Concept drift detection on social network data using cross-recurrence quantification analysis
Rodrigo F. de Mello, Ricardo A. Rios, Paulo A. Pagliosa, and Caio S. Lopes

Citation: Chaos 28, 085719 (2018); doi: 10.1063/1.5024241
View online: https://doi.org/10.1063/1.5024241
View Table of Contents: http://aip.scitation.org/toc/cha/28/8
Published by the American Institute of Physics

Articles you may be interested in
Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions

Extended recurrence plot and quantification for noisy continuous dynamical systems
Chaos: An Interdisciplinary Journal of Nonlinear Science 28, 085722 (2018); 10.1063/1.5025485

Social semantic networks: Measuring topic management in discourse using a pyramid of conceptual recurrence metrics

A recurrence quantification analysis-based channel-frequency convolutional neural network for emotion recognition from EEG

Deriving inferential statistics from recurrence plots: A recurrence-based test of differences between sample distributions and its comparison to the two-sample Kolmogorov-Smirnov test

Multivariate weighted recurrence network analysis of EEG signals from ERP-based smart home system
### I. INTRODUCTION

The last few decades have been characterized by the design and development of new technological tools to allow faster communication among people around the World. In that context, Twitter, a social network tool created in 2006, has a prominent success in providing short communications from one person to his/her followers. Currently, Twitter has more than 320 million of monthly active users (http://twitter.com/). Users’ activities are referred to as tweets and represent the context, Twitter, a social network tool created in 2006, has a prominent success in providing short communications from one person to his/her followers. Currently, Twitter has more than 320 million of monthly active users (http://twitter.com/). Tweets are written by users under a given nickname who may also add hashtags to their messages. Twitter provides an Application Programming Interface (API), so one can write a client-side application to monitor specific hashtags. By monitoring these subjects, one can attempt to model the public opinion (or ideological biases) along time and how they relate to different topics. For example, one could monitor the political or economical situation: (i) in Brazil to provide evidence about how people have been facing particular issues and (ii) in Syria to understand the war impacts on people and how other factors, such as health and refugees supports, have been provided.

In this context, we propose in this paper the comparison of trajectories of consecutive data windows in order to point out time series modifications along time, a.k.a. Concept Drift Detection in the context of Data Streams and Machine Learning. Concept Drift means the time instant in which the generation process, responsible for producing data observations, has changed. For instance, suppose a sinusoidal function was responsible for producing observations until a given time instant, then its frequency slightly changes (smooth drift) or the process is then ruled by a Logistic map (abrupt change). In such scenarios, we employ the Cross-Recurrence Quantification Analysis (CRQA) to measure the longest diagonal line \( L_{\text{max}} \) of two consecutive data windows to take conclusions whether they follow the same generating process along time.

We take any time series and set a given window length, which is used to slide along observations. Every data window is then provided as input to CRQA, which employs Takens’ immersion theorem to reconstruct such time series into the phase space. A consecutive window will, therefore, result in a further phase space to be compared against the publication of short texts (at most 140 characters), pictures, videos, links, and retweets (repost of someone else’s tweet).

While studying the generating processes of the time series collected from social network systems, we used Takens’ embedding theorem to reconstruct them into phase spaces and detect Concept Drifts using the Cross-Recurrence Quantification Analysis to compare how trajectories evolve by assessing the longest diagonal line. From that, we obtained a Concept Drift detection approach to point out the most relevant events associated with a given hashtag. For example, while analyzing the novelty level each new tweet adds to the historical information for hashtag dilmabr (the one used by the Brazil’s former president), we observed that the most relevant events were associated with the exact date the impeachment was voted in the House of Representatives, the exact date Petrobras (the main Brazilian oil company) confirmed the corruption to the stockholders, and the date Fidel Castro passed away (one of her main allies). We have now been constantly working on improving the TSViz (http://www.tsviz.com.br) project to provide additional features to end users.

