PNAS NEXUS

PNAS Nexus, 2025, 4, pgaf056

https://doi.org/10.1093/pnasnexus/pgaf056
Advance access publication 24 February 2025
Research Report

Highly engaging events reveal semantic and temporal compression in online community discourse

Antonio Desiderio [a,b], Anna Mancini [a,b], Giulio Cimini [a,b,†] and Riccardo Di Clemente [c,d,*,†]

[a]Physics Department and INFN, University of Rome Tor Vergata, Via della Ricerca Scientifica, 1, Rome 00133, Italy
[b]Centro Ricerche Enrico Fermi, Via Panisperna, 89a, Rome 00184, Italy
[c]Complex Connections Lab, Network Science Institute, Northeastern University London, 58 St Katharine's Way, London E1W 1LP, United Kingdom
[d]Complex Connections Lab, ISI Foundation, Via Chisola 5, Turin 10126, Italy
*To whom correspondence should be addressed: Email:riccardo.diclemente@nulondon.ac.uk
†These authors jointly supervised the work.
Edited By Derek Abbott

Abstract

People nowadays express their opinions in online spaces, using different forms of interactions such as posting, sharing, and discussing with one another. How do these digital traces change in response to events happening in the real world? We leverage Reddit conversation data, exploiting its community-based structure, to elucidate how offline events influence online user interactions and behavior. Online conversations, as posts and comments, are analyzed along their temporal and semantic dimensions. Conversations tend to become repetitive with a more limited vocabulary, develop at a faster pace, and feature heightened emotions. As the event approaches, the shifts occurring in conversations are reflected in the users' dynamics. Users become more active, and they exchange information with a growing audience, despite using a less rich vocabulary and repetitive messages. The recurring patterns we discovered are persistent across a wide range of events and several contexts, representing a fingerprint of how online dynamics change in response to real-world occurrences.

Keywords: computational social science, online social media, human interactions, digital discourse

Significance Statement

Online social networks are the main platforms where people today engage and exchange viewpoints. Leveraging data from Reddit, we quantify how online community discussions change due to offline events. Comment sequences reveal significant changes in discussion speed, becoming more repetitive in specific word usage while showing more diverse, statistically significant combinations of words. These changes are mirrored in the dynamics of individual users, whose semantic spaces compress as activity frequencies accelerate. Our findings suggest that across various events, increased community content production around offline events links to users' semantic redundancy at high activity frequencies, with effects varying by event type. Consequently, community discussions reflect an accelerating and redundant dynamic, hindering engagement in more meaningful conversations.

Introduction

In today's world of data, the detection of human interactions is increasingly being realized through the continuous stream of signals generated (1, 2) and the knowledge extracted from them can be fed into reliable predictive models (3, 4), continuously refining our portrait of human behavior (5, 6). As human beings, we are social animals living in a community (7–9), and we communicate social issues with others to share our ideas and views (10–12). Communication is a complex phenomenon, shaped by individuals' responses to external stimuli through various channels of communication (13, 14). Nowadays, online social networks represent the most popular means through which humans communicate (15, 16), digest information (17, 18), and engage in discussions about offline events (19). These digital discussions provide an unprecedented amount of data that can lead to a quantitative understanding of how we interact with each other (20) and, in turn help us address major socio-political challenges of our times (21). For instance, by collecting tweets related to climate change conferences (22) we can analyze the discussions and reveal a significant rise in ideological polarization due to the growing presence of right-wing activity.

Offline events such as political elections, championship sport matches, or large-scale epidemic outbreaks are characterized by a mass convergence of online attention and in turn these events

Competing Interest: The authors declare no competing interests.
Received: July 10, 2024. **Accepted:** January 21, 2025

can be used to precisely quantify collective behavioral dynamics (21, 23). The research literature typically characterizes users' attention as the amount of engagement with news. We understood that news propagates and fades away with a stretched-exponential law (24), using the news' popularity index submitted by users on Digg.com, and characterized the burst of activity followed by power-law relaxation using views of new Youtube videos (25, 26). Analyzing user-generated content and real-time communication platforms such as Twitter and Yahoo Research, we have measured that temporal patterns in users' tweet streams changes from the baselines during shocking events (27, 28), such as terrorist attacks, natural disasters or elections. However, human interactions around social issues consist of a continuous dynamical exchange of ideas typical of communities (29), defined as group of individuals who share common interests, characteristics, and interact with one another on a regular basis. How do events affect community discourse, interaction frequency, and writing style? Are these changes, along the temporal and semantic dimensions, specific to the type of event, or are they recurrent across events? Offline events shape the environment in which online conversations take place, thereby changing the direction and altering the dynamics of online discussions, influencing how individuals digest online information (21).

We address these questions using Reddit conversation data. Reddit is a public online forum whose users interact with each other by submitting posts and adding comments to existing posts or comments, thus creating conversation threads (30). The wide thematic spectrum of Reddit conversations enables us to deepen our comprehension of human communication within communities (31): for instance it was shown how users become more intuitive and express their sadness during COVID-19 warning and lockdown phase (32) or how users try to shift the point of view of their interlocutors according to their preconceptions (33). We investigate the temporal and semantic dimensions of online community discourse during highly engaging events (23, 27, 28).

By examining changes in the dynamic structure and content of these discussions, we gain insights into community responses to key events, with implications for how information spreads and how public discourse is shaped in response to such events. By analyzing the time sequence of comments, we can identify variations in discussion activity speed, while the semantic dimension uncovers unique patterns of words and statistically significant expressions within the conversation.

Shifts in community engagement marked by semantic redundancy and increased activity frequency, reflect the intensity and dynamics of collective responses to significant real-world events. Reddit conversations during these events indeed display extreme variations: the frequency of replies increases, conversations develop at a faster pace and are repetitive, the use of word combinations changes, and there is an increase of total emotions shared.

These conversations evolve as users exchange messages around the event, reflecting heightened engagement. Users express their opinions and thoughts on a given event through a comment, that is shared with the community at a specific *time* and with a *semantic* fingerprint.

During events, users begin to increase their activity frequency, while also interacting with more users. High frequency of activity involves a lack of diversified language, and extremely repetitive messages. As users interact with a growing audience, joining the debate and *de facto* broadening the exchange of information. The semantic diversity of a user's conversation peers increases as they occupy a larger semantic space (34, 35), shifting the dialog in practice.

The resulting picture reveals that the increased production of community content around offline events is accompanied by semantic redundancy among users, which emerges alongside high activity frequency. By dissecting communication dynamics in online communities, our approach enhances content moderation efforts to track the evolution of discourse over time, shedding light on how digital platforms function as spaces for collective knowledge-sharing and social cohesion in an increasingly online world.

Results and discussion

The Reddit platform consists of a vast collection of communities, each dedicated to a specific topic (36). Here, we focus on communities with a large user base that discuss US politics (r/politics) and European politics (r/europe), as well as US basketball (r/NBA) and football (r/NFL). Our Reddit dataset comprises over 60 million comments, with a time range spanning from 2020 January 01 to 2021 January 31. This period includes a broad range of events such as the COVID-19 pandemic, the US 2020 elections, NBA interruption, Kobe Bryant death, several NFL matches, etc…(see Supplementary material, Section 1 for the full list of the events considered).

We identified notable events on a weekly basis using a fixed time window, which allowed us to include events spanning multiple days while ensuring reliable statistical analysis of highly engaging events through aggregated signals. Details on the criteria and definitions can be found in the following section and Supplementary material, Section 2.

