and fundamental cultural incompatibility. This longitudinal approach would also capture whether arrest-era users maintained their linguistic alignment over extended periods or eventually diverged as the community evolved.

Comparative analysis across multiple platforms and communities would test the generalizability of the compositional filter effect beyond Reddit's specific affordances and *r/conspiracy*'s particular norms. Examining platforms with different architectural features would illuminate which aspects of the observed patterns derive from general visibility mechanisms versus platform-specific features. Similarly, studying diverse community types would reveal whether homepage visibility produces similar compositional effects across different social contexts or whether conspiracy communities respond uniquely to mainstream attention.

More rigorous causal inference would benefit from controlled experimental designs that directly manipulate visibility pathways. Partnerships with platforms could enable randomized assignment of homepage featuring, measuring downstream effects on user composition, retention, and integration while controlling for confounding factors like media coverage and content characteristics. Such experiments would eliminate the inference uncertainty inherent in observational studies, definitively establishing whether homepage visibility causes the compositional differences observed or whether other correlated factors drive the patterns.

Future research should also incorporate non-linguistic engagement signals to construct multidimensional profiles of user integration. Analyzing voting patterns, commenting frequency, network position, and temporal activity patterns alongside linguistic measures would capture users who demonstrate deep engagement through behavioral dimensions despite maintaining linguistic distinctiveness, suggesting partial rather than failed integration. This multidimensional approach would provide a more nuanced understanding of how newcomers participate in online communities beyond discourse alignment alone.

Implementing browser plugins or partnering with Reddit to track actual discovery pathways would eliminate inference uncertainty about how users found communities. Direct measurement of whether users arrived via homepage, search, cross-posts, or external links would separate homepage visibility effects from correlated factors like changing media coverage patterns or evolving recommendation algorithms. This individual-level tracking would enable more precise estimation of compositional filter effects by measuring the causal pathway rather than inferring it from temporal patterns.

The specificity of crisis-driven visibility in this study raises questions about whether alternative visibility mechanisms produce similar effects. Investigating gradual organic growth, algorithmic recommendation, and quality-based promotion would reveal whether the compositional filter effect specifically characterizes dramatic news events or applies more broadly to any form of increased visibility. Comparing multiple visibility pathways within single communities would isolate mechanism-specific effects, determining whether the selection process differs fundamentally across pathways or whether all forms of increased visibility attract less committed users.

Finally, examining how moderation strategies interact with visibility-driven compositional changes could help platforms balance openness with cultural coherence, designing moderation policies that facilitate healthy integration rather than merely controlling harmful content.

5.7. Final remarks

Despite these constraints, the convergent evidence from multiple analytical approaches strongly supports the core finding: platform visibility mechanisms fundamentally shape online community composition and cultural dynamics. The death of Jeffrey Epstein created a natural experiment demonstrating how algorithmic visibility operates not as a simple amplifier but as a sophisticated selection mechanism that determines who joins conspiracy communities, how long they stay, and whether they linguistically integrate with established discourse patterns.

Understanding these dynamics is essential for designing platform governance strategies that balance freedom of expression with mitigation of harmful content spread, recognizing that the pathways through which users discover communities may matter as much as the content they encounter upon arrival.

Bibliography

[1] Integrated Brier Score, sksurv.metrics.integrated_brier_score — scikit-survival 0.25.0, . URL https://scikit-survival.readthedocs.io/en/stable/api/generated/sksurv.metrics.integrated_brier_score.html.

[2] Introduction to Survival Analysis with scikit-survival — scikit-survival 0.25.0, . URL https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html.

[3] Key Facts to Understanding Reddit's Recent API Updates - Upvoted. https://redditinc.com/blog/apifacts, . URL https://redditinc.com/blog/apifacts.

[4] polars: Blazingly fast DataFrame library. https://www.pola.rs/, . URL https://www.pola.rs/.

[5] Pushshift Reddit API v4.0 Documentation — Pushshift 4.0 documentation., . URL https://reddit-api.readthedocs.io/en/latest/.

[6] RandomForestClassifier — scikit-learn 1.7.2 documentation, . URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[7] RandomUnderSampler — Version 0.14.0, . URL https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html.

[8] SentenceTransformers Documentation — Sentence Transformers documentation. https://sbert.net/, . URL https://sbert.net/.

[9] Subreddit r/conspiracy, www.reddit.com/r/conspiracy/, Jan. 2008. URL https://www.reddit.com/r/conspiracy/.