**While studying the generating processes of the time series collected from social network systems, we used Takens’ embedding theorem to reconstruct them into phase spaces and detect Concept Drifts using the Cross-Recurrence Quantification Analysis to compare how trajectories evolve by assessing the longest diagonal line. From that, we obtained a Concept Drift detection approach to point out the most relevant events associated with a given hashtag. For example, while analyzing the novelty level each new tweet adds to the historical information for hashtag dilmabr (the one used by the Brazil’s former president), we observed that the most relevant events were associated with the exact date the impeachment was voted in the House of Representatives, the exact date Petrobras (the main Brazilian oil company) confirmed the corruption to the stockholders, and the date Fidel Castro passed away (one of her main allies). We have now been constantly working on improving the TSViz (http://www.tsviz.com.br) project to provide additional features to end users.**
previous one by using the CRQA measurements. In particular, in this paper we employ the maximum diagonal line to point out similarities between consecutive data windows and, consequently, detect when a significant data drift happens.

Our approach was implemented as part of the TSViz project (available at http://www.tsviz.com.br), a Web tool that monitors tweets and builds up data streams to be processed later. TSViz was designed to plug new components which are responsible for generating time series. For example, consider tweets associated with hashtags “#Brazil” and “#impeachment” have been collected along one year. TSViz is capable of processing a sequence of tweets under a given subject to produce: (i) time series with information novelties, helping us to understand how tweet content changes along time; (ii) time series with the frequencies of tweets per day; (iii) time series with the positive or negative sentiments that users stated along their messages; (iv) geopositioning system of users according to their profiles; and (v) time series with the novelty degree of most common words used throughout messages. This list of features is not extensive but simply an example of the components provided by TSViz.

In addition to building up these series, TSViz also provides data analysis tools so that users are capable of understanding tendencies and relationships among different subjects. Moreover, TSViz offers a Web tool so that users will visualize data evolution. From this perspective, we intend to support researchers from different areas of knowledge to understand how real-world situations are mapped or impact in the Social Media and in the public opinion.

As main contribution, we decided to analyze the most recent data collected by TSViz in order to detect relevant drifts, related to hashtags from politics as well as others producing collateral impacts in Brazil. The results confirm that the drifts detected by our approach are directly related to the main fact published by TV channels, newspapers, and the Web in general.

This paper is organized as follows: Sec. II introduces the TSViz project; Sec. III provides the necessary background on the Cross Recurrence Quantification Analysis; the results are presented and discussed in Sec. VI; Sec. VII draws conclusions and, finally, references are listed.

II. THE TSVIZ PROJECT

Twitter’s users daily produce a huge amount of data which motivated several researchers and companies to design tools and algorithms to collect and analyze tweets, aiming at extracting patterns to understand users’ behavior and feelings about specific subjects or trends. However, most of such analyses focused on creating a dataset of collected tweets and analyzing them by using offline approaches. This is the main gap that motivated us to propose the TSViz (Time Series Visualization) project, a Web tool designed to analyze information gathered from Twitter.

Figure 1 illustrates the main components involved in the TSViz architecture: (i) TSViz Robot; (ii) TSViz Core; and (iii) TSViz User Interface. TSViz Robot online collects every tweet published by users according to monitored hashtags or usernames.

We are already in the second version of this robot designed using the Twitter API (http://developer.twitter.com/en.html), which provides read-and-write access from/to Twitter servers. Using this API, we developed a listener that registers, in the Twitter servers, a set of hashtags and users to be monitored (Fig. 1—Arrow 1). Hence, once any tweet around the World is published using the monitored hashtags and/or users, the robot receives a message containing all information about them (Fig. 1—Arrow 2). In this manner, every hashtag or username defines a new multidimensional data stream/time series. Finally, TSViz Robot stores the tweet information into our databases to be later analyzed by the TSViz Core (Fig. 1—Arrow 3).