Burst of activity and conversation characterization

A burst in the overall conversations' volume around an event is the hallmark of its attractiveness (24, 27).

We identify peak weeks of heightened engagement by ranking weekly bursts based on daily z-score variations in posts (see Supplementary material, Section 2 and Methods). To contextualize these peaks, we use nearby events taken from Wikipedia pages (see Supplementary material, Section 1 and Table S2 for the pages retrieved). We note that certain events—such as COVID-19 in the United States and the Capitol Hill incident in Europe—appear with shifts in time across geographic areas, likely due to delays in the public's response timing. We trace more than 20 highly engaging weeks in the chosen Reddit communities, as documented on the Wikipedia pages. On the contrary, we did not considered in our analysis events such as the first impeachment trial of US president Donald Trump (2020 January 16), and Bulgarian protests in July 2020—listed on the Wikipedia pages—since they received limited engagement from the respective communities on Reddit. Figure 1A shows the burst of activity within Reddit political communities, in terms of overall number of daily posts and comments generated around highly engaging events. In general, volumes of both posts and comments increase during the event, with some noteworthy exceptions (e.g. COVID-19 for the US politics community where comments grow much more than posts). To cross-check the events selected, we have integrated into our analysis Google Trends data (using *nba* and *nfl* as query terms for r/NBA and r/NFL, respectively). Figure 1B shows how the time series for the number of posts of the sport communities and the Google Trends are strongly correlated (NBA Correlation 0.7, NFL Correlation 0.8), and the peaks mostly coincide, meaning that people search for events (Google Trends) as they talk more about them (Reddit). For these communities, we observe events that

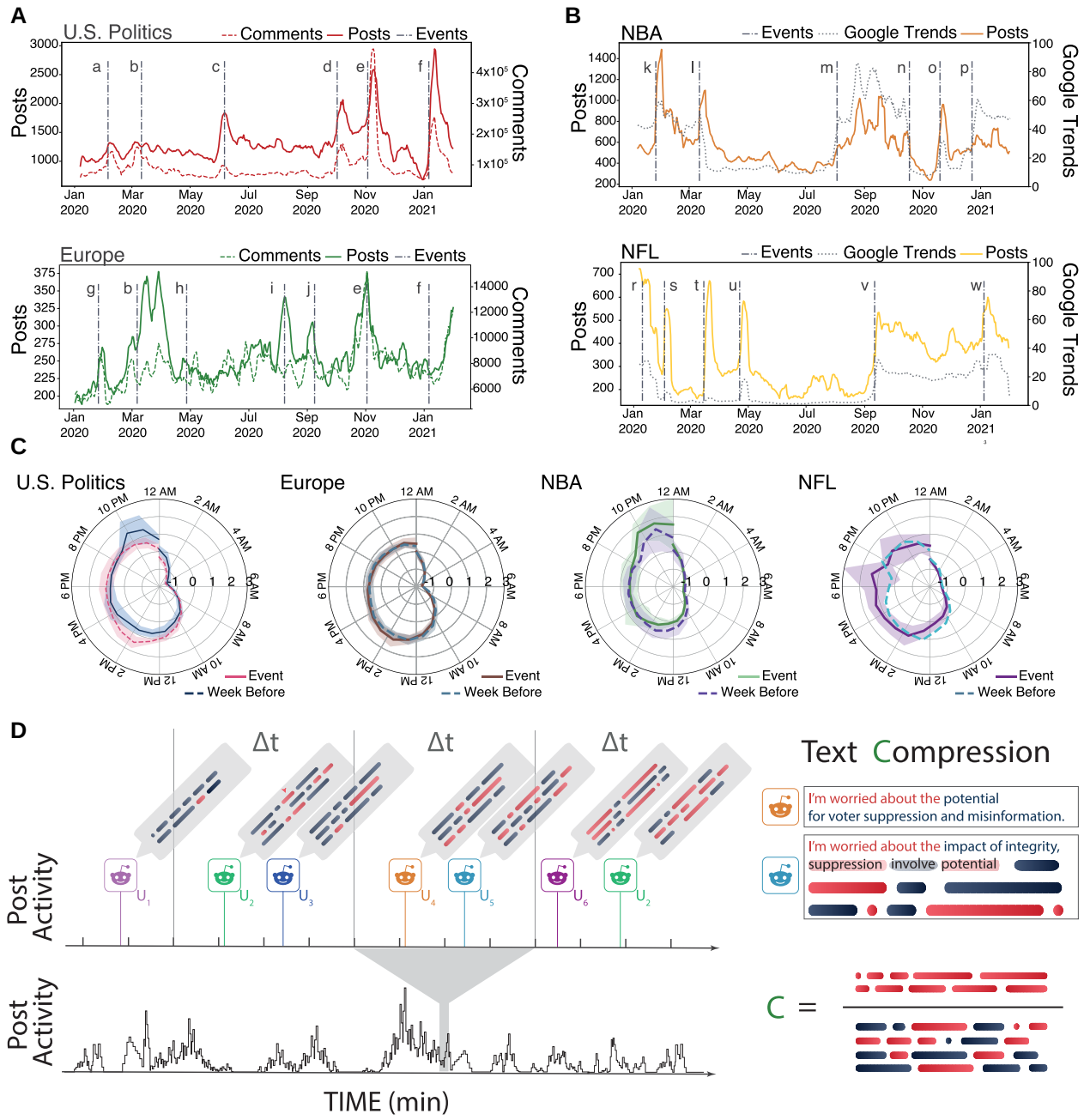

Fig. 1. Burst of activity and conversation characterization In subplots A and B, we apply a 7-day moving average to the time series. A) Number of posts (solid line) and comments (dashed lines) for the US politics community (upper panel) and European community (lower panel). The vertical dashed-dotted lines mark the highly engaging events and correspond to the peaks of the signals. B) Number of posts compared to the Google Trends for the NBA community (upper panel) and NFL community (lower panel). C) Radar plots showing, for each subreddit, the average Z-scored hourly activity in the week before (dashed line) and that of the event (solid line), with the shaded area representing the SD. For the European community, the Amsterdam timezone is selected, while the US/Eastern timezone is employed for all other communities. D) Schematic representation of how we characterize a conversation. For each post, we capture the temporal dimension as the time series extracted by counting the comments underneath within a $\Delta t = 5$ min time interval, and the semantic dimension by merging all the comments into a single text, whose compression is obtained as the ratio between the number of unique patterns of words and of all words (unique and repeated). Icons have been designed by Freepik. The schematic representation of a post have been created by the authors.

span several weeks, coinciding with weekly matches, and in this case, we consider the start and end of this period as an event.

To gain a more comprehensive understanding of the interplay between external events and the communication patterns within the Reddit communities, a common approach is to explore the users' behavior around the observed peaks (27, 28). Figure 1C displays the Z-scored hourly activity of the week before and that of the event. We observe that the digital circadian rhythm of content production is different around specific hours, most likely modified by the events (27). For instance, in sports-related events, the difference is most noticeable around the match kickoff and the end of the game. In the EU politics case, we do not observe a relevant gap, while in the US politics case there is a marked shift in the evening.