[10] Americans Say Murder More Likely Than Suicide in Epstein Case. Rasmussen Reports®, Aug. 2019. URL https://web.archive.org/web/20200519051748/https://www.rasmussenreports.com/public_content/politics/general_

politics/august_2019/americans_say_murder_more_likely_than_suicide_in_epstein_case.

[11] Most Now Think Jeffrey Epstein Was Murdered. Rasmussen Reports®, May 2020. URL https://web.archive.org/web/20200519050229/https://www.rasmussenreports.com/public_content/lifestyle/people/january_2020/most_now_think_jeffrey_epstein_was_murdered.

[12] Epstein didn't kill himself. Wikipedia, Oct. 2025. URL https://en.wikipedia.org/w/index.php?title=Epstein_didn%27t_kill_himself&oldid=1319780336. Page Version ID: 1319780336.

[13] Interrupted time series. Wikipedia, Sept. 2025. URL https://en.wikipedia.org/w/index.php?title=Interrupted_time_series&oldid=1312450828. Page Version ID: 1312450828.

[14] Kaplan–Meier estimator. Wikipedia, July 2025. URL https://en.wikipedia.org/w/index.php?title=Kaplan%E2%80%93Meier_estimator&oldid=1298262528. Page Version ID: 1298262528.

[15] Reddit. Wikipedia, Nov. 2025. URL https://en.wikipedia.org/w/index.php?title=Reddit&oldid=1321115996. Page Version ID: 1321115996.

[16] D. I. a. C. G. Ali Watkins. Inmate 76318-054: The Last Days of Jeffrey Epstein. *The New York Times*. URL https://www.nytimes.com/2019/08/17/nyregion/epstein-suicide-death.html.

[17] Andrea Salcedo. What We Know About Jeffrey Epstein's Death. *The New York Times*. URL https://www.nytimes.com/2019/08/15/nyregion/newyorktoday/jeffrey-epstein-suicide.html.

[18] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839, May 2020. ISSN 2334-0770. doi: 10.1609/icwsm.v14i1.7347. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7347.

[19] D. K. Brown, Y. M. M. Ng, M. J. Riedl, and I. Lacasa-Mas. Reddit's Veil of Anonymity: Predictors of engagement and participation in media environments with hostile reputations. *Social Media + Society*, 4(4):2056305118810216, Oct. 2018. ISSN 2056-3051. doi: 10.1177/2056305118810216. URL https://doi.org/10.1177/2056305118810216. Publisher: SAGE Publications Ltd.

[20] J. Chaffin. Epstein's death proves feeding ground for conspiracy theo-

ries. *Financial Times*, Nov. 2019. URL https://www.ft.com/content/8f406516-0c9e-11ea-b2d6-9bf4d1957a67.

[21] M. Crowley. Trump Shares Unfounded Fringe Theory About Epstein and Clintons. *The New York Times*, Aug. 2019. ISSN 0362-4331. URL https://www.nytimes.com/2019/08/10/us/politics/trump-epstein-conspiracy-theories.html.

[22] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318, Rio de Janeiro Brazil, May 2013. ACM. doi: 10.1145/2488388.2488416. URL https://dl.acm.org/doi/10.1145/2488388.2488416.

[23] L. de Wildt and S. Aupers. Participatory conspiracy culture: Believing, doubting and playing with conspiracy theories on Reddit. *Convergence*, 30(1):329–346, Feb. 2024. ISSN 1354-8565. doi: 10.1177/13548565231178914. URL https://doi.org/10.1177/13548565231178914. Publisher: SAGE Publications Ltd.

[24] E. G. Ellis. 'Epstein Didn't Kill Himself' and the Meme-ing of Conspiracy. Wired. *Wired*. ISSN 1059-1028. URL https://www.wired.com/story/epstein-didnt-kill-himself-conspiracy/. Section: tags.

[25] E. Fast. Ejhfast/empath-client. Github. https://github.com/Ejhfast/empath-client, Oct. 2025. URL https://github.com/Ejhfast/empath-client. original-date: 2016-04-16T23:56:42Z.

[26] A. Grucela. 60+ Facts About Reddit [Demographics, Trends & User Statistics]., Sept. 2022. URL https://passport-photo.online/blog/reddit-statistics/. Section: Statistics.