After monitoring and collecting tweets, the robot saves them as a multidimensional data stream/time series in our databases to be analyzed later. This data stream contains all tweets information, including creation timestamp, owner, geographic location, language, the tweet itself, and all other parameters provided by Twitter. Then another TSViz component, called TSViz Core, is responsible for performing two main tasks: (i) first, a producer module periodically looks for and incrementally converts all tweets associated either with a hashtag or with a username into time series (Fig. 1—Arrow 4). Then, the resultant time series is also stored into our databases using different time resolutions through a Discrete Wavelet Transform (DWT—Fig. 1—Arrow 5), so they can be analyzed later by the producer module. DWT is used to process time series in order to obtain different resolutions for end-users, so they are served fast (in terms of charts produced by TSViz) with such information without any additional processing stage. Collected messages are then processed using our Data Analysis modules (Fig. 1—Arrows 6 and 7) to make them available to end-users through the Web interface (Fig. 1—Arrows 8 and 9).

Currently, we have the following producer modules: (i) a novelty detection module based on the Normalized Compression Distance (NCD), to inform about binary changes along messages under the same context; (ii) a cumulative module that sums up series observations along time to discover data tendencies. This can be employed in conjunction with any of the other components; (iii) a module to extract the most relevant words in a given context along time; (iv) a module to perform the correlation among different subjects using the set of their most tweeted words along time; (v) a novelty detection module, based on Shannon’s entropy, to point out
changes in the set of words under a subject; (v) a geopositioning module that takes users’ profiles and associates them with the most probable place of tweeting; (vi) a sentiment analysis module to measure the level of positiveness of negativeness of posted messages; and, finally, (vii) a concept drift detection module based on the Cross-Recurrence Quantification Analysis (CRQA) that informs the most important moments in the generation processes. Section III details CRQA, and afterwards our approach to detect drifts in time series from social networks is introduced.

III. CROSS-RECURRENCE QUANTIFICATION ANALYSIS

Recurrence analysis tools are useful to study and characterize the behavior of dynamical systems by reconstructing the produced data in phase spaces. The reconstruction of such spaces is typically performed using Takens’ immersion theorem that allows one to analyze system states from which one may obtain a regression function to model and forecast future states. Among those tools, Recurrence Plot (RP) supports the study and analysis of time series recurrences. In summary, RP produces an output matrix, referred to as Recurrence Matrix, representing how close two states are in phase space, thus allowing us to understand how states evolve along time. Equation (1) is employed to measure nearby phase space, thus allowing us to understand how states evolve along time. Equation (1) is employed to measure nearby

\[ R_{ij}(\varepsilon) = \Theta(\varepsilon - ||\vec{x}_i - \vec{x}_j||), \quad \{i,j = 1,2,\ldots,N\}, \]  

(1)

where \( \Theta(\cdot) \) is a heaviside function [Eq. (2)]:

\[ \Theta(\alpha) = \begin{cases} 0, & \alpha < 0 \\ 1, & \alpha \geq 0 \end{cases} \]  

(2)

To proceed with the RP analysis, the recurrence matrix is plotted by using black dots when \( R_{ij} = 1 \) and white ones when \( R_{ij} = 0 \). Structures present in RP provide important information on the dynamical systems under study. For instance, isolated points mean system states are rarely repeated, emphasizing a stochastic behavior. On the other hand, diagonal lines occur when there is persistent behavior, showing that system states are highly recurrent.

As previously discussed, the RP analysis is performed on a single time series unfolded on its phase space. However, there exists scenarios in which it is important to analyze the relationship between states from two different systems. Aiming at overcoming this situation, Cross Recurrence Plots (CRPs) are a natural complement of RPs, which intend to study the dynamic of two systems by representing them into the same phase space, looking for occurrences when state of a system recurs to state of the other. Similar to RP, CRP also produces a matrix, whose recurrence is defined by Eq. (3), in which \( \vec{x} \) and \( \vec{y} \) represent the trajectories of two dynamical systems in a \( d \)-dimensional phase space:

\[ CR_{ij}^{xy}(\varepsilon) = \Theta(\varepsilon - ||\vec{x}_i - \vec{y}_j||), \quad \{i = 1,2,\ldots,N\}, \]  

(3)

\[ j = 1,2,\ldots,M \].