Given that offline events influence users' online activity (27, 37–39), how do these events change the way people communicate with each other? On Reddit, users discuss about specific topics

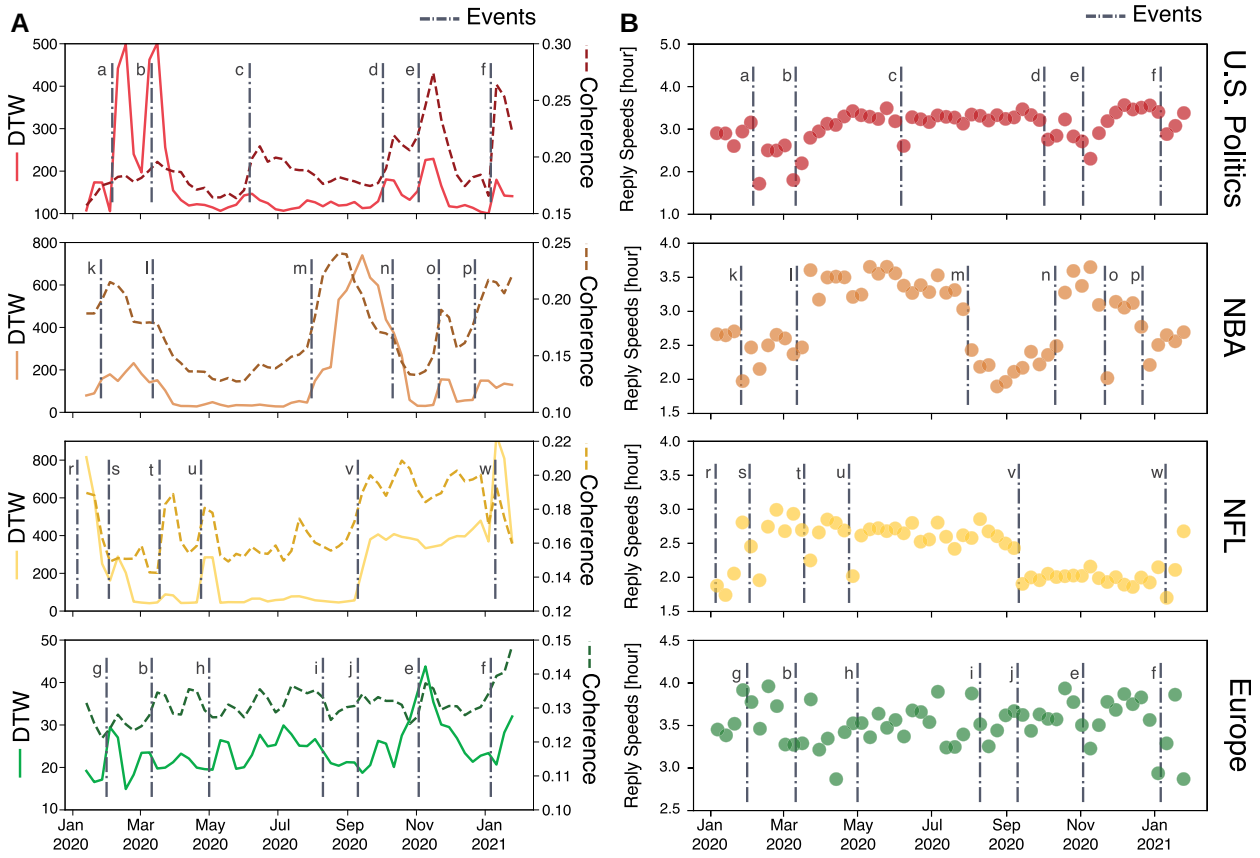

Fig. 2. Temporal dimension. A) Average DTW (solid) and coherence (dashed) distances between the conversations of a week and of the previous one, for each subreddit. B) Average reply speeds, for each subreddit. In all panels, the gray vertical dashed-dotted lines mark the events.

or events by writing posts and commenting to other posts or comments. Hence, we can consider a post and all its underneath comments as a single conversation. Figure 1D illustrates how Reddit posts are depicted and compared along the temporal and semantic dimensions. From each post, we extract a time series by considering time intervals of length Δt, starting from the creation of the post, and counting how many comments are written within each of these intervals. We also extract a text, or document, for each post by joining all the comments underneath and compute its compression as the fraction of unique words to the total number of words in the document.

Accelerating online conversational pace during offline events

We measure dynamic time warping (DTW) and coherence distance between conversations of 1 week and the week before to capture the temporal shift of conversation dynamics (see Methods). The aim of DTW (40) is to find the optimal alignment between two time series by warping one of them in a nonlinear way. This alignment process captures stretching and compression between the two series. When the DTW distance increases, the two series become less similar (temporal mismatch/distortions), as the alignment process requires more warping or compressing of the sequences to work properly. Coherence, instead, allows to measure possible enhancing time series' relationships between weeks by computing the frequency spectra, and detecting common frequency patterns (41, 42). The purpose of coherence is to measure the degree of linear synchronization between time series, providing insights into how much two time series are correlated at different frequencies and it indicates how well the phases align at different frequencies (consistency of their temporal shifts). Conversely, low coherence values point to inconsistent or random temporal shifts (see Methods).

Coherence increases significantly during event weeks compared to baseline periods (weeks without events). This increase reflects a structural shift in conversations, which can either enhance (constructive) or disrupt (destructive) their dynamics.

Figure 2A illustrates the average distances of DTW and coherence across events. For most of the analyzed events, significant changes are evident, with an average variation exceeding 39%. However, variations differ across events and communities. For instance, sharp spikes are observed during COVID-19 discussions in US politics and the NFL Playoffs in January 2021. In sports-related cases, large variations occur at the start of tournaments, but the average distances stabilize once the tournament progresses. Notably, events like the NFL and NBA show identical variations, both exceeding 80%. Meanwhile, the European community displays a marked variation of 35% only for the US 2020 election.

Overall, coherence and DTW variations display consistent patterns, indicating shifts in conversation dynamics. However, coherence values remains consistently lower (below 0.35) than DTW, suggesting that while events affect conversation flow, shifts are not entirely synchronized in frequency and phase. This implies that DTW detects more immediate and intense shifts, while coherence highlights more gradual, frequency-based alignment in conversational patterns. Since DTW is sensitive to time distortion and different speeds, we validate the results by testing