[27] C. Guo and K. Caine. Throwaway Accounts and Moderation on Reddit, Jan. 2025. URL http://arxiv.org/abs/2501.17430. arXiv:2501.17430 [cs].

[28] L. Hanu and t. Unitary. Detoxify. Github. https://github.com/unitaryai/detoxify, Nov. 2020. URL https://github.com/unitaryai/detoxify.

[29] F. Hoffa. The most popular languages on Reddit, analyzed with Snowflake and a Java UDTF., Oct. 2021. URL https://medium.com/data-science/the-most-popular-languages-on-reddit-analyzed-with-snowflake-and-a-java-udtf-4

[30] M. Horta Ribeiro, H. Hosseinmardi, R. West, and D. J. Watts. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus*, 2(3):pgad035,

Mar. 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad035. URL https://doi.org/10.1093/pnasnexus/pgad035.

[31] A. Hyzen and H. V. d. Bulck. Conspiracies, Ideological Entrepreneurs, and Digital Popular Culture | Article | Media and Communication. Sept. 2021. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/4092.

[32] Jacob Shamsian. Almost half of Americans now believe the conspiracy theory that sex offender Jeffrey Epstein was murdered. Business Insider., Nov. 2019. URL https://web.archive.org/web/20200609021458/https://www.businessinsider.de/international/jeffrey-epstein-kill-himself-poll-2019-11/?r=US&IR=T.

[33] James B. Stewart. The Day Jeffrey Epstein Told Me He Had Dirt on Powerful People. *The New York Times*. URL https://www.nytimes.com/2019/08/12/business/jeffrey-epstein-interview.html.

[34] C. Klein, P. Clutton, and A. G. Dunn. Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PLOS ONE*, 14(11):e0225098, Nov. 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0225098. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0225098. Publisher: Public Library of Science.

[35] C. Monti, M. Cinelli, C. Valensise, W. Quattrociocchi, and M. Starnini. Online conspiracy communities are more resilient to deplatforming. *PNAS Nexus*, 2(10): pgad324, Oct. 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad324. URL https://doi.org/10.1093/pnasnexus/pgad324.

[36] S. Phadke, M. Samory, and T. Mitra. Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 770–781, May 2022. ISSN 2334-0770. doi: 10.1609/icwsm.v16i1.19333. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19333.

[37] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2):20563051211019004, Apr. 2021. ISSN 2056-3051. doi: 10.1177/20563051211019004. URL https://doi.org/10.1177/20563051211019004. Publisher: SAGE Publications Ltd.

[38] G. Russo, L. Verginer, M. H. Ribeiro, and G. Casiraghi. Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning. *Proceedings of the International AAAI Conference on Web and Social Media*,

17:742–753, June 2023. ISSN 2334-0770. doi: 10.1609/icwsm.v17i1.22184. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22184.

[39] M. Samory and T. Mitra. Conspiracies Online: User Discussions in a Conspiracy Community Following Dramatic Events. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), June 2018. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v12i1.15039. URL https://ojs.aaai.org/index.php/ICWSM/article/view/15039.

[40] T. Schatto-Eckrodt, L. Clever, and L. Frischlich. The Seed of Doubt: Examining the Role of Alternative Social and News Media for the Birth of a Conspiracy Theory. *Social Science Computer Review*, 42(5):1160–1180, Oct. 2024. ISSN 0894-4393. doi: 10.1177/08944393241246281. URL https://doi.org/10.1177/08944393241246281. Publisher: SAGE Publications Inc.

[41] A. C. A. b. h. f. w. a. t. a. o. . S. then, H. R. a. M. Agency, M. C. M. C. f. . F. Brands, and g. B. W. t. o. . m. y. v. F. f. A. c. s. t. a. o. a. y.-y. S. t. A. Y. channel. Reddit User Age, Gender, And Key Demographics Statistics (2025 Data)., Apr. 2025. URL https://adamconnell.me/reddit-statistics/.

[42] A. Trujillo and S. Cresci. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_donald. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2):526:1–526:28, Nov. 2022. doi: 10.1145/3555639. URL https://dl.acm.org/doi/10.1145/3555639.

[43] J.-W. van Prooijen, J. Ligthart, S. Rosema, and Y. Xu. The entertainment value of conspiracy theories. *British Journal of Psychology*, 113 (1):25–48, 2022. ISSN 2044-8295. doi: 10.1111/bjop.12522. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12522. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12522.

