9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Recurring patterns within Reddit communities during exogenous events

Keywords: group behaviour, text analysis, Reddit, network science, collective patterns

Extended Abstract

In today's world of data, our scientific understanding of human interactions is on the rise. Human beings generate a continuous stream of detectable signals and the knowledge extracted from them could be fed in reliable predictive models, continuously refining our portrait of human behaviour. People actively perceive events of the society in which they live in and express their views, interacting with each other and with the event itself. It is an established fact that online social networks are the nowadays spaces where humans communicate and organise with each other. The fingerprints of these digital discussions help us capture how the external events are perceived and how people interact with each other. We are driven by the event to join the debate, but we communicate with others differently; not only our interactions are faster but also the messages are poor and repetitive. In the context of online social networks it has been measured that temporal patterns in users' tweet streams shift from the baselines during shocking events [1]. These studies are restricted only to the temporal data dimension, users are considered as single entities and the topics considered are too homogeneous. Moreover, human interactions around novel news develop with a continuous dynamical exchange of ideas typical of communities [2]. In this work we explore the whole dimension of discussions, measuring and framing conversation dynamics during special events, in order to understand how a community's interest is driven by the events and how users' behaviours are impacted.

We benchmark changes of human interactions during a set of established events, extracted from Wikipedia and crosschecked with Google Trends, on Reddit, which is a public *community*-based platform whose users chat with each other by submitting new *posts* and adding *comments* to existing posts or comments. The forum is organised into various subreddits, each dedicated to a specific topic; as a result Reddit is a place where people can dive through the lens of their interest. We focus on communities that discuss about American and European politics, NBA and NFL, during the period from January 01, 2020 to January 31, 2021. We describe each conversation using a time series and text (Figure 1A), proposing the *semantic space-time* framework to weight the weeks and quantify the impact of the events. The conversations' temporal dimension quantifies differences in frequency of discussion activity, gauging the shift of the events. We measure the Dynamic Time Warping distance between the time series of one week and the week before: these distances increase during the events. Moreover, we compute coherence, statistically validated by randomizing the comments' timestamps, between two weeks to gauge if there is causal relationship between the time series (Figure 1B). The distributions of answering times, i.e. the elapsed time between a comment and its response, change during the events (kurtosis increases). The semantic dimension captures the fingerprints of the conversations' content. It is defined as the statistically relevant unique patterns of words used in the conversation. First, we measure the repetitiveness of the conversations' words sequence by counting unique patterns using Lempel-Ziv complexity that provides a measure of the conversations' compression [3]. Notably the compression between the week of the event

and the week before is decreasing (Figure 1C), marking that conversations have become repetitive. Compression focuses only on the unique structures in the text, not capturing the possible recurring and therefore significant structures in it. We, then, randomize the words in a text to extract the statistically relevant bigrams (or patterns of Lempel-Ziv), outlining the textual fingerprints. To assess whether two weeks are statistically similar along the patterns dimension, we consider for each week the Top 15% bigrams and compute cosine similarity between a week and the week before. The cosine similarity decreases during the events: even tough we have uniformity of content (compression), the non match of the fingerprints indicates semantic difference due to the event. The research literature characterizes sentiment as a means of assessing the opinions of social media users, thus we compute the overall tone during each week to measure the emotions shared and find that extreme variations are always related with exogenous events.

These semantical-temporal patterns of social activity are general across heterogeneous topics and subjects, showing common dynamics in the shifts of users' behaviors. Chats are collections of messages exchanged by users, which express their views on the event at a given *time* and *semantic-space* position. All the comments sent by each user outline their trajectories, from which we extract the users' temporal activity, semantic compression and diversity to study the single user dynamics. We filter users to have only the ones that interact more and many times within the subreddit, as they are the ones that will readily perceive the event. The temporal dimension is defined as the frequency of activity, the inverse of active time computed from the ordered sequence of comments, to capture changes in the periods of activity. We have outlined that as a user is more active, more users are interacting with her (Figure 2A). High frequency of activity involves a lack of language, and extremely repetitive messages. The users' compression has been computed via Lempel-Ziv complexity as in the conversations' analysis (Figure 2C). We train Word2Vec on all the dataset and associate a semantic vector to each user related to her comments, and compute the semantic diversity to grasp users that are more semantic explorers [4]. During events, users connect with a growing number of users (shift in Figure 2A), joining the debate and de facto broadening the exchange of information. By focusing on the peers of each user, we investigate their semantic diversity and observe that they are the ones filling up more semantic space, shifting the dialogue in practice (Figure 2B). The resulting picture tells us that the increased production of community content surrounding special events is accompanied by users' semantic fragmentation and redundancy, which develops over high frequency of activity.

References

[1] He, Xingsheng and Lin, Yu-Ru, *Measuring and monitoring collective attention during shocking events*, EPJ Data Science **6**, 1-22 (2017).