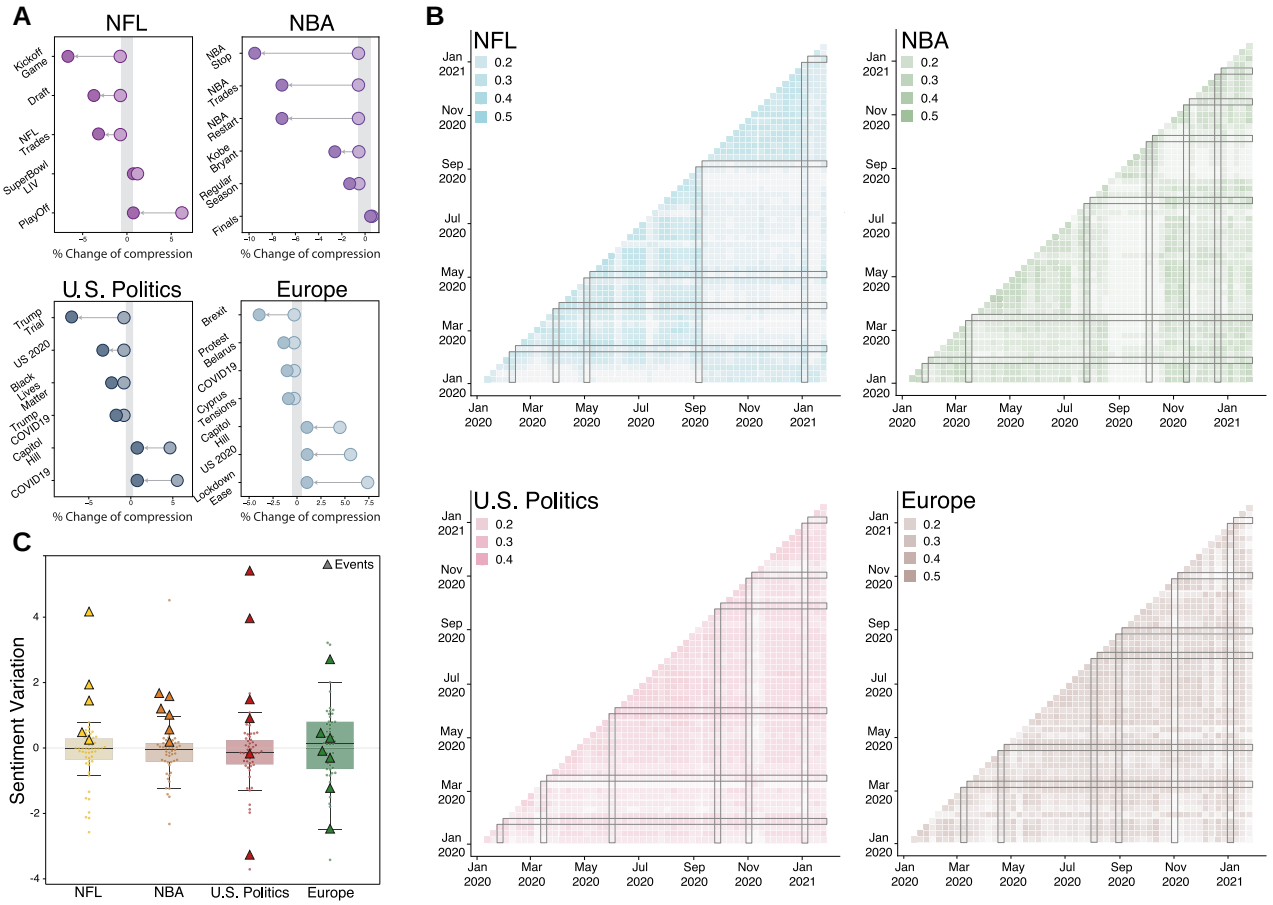

Fig. 3. Semantic dimension. A) Percentage change of compression between the week associated with the event (darker) and the week before (lighter) for each subreddit. The gray shaded vertical area is the SD of the mean change between 1 week and the preceding week. B) Jaccard index among statistically relevant bi-grams between all weeks, the lighter the color the more the weeks are dissimilar. Events are marked with gray lines. C) Emotion variation for each subreddit between consecutive weeks. The triangles mark the variation associated to the events.

against null models obtained through randomization of the timestamps of the comments to statistically validate the changes in the way conversations are structured along the temporal dimension (Supplementary material, Section 3).

Another interesting quantity to look at is the reply speed, defined as the temporal distance between a comment and its response. We find that the weekly distributions of reply speeds are well approximated by Log-Normal distributions, in agreement with other analyses of human temporal patterns (43). We observe that there is a decrease of the reply speed during the events (see Fig. 2B), exceeding 30% in the majority of cases: the peaks of the distributions during the events are getting sharper and shifting to lower values (see Supplementary material, Section 4 for the SD of the reply speeds). These variations are not the same for all the events, due to their heterogeneous attractiveness. Furthermore, we observe that the reply speed remains consistently low—approximately at the same level—for events that span several weeks, such as NFL matches or the initial lockdown period in Europe. We can conclude that, during highly engaging events, conversations along the temporal dimension are structured differently and take place with an overall faster pace.

Reduced focus yet increased diversity in conversation

To measure the focus of conversations on a specific topic, we explore the information content of the text associated to each conversation thread. We measure the compression of the conversation using the Lempel–Ziv complexity index, which measures the repetitiveness of the content (see Methods for further details). The idea behind Lempel–Ziv complexity is to measure the amount of information in a sequence or a string of symbols by identifying and encoding repeated patterns. Figure 3A shows the compression's variation between the week of the event and the week before: a negative variation outlines that conversations have become more repetitive.

The gray shaded area represents the SD of the variation in compression, which is generally close to zero, indicating stability over time, with spikes corresponding to highly engaging events. Most of the events are characterized by a large variation in terms of compression level with respect to the preceding period; however, while the discussions about sports become in general more repetitive, the political discussions tend in the opposite direction.

For instance, events like the NBA Finals and NFL Kickoff show a variation in compression lower than −6%, indicating an increase in repetitive content as users converge on specific recurring topics within the week. Conversely, political events such as Capitol Hill and Lockdown Ease demonstrate positive variations, with changes exceeding +3%, suggesting an evolving discourse during these periods.

Compression, however, focuses only on words, while people tend to repeat certain structures such as word sequences or phrases, which can be important for conveying meaning or

establishing a sense of belonging of a user to the community, especially during a particular event. To capture changes in language before and after events, we detect the statistically significant structures within conversations. We generate an ensemble of document realizations for each week by randomizing the order of words, and compute the relevant bi-grams against the ground truth to assess their statistical significance. We limit our analysis to the top bi-grams, and we exploit them to compare the weeks using Jaccard similarity index among bi-grams (see Fig. 3B and Methods).

For most events, regardless of the topic, Jaccard similarity index values remain below 0.3 when comparing event-related weeks with other weeks. This indicates the presence of distinct statistically relevant bi-grams during event weeks. In sports events, match weeks consistently exhibit a Jaccard similarity index above 0.4 when compared to each other, but below 0.25 when compared to the other weeks. This generates distinct clusters of linguistically similar weeks, visible in Fig. 3B as areas with similar Jaccard index values. A comparable pattern emerges during the US 2020 election weeks (October 2020), where Jaccard indices remain above 0.36, reflecting a clustering of linguistically similar weeks. This linguistic consistency aligns with observations from the temporal analysis. We derive the dissimilarity index from the Jaccard indices, which measures the number of weeks where the Jaccard index falls below the median value for a given week (see Supplementary material, Section 5).

We further perform sentiment analysis to provide a more complete understanding of conversation content and of people's perceptions and attitudes towards an event. Sentiment analysis is a standard technique in online social network analysis to capture the polarity of a text (44–46). First, we compute the sentiment of each post and comment using *Valence Aware Dictionary and sEntiment Reasoner* (VADER). Sentiment varies between −1 (negative) and +1 (positive). We binned this interval and compute, for each week, the histogram of post/comment sentiment values. Then, we compute the Z-score of each bin by using the average value and SD of all weeks. The total "emotion" of a week is defined as the sum of Z-scores over all bins (i.e. the area of the denoised histogram); this represents how distant the week is from the baseline and thus captures the sentiment variation that is possibly associated to an event (see Methods). Finally, we compute the variation of the emotion between a week and the week before.

Generally, there is a consistent positive emotion shift of more than 1.2 SD between the week of the event and the preceding week, compared to the variation observed between two consecutive weeks prior to the event (see Fig. 3C). As in the previous results, all emotion changes for the weeks of NBA and NFL matches lie on the upper tail of the distribution. Meanwhile, in the US politics community we find significant variations for the election weeks and the entire Black Lives Matter protest period, while in the EU case during the first COVID-19 lockdown.