[44] T. Venturini. Online Conspiracy Theories, Digital Platforms and Secondary Orality: Toward a Sociology of Online Monsters. *Theory, Culture & Society*, 39(5):61–80, Sept. 2022. ISSN 0263-2764. doi: 10.1177/02632764211070962. URL https://doi.org/10.1177/02632764211070962. Publisher: SAGE Publications Ltd.

[45] J. Wagner. 'So I think I was fine': Trump defends promoting baseless conspiracy theory about Epstein's death. *The Washington Post*, Aug. 2019. ISSN 0190-8286. URL https://www.washingtonpost.com/politics/so-i-think-i-was-fine-trump-defends-promoting-baseless-conspiracy-theory-about-2019/08/13/75547346-bde2-11e9-9b73-fd3c65ef8f9c_story.html.

[46] J. Zeng, M. S. Schäfer, and T. M. Oliveira. Conspiracy theories in digital environments: Moving the research field forward. *Convergence*, 28(4):929–939, Aug. 2022. ISSN 1354-8565. doi: 10.1177/13548565221117474. URL https://doi.org/10.1177/13548565221117474. Publisher: SAGE Publications Ltd.

A | Appendix A

Here we provide a detailed account of the specific fields retained during the data cleaning process.
For comments, we preserved the following fields:

- **id**: unique identifier;
- **author**: author username;
- **body**: comment text;
- **created_utc**: UTC timestamp of creation;
- **score**: vote score;
- **subreddit**: subreddit name;
- **parent_id**: parent content identifier;
- **link_id**: submission identifier;
- **permalink**: permanent URL.

For submissions we retained:

- **id**: unique identifier;
- **author**: author username;
- **title**: submission title;
- **selftext**: selftext content;
- **created_utc**: UTC timestamp of creation;
- **score**: vote score;
- **subreddit**: subreddit name;
- **permalink**: permanent URL;
- **url**: external link URL;

- **num_comments**: number of comments;
- **is_self**: boolean flag indicating whether the submission was a self-post.

These fields provided sufficient information to conduct temporal, linguistic, and network analyses while discarding metadata irrelevant to our research objectives.

To systematically identify submissions and comments discussing Jeffrey Epstein and Ghislaine Maxwell, a case-insensitive regular expression pattern was applied to filter the corpus. The pattern used was:

```
r"\b(epstein|jeffrey epstein|ghislaine|maxwell)\b"
```

where `\b` denotes word boundaries that ensure matches occur only for complete words rather than substrings within larger terms. Word boundaries match at positions between word characters (letters, digits, and underscores) and non-word characters (spaces, punctuation, or string boundaries), preventing false positives such as matching `epstein` within hypothetical terms like `epsteinian`. The pipe operator `|` implements alternation, matching any of the four specified terms: `epstein`, `jeffrey epstein`, `ghislaine`, or `maxwell`. This pattern was applied to submission titles and text content to identify Epstein-related threads. Subsequently, the `parent_id` field was used to retrieve all comments posted within these identified threads, establishing the complete set of Epstein-related discourse regardless of whether individual comments explicitly mentioned the search terms. This two-stage filtering approach ensured comprehensive capture of both direct mentions and contextual discussions within Epstein-focused conversation threads throughout 2019.

Here we detail the Empath lexical categories employed in our linguistic analysis. We utilized a set of predefined categories relevant to conspiracy discourse and community dynamics:

- trust;
- conflict;
- power;
- knowledge;
- government;

- politics;
- fear;
- confusion;
- community;
- identity;
- crime;
- violence;
- religion;
- technology;
- secrecy;
- and dispute.

Additionally, we created a custom category labeled "paranoia" to capture language patterns specific to conspiratorial thinking and perceived surveillance. This custom category was constructed using the following seed words:

- suspicion;
- anxiety;
- paranoid;
- watching;
- followed;
- surveillance;
- threat;
- danger;
- plotted;
- and conspiracy.

For computing sentence embeddings to measure semantic similarity between user posts, this research employed the `all-MiniLM-L6-v2` pre-trained model from the Sentence Transformers library. This model is based on the MiniLM architecture with 6 transformer layers

and approximately 22 million parameters, representing a compact yet effective variant optimized for semantic similarity tasks. The model maps input sentences and paragraphs to a 384-dimensional dense vector space through mean pooling of token embeddings, with a maximum sequence length of 512 tokens. Originally trained on Natural Language Inference (NLI) and Semantic Textual Similarity (STS) datasets, `all-MiniLM-L6-v2` is specifically designed to capture semantic meaning rather than lexical overlap, making it well-suited for measuring conceptual alignment between user posts in online communities. The model's relatively small size and computational efficiency enabled the generation of embeddings for over $2,3$ million posts while maintaining strong performance on semantic similarity tasks.