[2] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L. and Hidalgo, C. A., *The universal decay of collective memory and attention*, Nat. human behaviour 3, 82–91 (2019).

[3] Aboy, M. and Hornero, R. and Abasolo, D. and Alvarez, D., *Interpretation of the Lempel-Ziv Complexity Measure in the Context of Biomedical Signal Analysis*, IEEE Transactions on Biomedical Engineering **53**, 2282-2288 (2006).

[4] Gonzalez, M. C., Hidalgo, C. A. and Barabasi, A.-L., *Understanding individual human mobility patterns*. Nature 453, 779–782 (2008).

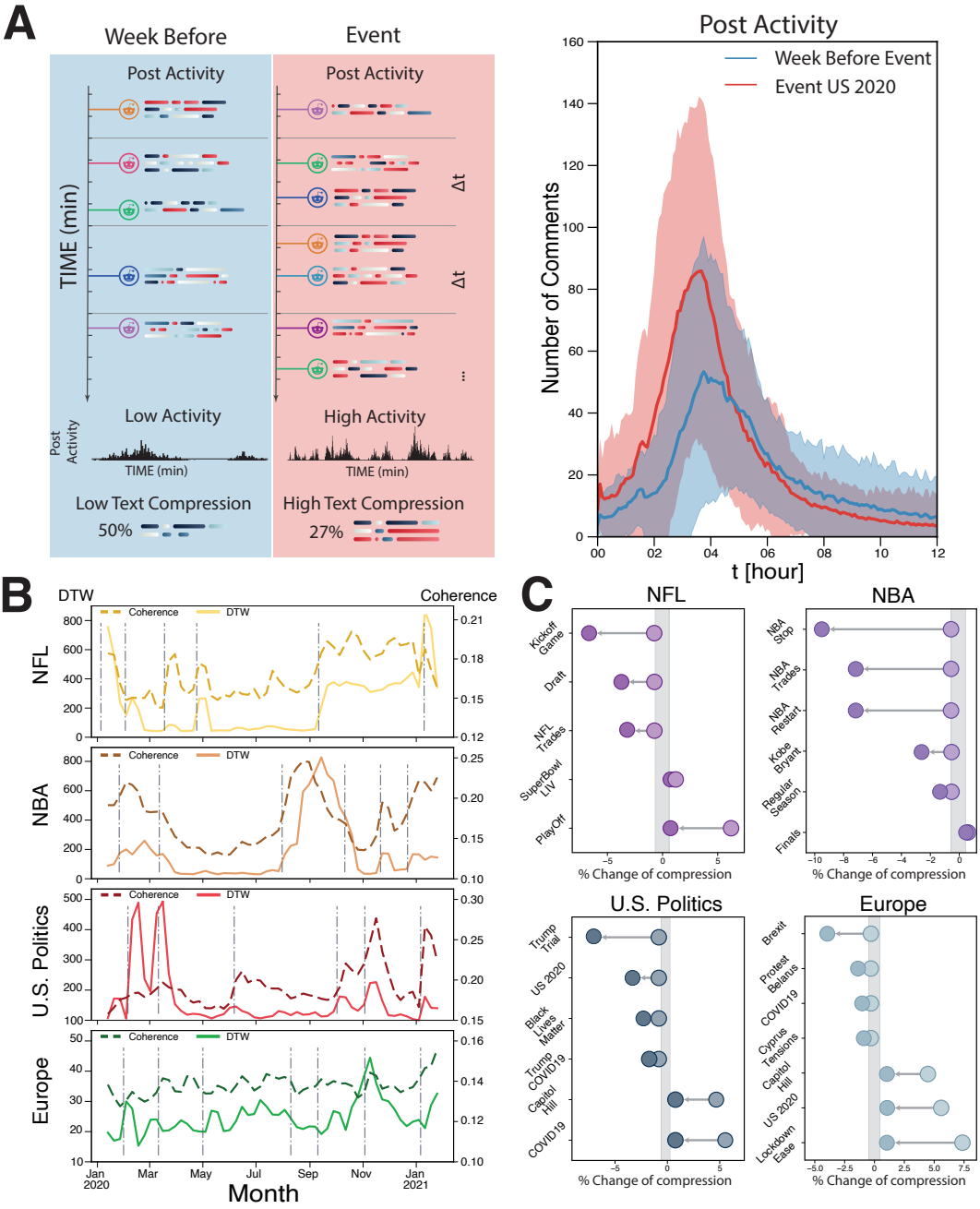

Figure 1: **Semantic space-time: compression and frequency of conversations.** A) Schematic representation of the framework. From each post we extract a time series by aggregating and counting the comments underneath within a 5 minutes time interval, and a text by joining all the comments. On the right it is shown the average time series of the US 2020 election (red) and the week before (blue). The shaded area represents the standard deviations. In the infographics on the left the repeated words are marked in red and the compression of the chat is reported at the bottom. B) Each panel shows the average Dynamic Time Warping (solid) and coherence (dashed) distances between the conversations of a week and the previous for each subreddit. The grey vertical dashed-dotted lines mark the events. C) Each panel shows the percentage change of compression between the week associated with the exogenous event (darker) and the week before (lighter) for each subreddit. The grey shaded vertical area is the standard deviation to the mean of the changes between one week and the preceding week.