Overall, we observe consistent sentiment variations within topics, such as between NFL and NBA, while noting dissimilarity across different categories, particularly between US politics and the NBA. Notably, this result aligns with the compression analysis, suggesting a consistent pattern within topics, where highly engaging events lead to increased linguistic predictability and shared emotional structures in communities. We can conclude that communities express their views and feelings towards the event in a multifaceted manner, with large variations in sentiment and expressions defined by different combinations of words.

User dynamics reveal amplified repetition along with heightened speed

When people engage in a conversation, they exchange comments with one another and within the community, giving rise to a dynamic process of communication. The dynamical changes of conversations as a whole due to the occurrence of a particular event, that we observed in the previous results, also imply the existence of shifts in temporal activity and semantic structure at the level of individual users. Hence, in this section, we focus our analysis on individual behaviors. Events typically involve an higher number of users who are active solely during the event. As a result, the observed conversational shifts in the previous sections—even if validated with null models—may be attributed to these random users (23). Consequently, in this analysis, we consider only recurrent or dutiful users—those who actively and consistently engage in the community over several weeks (see Methods for details). Furthermore, we consider an additional dimension given by the number of conversation peers of each user, that is, how many neighbors she has in the network of social interactions.

We characterize the individual temporal dimension using the frequency of activity, considering both comments and posts contributed by each user, as an indicator of her level of time-based engagement with the community. We find that there is a power-law correlation between the users' activity frequency and the number of users with whom it interacts, the users' degree, regardless of the event ($R^2 > 0.7$, see Fig. 4A).

The relationship is sublinear, with an average exponent of 0.8 across all weeks, indicating that as users increase their activity frequency, they tend to interact with a disproportionately smaller number of peers. However, the exponent increases during events, reflecting broader engagement with more users while still maintaining sublinear growth (See Supplementary material, Section 6 and Table S6). Hence, during an event, users tend to increase their activity frequencies and they engage in conversations with an expanding group of users (see also the distributions shift in Fig. 4A and Supplementary material, Section 6 for more examples). As more users join the discussion surrounding the event, they become more engaged and reach a larger audience.

We analyze user dynamics to track changes in activity frequency and degree, providing a complete picture of their engagement. We use the Wasserstein distance (47) to compare distributions with varying supports and capture dynamic shifts. We consider the two political communities during the shared events (US 2020 election and the Capitol Hill riot), and we observe that there are no changes (Wasserstein distance near zero) across events in the European case, contrary to the US case (Wasserstein distance >0.3), showing that the users' engagement level is not simply related to the community volume production (see Supplementary material, Sections 6 and 7).

We then move to the analysis of the semantic dimension, considering all the posts and comments contributed by a user in a given week. We found that at the conversation level the combinations of words chosen by users to express their feelings changes during the events (Fig. 3). By mapping the text of comments into the Mikolov semantic space (48), where words that share similar contexts in the corpus are located in close proximity, we can capture users' movements in the conversation by measuring their semantic diversity (see Methods). Due to the shorter text data at the individual user level, statistically validated bi-grams can be noisy in capturing semantic diversity. We find that during events, users' peers become more semantically dispersed, and connected at the same time, as shown by the shifts in Fig. 4B.

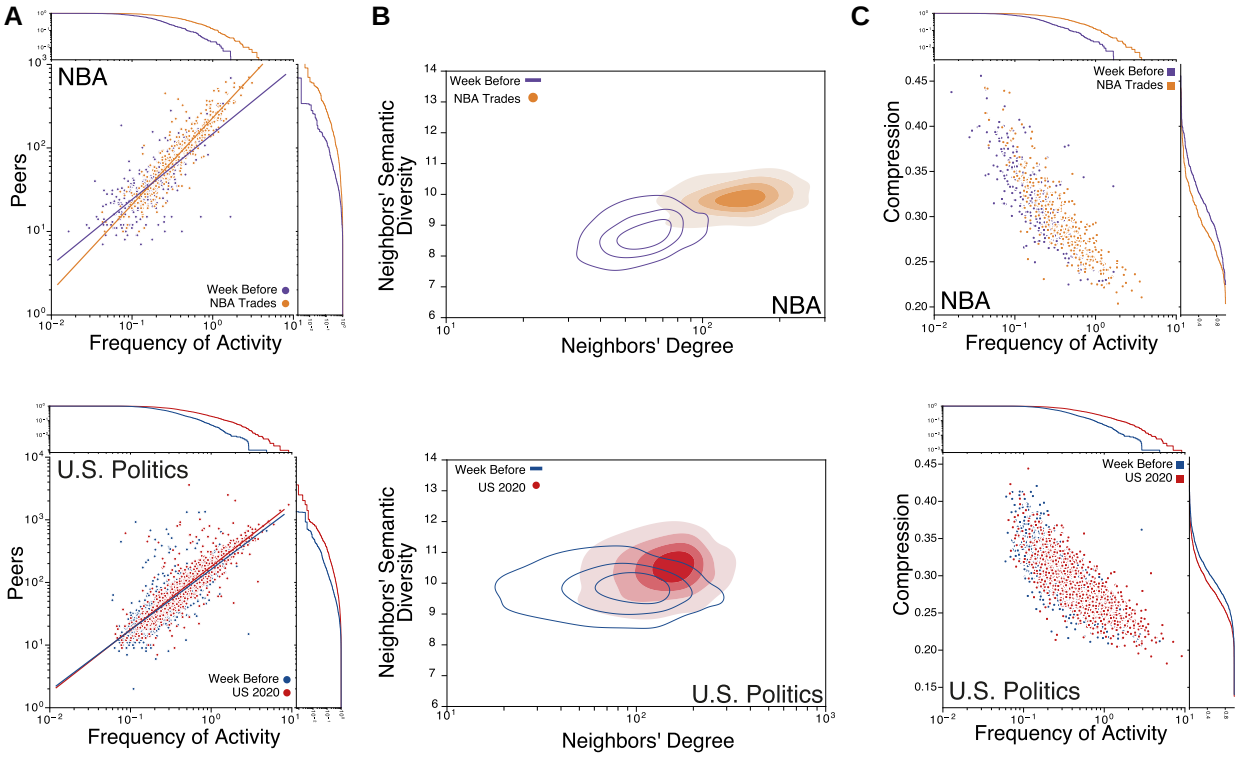

Fig. 4. Users' dynamics. In the following subplots, the data used on the left panels are of the users active on the subreddit r/NBA during the NBA Trades, while on the right of the users on r/politics during the US 2020 election. A) The central panels show the relation between the frequency of activity of each user and the number of interacting peers (the degree). The marginal plots report the survival function of each variable for the 2 weeks. B) The density plots show the variations of the peers' degree and semantic diversity. C) The panels show the relation between user's compression and frequency of activity. Marginal plots report the survival function of each variable for the 2 weeks.

Notably, semantic diversity tends to increase during events, with high values exhibiting further growth while maintaining a stable spread (see Supplementary material, Section 6 and Table S6). Other events can be found in Supplementary material, Sections 6 and 7. Additionally, we find that the average semantic displacement of each post, defined as the average distance in the semantic space between a comment and the succeeding one, tends to increase as the semantic diversity of the user also increases (see Supplementary material, Section 8).