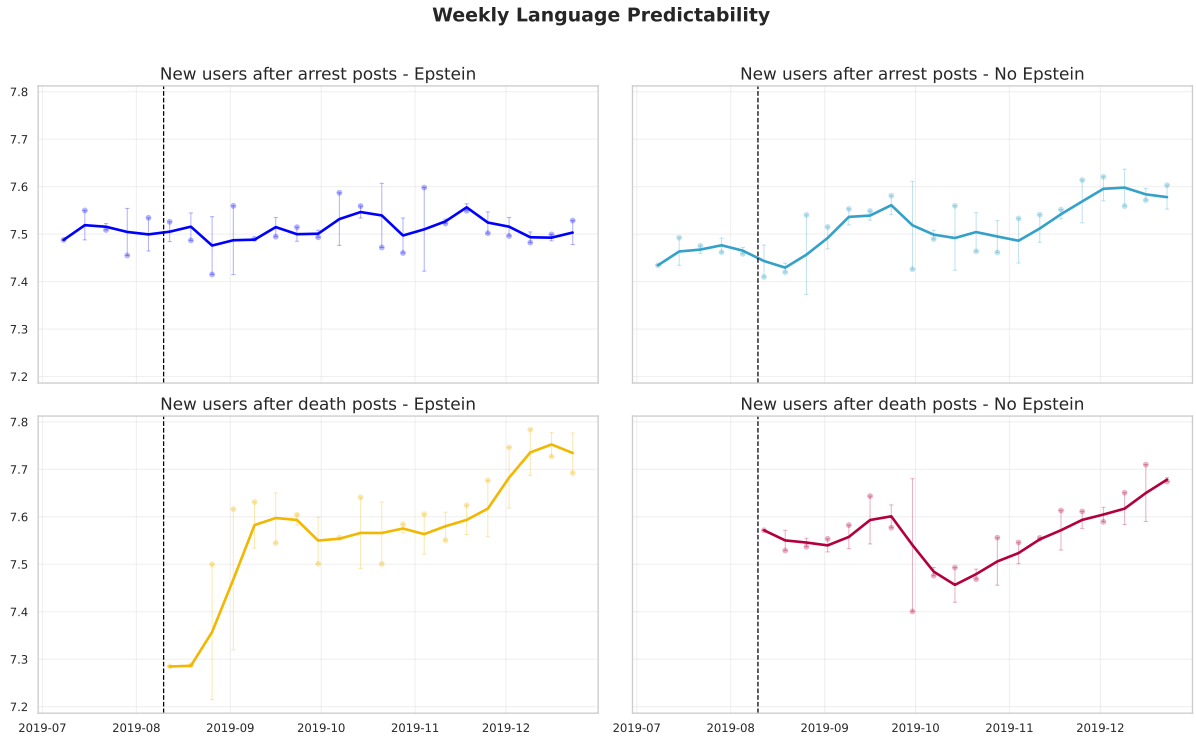


Figure A.1: Temporal evolution of semantic alignment for new user cohorts, measured as cross entropy between the weekly snapshot language model of core users' posts. Solid curves represent three-sample rolling means of weekly average cross entropies; dots indicate unsmoothed weekly means, with vertical lines showing absolute deviation from the smoothed trajectory.

Before conducting the linguistic analysis with SBERT embeddings and cosine distance, we performed a preliminary analysis using bigram probabilities and cross-entropy to quantify semantic alignment between new users and the core community.