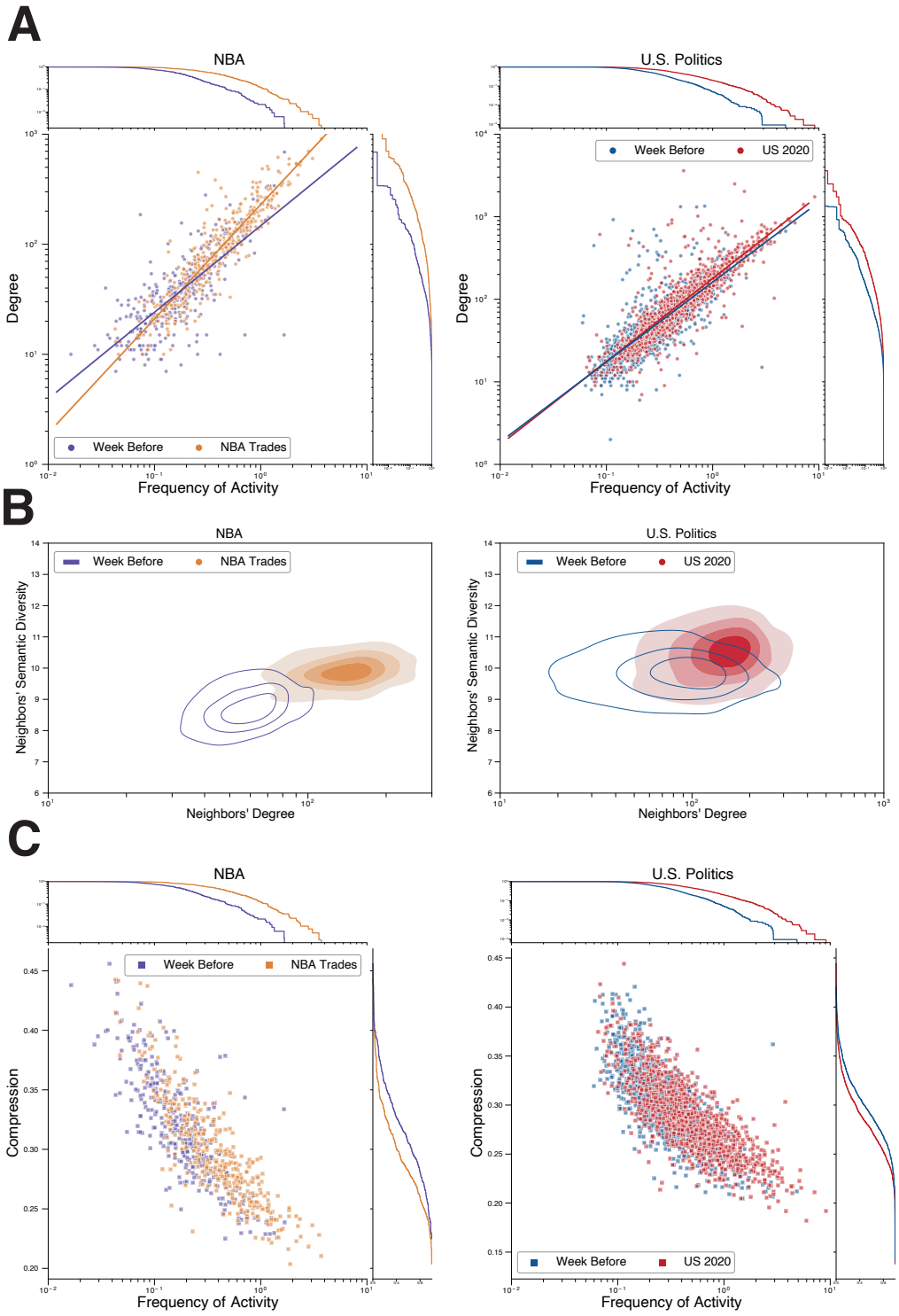

Figure 2: **Users' trajectories in the semantic space-time.** In the following subplots the data used on the left panels are of the users active on the subreddit r/NBA during the NBA Trades, while on the right of the users on r/politics during the US 2020 election. A) In the central panels it is shown the relation between the frequency of activity of each user and the interacting peers (degree). In the marginal plots it is reported the survival function of each variable and each week. B) The density plots show the variations of the peers' degree and semantic diversity. C) In the central panels it is shown the relation between the compression and the frequency of activity. In the marginal plots it is reported the survival function of each variable and each week.