We observe no difference in post displacements between the communities; conversely, NBA and NFL cluster at lower semantic diversity values, US Politics exhibits the highest variability and range, and Europe overlaps with the other political community. In other words, as the users' peers become more semantically dispersed and connected, they are introducing new and varied semantic structures into the conversations. Finally, we notice that as the users' activity frequency increases, their semantic compression also grows (Fig. 4C).

We confirmed that this trend is not solely due to text size growth, as predicted by Heaps' Law, which predicts that larger texts introduce fewer new words (see Supplementary material, Section 9). The observed increase in semantic compression and activity frequency persists across text sizes during highly engaging events, indicating shifts in communication structure. Overall, during events of highly engaging events, users tend to interact with a greater number of peers and the messages they exchange become even more repetitive.

Conclusion

The fingerprints extracted from the digital discussions on Reddit provide evidence of how offline events are perceived by online users. The increased production of online content regarding in-person events is characterized by discussions marked by semantic redundancy, which develop over time at an accelerated pace, regardless of the event type. The observed changes in online social media discussions are reflected in the dynamics of users, where their semantic spaces shrink as activity frequencies rise. By examining the language used by each user's peers, we discover that individuals with broader vocabularies engage more frequently, hence influencing the direction of the conversation.

Our framework for evaluating the impact of offline events on the digital discourse of a community is subject to certain constraints.

First, using Reddit—an online social network where users communicate anonymously, predominantly in English, and which is not a mainstream platform—limits the range of our findings. The development of conversations can be influenced by various factors, including the nature of the topic, the language employed, and the characteristics of the participants (49). To overcome the issue of topic specificity, we focused on various US communities with large user bases, such as politics, NBA, and NFL. Conversely, anonymity may encourage users to express more genuine opinions, as they are less constrained by social consequences (50). Users who engage in specific discussions may not represent the broader population, as their participation is shaped by the unique context of each thread. To overcome this issue when

analyzing user dynamics, we filtered out random users who interacted solely because of the events. Second, we have selected the timeframe of 2020, which was characterized by a massive increase in the usage of online platforms by individuals due to the COVID-19 containment measures (32). Given this limitation, we have explored the timeframe of June 2016 specifically for NBA and US politics, and the findings substantiate our main results (refer to Supplementary material, Section 10). Finally, our results are derived from a single social media platform, Reddit, due to its community structure. Yet our framework, which relies solely on semantics and temporal aspects of online interactions, can be readily applied to other platforms, thereby corroborating our findings.

Our analysis contributes to a deeper understanding of how offline events are discussed by online communities (29). The dissemination of knowledge and the consumption of news (51, 52) are crucial aspects of modern societies (53), fostering social cohesion by providing a shared awareness of nowadays events and promoting exchange of perspectives. With advancements in technology, news receive more collective attention but individual exposure is shortening (37) and individual daily activity is more fragmented (54). Here, we explore the semantic component of online debate, demonstrating that semantic redundancy is always coupled with higher activity frequency. This could result from users repeatedly expressing the same concept at a higher frequency, thus not fostering much deeper conversation.

A certain degree of variability is observed across individual communities, but this is resolved when communities are grouped by topic (e.g. sports and politics). This pattern suggests that the topic under discussion may be useful in further characterizing user behaviors. Studying semantic recurrences over longer time scales can reveal how language and culture change and adapt over time (38), which can have valuable implications for fields such as linguistics (55) and anthropology (39). Furthermore, our framework holds the potential to identify events and remove biases within corpora employed for machine learning pipelines, specifically by identifying and excluding event-related data (56).

Methods

Dataset

We retrieved Reddit conversation data from Pushshift (57), an application programming interface that regularly copies activity data of Reddit and other social media. We queried the service to retrieve information about the chosen subreddits' posts and comments from 2020 January 01 to 2021 January 31. The datasets was cleaned by removing posts/comments made by users with username ending with *bot* and *AutoModerator* (see Supplementary material, Section 1 and Table S3). Google Search engine data were generated by the Google Trends platform and were retrieved via the Python package *pytrends* (see Supplementary material, Section 1 and Table S2 for the keywords used). Highly engaging events are selected according to the daily Z-score variations of the Reddit posting activity, where the mean $\mu_s(t) = \frac{1}{t}\sum_{t'=1}^{t} x_s(t')$ and variance $\sigma_s^2(t) = \frac{1}{t}\sum_{t'=1}^{t} [x_s(t') - \mu_s(t)]^2$ of the time series $x_s(t)$ are used to compute it (58). This quantity captures the variation of engagement of the community in a given week, thereby allowing us to rank weeks according to it. The daily Z-scores variations are reported in Supplementary material, Section 2. The events were contextualized with Wikipedia by manually inspecting the corresponding page of the subreddit and matching with the bursts (see Supplementary material, Section 1 and Table S2 for the pages). The events considered for each community are reported in Supplementary material, Section 1 and Table S1.

Temporal analysis

To compute the hourly activity, we have first counted the comments/posts for each hour within a week, then we have computed the hourly Z-score with respect to the average hourly activity of the overall period. We have extracted a time series from each post by considering time intervals of length Δt, starting from the creation of the post, and counting how many comments are written within each of these intervals. We consider a post lifetime of 24 h and discarded comments written afterwards (¡5% of the total, on average). For each week, we have considered only the top 100 posts by number of comments (accounting for over 50% of comments) and we measured DTW distance (59) and coherence between all the possible combinations of conversations of 1 week and the week before. Coherence has been computed via Welch's method (60) using Hann window, with an overlap of 50% between the two time series (61). If we have two time series, $y(t)$ and $x(t)$, that are linked by a convolution relation and additive white noise $w(t)$ such that $y(t) = H \otimes x(t) + w(t)$, we can compute coherence as follows

$$C_{xy}(\omega) = \frac{|S_{xy}(\omega)|^2}{S_{xx}(\omega)S_{yy}(\omega)} = \left(1 + \frac{S_{ww}}{S_{xx}^2|H|^2}\right)^{-1} = \begin{cases} S_{ww} \gg S_{xx}^2|H|^2 \Rightarrow C_{xy} \sim 0 \\ S_{xx}^2|H|^2 \gg S_{ww} \Rightarrow C_{xy} \sim 1 \end{cases}, \quad (1)$$

where S_{xy} is the cross-spectral density between x and y, and S_{xx} the auto spectral density (same for y). If coherence increases, then the impulse response function H is greater than white noise w; this means that the variability of y can be well explained by the variability of x. DTW is a technique mainly used to find the optimal match between two time series with different lengths by nonlinearly mapping one signal to the other (40). The key idea is to create a matrix M_{ij}, where the entries are the distances between each point i in the signal x(t) and each point j in the other signal y(t). The matrix M_{ij} can be interpreted as the weighted adjacency matrix of a graph, where the point i is connected to the point j with a weight M_{ij}. We can use the Dijkstra's algorithm to find the weighted shortest path through the graph (cumulative distances between each point), which corresponds to the optimal DTW path between the two time series (40). The reply speeds have been computed as the elapsed time between a comment and its response, in this case all the comments have been considered within a week.