The structures produced by CRP are analyzed by the Cross Recurrence Quantification Analysis (CRQA), which provides a set of measures to quantify, for example, the determinism rate, the recurrence rate, the maximal diagonal line, etc. In this article, we are particularly interested in the maximal diagonal line \( (L_{\text{max}}) \) measure, whose results are shown in Sec. VI. The length of a diagonal line \( (l) \) is related to the time steps in which two trajectories in phase space are close to each other, emphasizing similar behavior. By measuring the maximal diagonal line in a RP plot, we can analyze the exponential divergence of the phase space trajectory. In summary, the faster the trajectory segments diverge, the shorter the diagonal lines are.

The maximal diagonal line \( (L_{\text{max}}) \) is calculated by Eq. (4), in which \( P(l) \) is the frequency distribution of diagonal lines with length equal to \( l \):

\[ L_{\text{max}} = \max \left( \{ l_i \}_{i=1}^{N_l} \right), \]  

(4)

\[ N_l = \sum_{l \leq l_{\text{max}}} P(l). \]  

(5)

In Sec. IV, we present our approach using CRQA to analyze how the public opinion on specific topics has changed over time.

IV. A NOVEL CONCEPT DRIFT APPROACH

In this work, we designed a new approach to detect concept drift in time series. Its main objective is to detect when the generating process, responsible for producing data observations, changes along time. For example, the observations of a time series may be produced by a Logistic map under some parametrization, which slightly changes along time (smooth drift) or significantly modifies to another process, such as a sinusoidal function (abrupt drift). In this context, we decided to take into consideration the time series produced by the TSViz project (http://www.tsviz.com.br) after running the NCD and the Sentiment Analysis modules.

NCD allows one to compute the similarity among texts published in Twitter. By using this producer, we can measure if new textual content has been shared or basically the same (as retweeting). The resultant time series is obtained by applying NCD on pairs of consecutive tweets as defined below:

\[ X_N = \{ \text{NCD}(t_1,t_2), \ldots, \text{NCD}(t_i,t_{i+1}), \ldots, \text{NCD}(t_{T-1},t_T) \}. \]  

(6)

In this equation, \( T \) is the total number of collected tweets at the current time instant. The \( \text{NCD}(-) \) function is presented in Eq. (7) such that \( Z(\cdot) \) is the binary length of a compressed file in bytes. The compression tool adopted in our experiment was BZip2. In our case, a file contains either the text of a single tweet \( Z(t_i) \) or concatenated texts of two consecutive tweets \( Z(t_i,t_j) \):

\[ \text{NCD}(t_i,t_j) = \frac{Z(t_j) - \min[Z(t_i),Z(t_j)]}{\max[Z(t_i),Z(t_j)]}. \]  

(7)

The Sentiment Analysis module was implemented using a Naïve Bayes classifier. We implemented such a classifier to employ a training stage, in which the probabilities of words
published along tweets are used to label unknown posts as being negative or positive in a continuous range $[-1, 1]$. Then, we only classify tweets daily published, calculating their probability of positiveness and negativeness by using Eqs. (8) and (9), respectively. They compute the product of probabilities that every word $w$ has for wither supervised label (positive or negative). As a consequence, we obtain a time series summarizing positive and negative feelings about a given hashtag or username along time using Eq. (12), in which $SA(t_i) \in [-1, 1]$ corresponds to the overall sentiment analysis for a single tweet. Equations (10) and (11) are used to compute such time series, in which we basically normalize the positive and negative probabilities to sum up to one:

$$P(t_i, \text{Positive}) = \prod_{w \in t_i} P(w, \text{Positive}),$$  \hspace{1cm} (8)$$

$$P(t_i, \text{Negative}) = \prod_{w \in t_i} P(w, \text{Negative}),$$  \hspace{1cm} (9)$$

$$NP(t_i) = \frac{P(t_i, \text{Positive})}{P(t_i, \text{Positive}) + P(t_i, \text{Negative})},$$  \hspace{1cm} (10)$$

$$NN(t_i) = \frac{P(t_i, \text{Negative})}{P(t_i, \text{Positive}) + P(t_i, \text{Negative})},$$  \hspace{1cm} (11)$$

$$SA(t_i) = NP(t_i) - NN(t_i).$$  \hspace{1cm} (12)$$

Equations (10)–(12) are responsible to provide a sentiment degree between $-1$ and $1$, not only strictly classifying a post as either positive or negative. This was designed to behave as similarly as a correlation index, in which $-1$ means totally negative, $+1$ means totally positive, and 0 no relevant sentiment.