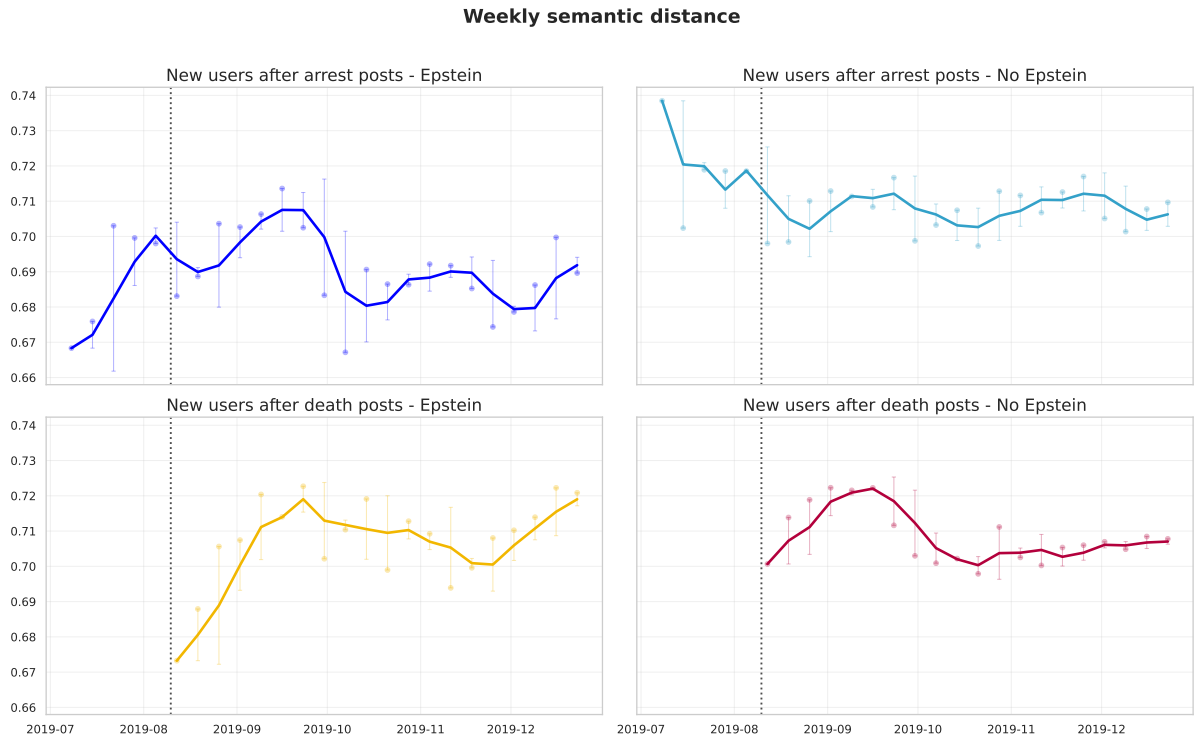


Figure A.2: Temporal evolution of semantic alignment for new user cohorts, measured as cosine distance from weekly core user language centroids. Solid curves represent three-sample rolling means of weekly average cosine distances; dots indicate unsmoothed weekly means, with vertical lines showing absolute deviation from the smoothed trajectory.

This analysis employed the same four new user cohorts described in Section 3.3, allowing for direct comparison with the SBERT-based approach. Rather than computing weekly centroid vectors, we constructed weekly snapshot language models (SLMs) for each week of the observation period. Each SLM was a bigram language model with Katz back-off smoothing, trained on all tokenized posts authored by core users during the corresponding week. These models captured the probabilistic structure of core community language at each point in time.

For each post authored by a new user, we computed a cross-entropy score measuring the post's alignment with the core community's linguistic patterns during the week it was published. First, we tokenized the post and identified the week to which it belonged based on its timestamp. We then extracted all bigrams from the post and computed their probabilities under the corresponding weekly snapshot language model. The cross-entropy of post p was defined as:

$$H(p, SLM_{w(p)}) = -\frac{1}{N}\sum_{i} logP_{SLM_{w(p)}}(b_i)$$

Where $b_1, \ldots, b_N$ are the bigrams comprising p and $P_{SLM_{m(p)}}(b_i)$ denotes the probability of the bigram b_i under the snapshot language model of the post's week $w(p)$. Cross-entropy quantifies the average log-probability of the post's bigrams under the model, with lower values indicating greater alignment with core community language and higher values reflecting linguistic divergence.

As with the SBERT analysis, we aggregated individual post-level cross-entropy scores into weekly cohort-level means and applied three-sample rolling mean smoothing to reduce noise and reveal underlying temporal trends in linguistic alignment.

With both approaches, all cohorts exhibited comparable patterns except for the post-arrest Epstein-posting group; consequently, we exclude this line from consideration in Section 3.8.2. Its lack of robustness across methodological variations suggests that this trajectory may be sensitive to the method chosen and should be interpreted with caution.