Semantic analysis

For each post, we joined all the (lower-cased text of) comments underneath, respecting their temporal order, to obtain a document. For each week, we have considered only the top 100 posts by number of comments, and we have computed the Lempel–Ziv complexity index (62). The algorithm works by scanning a string sequence and identifying repeated patterns or substrings, and then encoding those patterns using a dictionary of previously seen substrings. The number of distinct sequences found is the Lempel–Ziv index (62). In our case, we have removed the substrings of length ¡2 as they are uninformative. Regular signals can be characterized by a small number of patterns and hence have low complexity, while irregular signals are content-rich and therefore less predictable. Lempel–Ziv complexity was introduced to study binary sequences and the ideas introduced were later extended to become the basis of the well-known zip

compression algorithm (63). We have computed compression of a post as the ratio between its Lempel–Ziv complexity index and the total length of the document. To find the significant structures within a document, we have generated an ensemble of 100 documents for each post by randomizing the order of words. We have employed such ensemble as benchmarks to extract the statistically relevant bi-grams for each week by computing the residual occurrence. We have considered only the statistically relevant bi-grams with respect to the average residual (between 30 and 40% of the total bi-grams) and computed the Jaccard similarity index among weeks to assess whether 2 weeks are statistically similar, i.e. they have the same semantic structures. In this case, we have cleaned the text by removing stop-words and punctuation, and considered only the bi-grams with at least 25 occurrences.

Sentiment analysis has been carried out via VADER (64), a Python tool that assigns to each piece of text a score s between −1 (very negative) and +1 (very positive). For each comment/post within a week, we have applied VADER to the text and extracted the associated sentiment. The total emotion of each week has been computed as the total area of the denoised histogram of sentiment, thereby aggregating across bins to potentially account for extreme variations in both positive and negative directions. The denoising of each bin has been carried out by using all the weeks by computing the Z-score, thus revealing weeks with intense sentiment.

Users analysis

For each week, we have reconstructed the network of social interactions by considering posts and comments. Each user who contributed at least five of these posts/comments during that week is represented as a node, and a direct link between user i and j is present if i commented on posts/comments by j. User degree is defined as the number of first neighbors (in both directions) in the network. To frame the changes in the structure of thematic dialogs, we focused on dutiful users that interact persistently with more than 10 posts/comments per week and at least in 70% of the weeks considered. We report the number of users in Supplementary material, Section 1 and Table S4. To compute the activity frequency of each user, we have considered the ordered sequence of comments and posts of the user. The mean temporal distance between two consecutive contributions by the user gives the activity period, whose inverse defines the activity frequency. The semantic compression of each user has been computed via the Lempel–Ziv complexity index, as described in the conversations' analysis but on the document obtained by joining all the comments and posts of the user. To compute the semantic diversity of each user, we have, firstly, trained Word2Vec (48) on all the subreddits, using the Python package *gensim* (65). Word2Vec has been trained using the continuous bag of words model to learn word embeddings. In this neural network model, the goal is to predict a target word given a set of context words, where the target is the middle word of the context. The context words, represented as one-hot encoding vectors, are fed into an embedding layer, which serves as a lookup table for the corresponding word embeddings (dense vectors). The embeddings are then fed into a shallow neural network to predict the probability distribution over the vocabulary for the target word, and the weights are updated using back-propagation; thus refining the word embeddings of the first input layer (embedding layer). In this case, we have cleaned the text by removing punctuation and stop-words, lowering and stemming it. We have considered an embedding vector of 100 dimensions, with word window 3 and we have ignored all words with total occurrence lower than 4. The total number of words on which the model is trained is ∼850 M and we have trained the neural network till the loss reached a plateau (max 100 epochs). We have, then, mapped each comment/post to a point in the semantic space, by averaging the embeddings of the words appearing in a given text. The semantic diversity has been computed as

$$d_u = \sqrt{\frac{1}{N_u}\sum_{i=1}^{N_u} \|v_{i,u} - \langle v\rangle_u\|^2}, \quad (2)$$

where $v_{i,u}$ is the semantic vector of post/comment i by user u and $\langle v\rangle_u$ is the average semantic vector over the possible N_u posts/comments made by user u during the week considered.

Acknowledgments

R.D.C. acknowledges Sony CSL Laboratories in Paris for hosting him during part of the research.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

G.C. acknowledges support from "Deep'N Rec" Progetto di Ricerca di Ateneo of University of Rome Tor Vergata.

Author Contributions

A.D. and A.M. gathered the data. A.D. performed the analysis. A.D. and A.M. realized the figures. G.C. and R.D.C. designed and supervised the analysis. All the authors discussed the results, wrote the article, and approved the final manuscript.

Preprints

A preprint of this article is published at [https://arxiv.org/abs/2306.14735].

Data Availability

Reddit conversation data used in this study can be retrieved from the Reddit or Pushshift API at https://www.reddit.com/r/pushshift/ and were retrieved before October 2022. The code to reproduce the analysis is released on GitHub.

References

1 Sapiezynski P, Stopczynski A, Lassen DD, Lehmann S. 2019. Interaction data from the Copenhagen networks study. *Sci Data*. 6(1):1–10.
2 Yang Y, Pentland A, Moro E. 2023. Identifying latent activity behaviors and lifestyles using mobility data to describe urban dynamics. *EPJ Data Sci*. 12(1):15. doi: 10.1140/epjds/s13688-023-00390-w.
3 Eagle N, Pentland AS, Lazer D. 2009. Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci U S A*. 106(36):15274–15278.
4 Lu X, Bengtsson L, Holme P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci U S A*. 109(29):11576–11581.

5 Lazer D, *et al.* 2009. Social science: computational social science. *Science*. 323(5915):721–723.
6 Lazer DMJ, *et al.* 2020. Computational social science: obstacles and opportunities. *Science*. 369(6507):1060–1062.
7 Mowlana H. 2018. On human communication. *Javnost*. 25(1–2): 226–232.
8 Emery NJ, Clayton NS, Frith CD. 2007. Introduction. Social intelligence: from brain to culture.
9 Littlejohn SW, Foss KA. *Theories of human communication*. Waveland Press, 2008. https://www.waveland.com/browse.php?t=270.
10 Lee NJ, Shah DV, McLeod JM. 2013. Processes of political socialization: a communication mediation approach to youth civic engagement. *Commun Res*. 40(5):669–697.
11 Kahne J, Bowyer B. 2018. The political significance of social media activity and social networks. *Polit Commun*. 35(3):470–493.
12 Jin SV, Ryu E. 2020. "I'll buy what she's #wearing": the roles of envy toward and parasocial interaction with influencers in instagram celebrity-based brand endorsement and social commerce. *J Retailing Consum Serv*. 55:102121.
13 Stevens SS. 1950. Introduction: a definition of communication. *J Acoust Soc Am*. 22(6):689–690.
14 Teresa Anguera M, Izquierdo C. 2012. Methodological approaches in human communication: from complexity of perceived situation to data analysis. *Emerg Commun Stud New Technol Pract Commun*. 9:203–222.
15 Heidemann J, Klier M, Probst F. 2012. Online social networks: a survey of a global phenomenon. *Comput Netw*. 56(18):3866–3878.
16 Segerberg A, Bennett WL. 2011. Social media and the organization of collective action: using twitter to explore the ecologies of two climate change protests. *Commun Rev*. 14(3):197–215.
17 Cinelli M, *et al.* 2020. The COVID-19 social media infodemic. *Sci Rep*. 10(1):1–10.
18 Lazer DMJ, *et al.* 2018. The science of fake news: addressing fake news requires a multidisciplinary effort. *Science*. 359(6380): 1094–1096.
19 Halu A, Zhao K, Baronchelli A, Bianconi G. 2013. Connect and win: the role of social networks in political elections. *Europhys Lett*. 102(1):16002.
20 Omodei E, De Domenico M, Arenas A. 2015. Characterizing interactions in online social networks during exceptional events. *Front Phys*. doi: 10.3389/fphy.2015.00059.
21 Lorenz-Spreen P, Oswald L, Lewandowsky S, Hertwig R. 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat Hum Behav*. 7(1): 74–101.
22 Falkenberg M, *et al.* 2022. Growing polarization around climate change on social media. *Nat Clim Chang*. 12(12):1114–1121.
23 Szell M, Grauwin S, Ratti C. 2014. Contraction of online response to major events. *PLOS ONE*. 9(2):e89052. doi: 10.1371/journal.pone.0089052.
24 Wu F, Huberman BA. 2007. Novelty and collective attention. *Proc Natl Acad Sci U S A*. 104(45):17599–17601.
25 Crane R, Sornette D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proc Natl Acad Sci U S A*. 105(41):15649–15653.
26 Yang J, Leskovec J. 2011. Patterns of temporal variation in online media. In: *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*. p. 177–186. doi: 10.1145/1935826.1935863.
27 Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M, Aihara K. 2013. Correction: quantifying collective attention from tweet stream. *PLoS One*. 8(5):e61823.
28 He X, Lin YR. 2017. Measuring and monitoring collective attention during shocking events. *EPJ Data Sci*. 6(1):1–22.
29 Candia C, Jara-Figueroa C, Rodriguez-Sickert C, Barabási AL, Hidalgo CA. 2019. The universal decay of collective memory and attention. *Nat Hum Behav*. 3(1):82–91.
30 Choi D, *et al.* 2015. Characterizing conversation patterns in reddit: from the perspectives of content properties and user participation behaviors. In: *Proceedings of the 2015 ACM on Conference on Online Social Networks. ACM*. p. 233–243. doi: 10.1145/2817946.2817959.
31 Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. 2021. Studying reddit: a systematic overview of disciplines, approaches, methods, and ethics. *Soc Media Soc*. 7(2):20563051211019004.
32 Ashokkumar A, Pennebaker JW. 2021. Social media conversations reveal large psychological shifts caused by COVID-19's onset across U.S. cities. *Sci Adv*. 7(39):eabg7843.
33 Monti C, Aiello LM, De Francisci Morales G, Bonchi F. 2022. The language of opinion change on social media under the lens of communicative action. *Sci Rep*. 12(1):1–11.
34 González MC, Hidalgo CA, Barabási AL. 2009. Understanding individual human mobility patterns (Nature (2008) 453, (779–782)). *Nature*. 458(7235):238.
35 Lombardo G, Tomaiuolo M, Mordonini M, Codeluppi G, Poggi A. 2022. Mobility in unsupervised word embeddings for knowledge extraction—the scholars' trajectories across research topics. *Future Internet*. 14(1):25.
36 Olson RS, Neal ZP. 2015. Navigating the massive world of reddit: using backbone networks to map user interests in social media. *PeerJ Comput Sci*. 2015(5):e4.
37 Lorenz-Spreen P, Mønsted BM, Hövel P, Lehmann S. 2019. Accelerating dynamics of collective attention. *Nat Commun*. 10(1):1759.
38 Fortier I, Castellanos Juarez M. 2017. How hypermodern and accelerated society is challenging the cultural sector.
39 Malley B, Knight N. 2008. Some cognitive origins of cultural order. *J Cogn Cult*. 8(1–2):49–69.
40 Berndt D, Clifford J. 1994. Using dynamic time warping to find patterns in time series. In: *Workshop on knowledge discovery in databases*. Vol. 398. Seattle, WA, USA. p. 359–370. http://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf.
41 Grinsted A, Moore JC, Jevrejeva S. 2004. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process Geophys*. 11(5/6):561–566.
42 Maharaj EA, D'Urso P. 2010. A coherence-based approach for the pattern recognition of time series. *Physica A Stat Mech Appl*. 389(17):3516–3537.
43 Kaltenbrunner A, *et al.* 2008. Homogeneous temporal activity patterns in a large online communication space. *Int J WWW/INTERNET*. 6:61–76.
44 Box-Steffensmeier JM, Moses L. 2021. Meaningful messaging: sentiment in elite social media communication with the public on the COVID-19 pandemic. *Sci Adv*. 7(29):eabg2898.
45 Bovet A, Morone F, Makse HA. 2018. Validation of twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Sci Rep*. 8(1):8673.
46 Matalon Y, Magdaci O, Almozlino A, Yamin D. 2021. Using sentiment analysis to predict opinion inversion in tweets of political communication. *Sci Rep*. 11(1):1–9.
47 Kantorovich LV. 1960. Mathematical methods of organizing and planning production. *Manage Sci*. 6(4):366–422.
48 Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z,

Weinberger KQ, editors. *Advances in neural information processing systems*. Vol. 26. Curran Associates, Inc., 2013 Oct. p. 1–9. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf http://arxiv.org/abs/1310.4546.

49 Stans SEA, Dalemans RJP, de Witte LP, Smeets HWH, Beurskens AJ. 2017. The role of the physical environment in conversations between people who are communication vulnerable and health-care professionals: a scoping review. *Disabil Rehabil*. 39(25):2594–2605.

50 Bernstein M, *et al.* 2021. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. p. 50–57. doi: 10.1609/icwsm.v5i1.14134.

51 Del Vicario M, *et al.* 2016. The spreading of misinformation online. *Proc Natl Acad Sci U S A*. 113(3):554–559.

52 Gravino P, Prevedello G, Galletti M, Loreto V. 2022. The supply and demand of news during COVID-19 and assessment of questionable sources production. *Nat Hum Behav*. 6(8):1069–1078.

53 Watts DJ, Rothschild DM, Mobius M. 2021. Measuring the news and its impact on democracy. *Proc Natl Acad Sci U S A*. 118(15): e1912443118.

54 Sullivan O, Gershuny J. 2018. Speed-up society? Evidence from the UK 2000 and 2015 time use diary surveys. *Sociology*. 52(1): 20–38.

55 Foster I. 2019. The future of language learning. *Lang Cult Curriculum*. 32(3):261–269.

56 Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In: *Proceedings of the 60th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, Online, 2022 May.

57 Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. 2020. The Pushshift Reddit Dataset. In: *Proceedings of the International AAAI Conference on Web and Social Media*.

58 Mancini A, Desiderio A, Di Clemente R, Cimini G. 2022. Self-induced consensus of reddit users to characterise the GameStop short squeeze. *Sci Rep*. 12(1):13780.

59 Bellman R, Kalaba R. 1958. On adaptive control processes. *IRE Trans Automat Control*. 4(2):1–9.

60 Welch PD. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans Audio Electroacoust*. 15(2):70–73.

61 Virtanen P, *et al.* 2020. Author correction: sciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 17(3): 352 .

62 Lee DS, Kang S. 2014. On the complexity of finite sequences over a finite set. *Far East J Math Sci*. 87(2):133–147.

63 Network Working Group, Deutsch P, Aladdin Enterprises. 1996. DEFLATE Compressed Data Format Specification version 1.3. Technical report.

64 Hutto CJ, Gilbert E. 2014. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*.

65 Rehurek R, Sojka P. 2010. Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.