After obtaining the time series generated by NCD and Sentiment Analysis, we designed an approach that uses CRQA to detect concept drifts. In summary, this approach runs two straightforward steps. First, we define an initial time window $\omega_i$ containing a set of observations from the time series produced by either NCD or Sentiment Analysis. Then, the window is slid considering a time interval $\omega_i + \tau$ and a new set of observation is extracted. Secondly, the two windows are analyzed by using CRQA $\left[CR_{t_i-j, t_i-j+\tau}(\epsilon)\right]$, and the recurrence rate is stored into a new time series. The process is, finally, repeated by sliding the windows up to reach the last observation. The new time series will keep the similarity between pairs of windows and, if a difference is noticed, then a concept drift detected, i.e., a new information was detected or the people’s sentiment has changed.

In order to exemplify the advantage of using our approach, consider a time series created by two different generation processes. The first observations were created using a sine function, whereas the last ones generated from a Logistic map, as shown in Fig. 2(a).

![Fig. 2](image_url)

**Fig. 2.** (a) A time series created using a sine function and a Logistic map. (b) $L_{\text{max}}$ values obtained while comparing consecutive time windows. (c) The red-dashed line represents the concept drift obtained with our approach, while the blue-dashed one corresponds to Pettitt’s test, a common baseline method widely employed in the literature.
Using our approach, we first calculate the maximal diagonal line $L_{\max}$ by comparing pairs of windows $\mathbf{CR}_{ij}^{\omega_i} - \mathbf{CR}_{ij}^{\omega_i + \tau} (\epsilon)$. The value of every comparison is organized as a new time series as shown in Fig. 2(b). As one can notice, at a given time instant, this new time series presents a strong decay, emphasizing the moment when the general behavior had changed. To clarify, we zoomed in the original time series to highlight the data drift, as shown in Fig. 2(c). In this figure, we also plotted a red-dashed line representing the time instant in which our approach has detected a new concept drift. The blue-dashed line shows the concept drift detected using Pettitt’s test.\textsuperscript{18} This is a well-known test based on the alternative hypothesis that some change has happened along the time series.

As presented in this figure, the main drawback faced by general concept drift methods, including Pettitt’s test, is to detect changing points when the time series is stationary, i.e., they typically state a new behavior occurs when parameters are modified over time, as mean and variance.

In Sec. V, we present some details about the dataset and parameters used in our experiments.

V. EXPERIMENTAL SETUP

As previously discussed in Sec. I, we used CRQA to understand how people reactions in Twitter evolve along time and how that is connected to important facts from the real world. In this sense, we selected a set of Twitter hashtags and users to be monitored by TSViz (see Table I). It is important to highlight that such hashtags and users are not related to the authors’ political view but to the main terms used by the general media. The main motivation to select them was to understand how they are connected to the severe issues faced in Brazil, leading to a strong and highly frequent reaction on Twitter.

<table>
<thead>
<tr>
<th>Hashtag/user</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dilmabr</td>
<td>Twitter account of the former Brazil’s president Dilma Rousseff</td>
</tr>
<tr>
<td>LulaPeloBrasil</td>
<td>Twitter account of Luís Inácio Lula da Silva, also a former president</td>
</tr>
<tr>
<td>MichelTemer</td>
<td>Twitter account of the current president of Brazil Michel Temer</td>
</tr>
<tr>
<td>ForaTemer</td>
<td>Hashtag against Michel Temer</td>
</tr>
<tr>
<td>jairbolsonaro</td>
<td>Twitter account of a controversial figure in Brazil’s political scenario</td>
</tr>
<tr>
<td>petrobras</td>
<td>a semi-public Brazilian multinational corporation in the petroleum industry</td>
</tr>
<tr>
<td>SenadoFederal</td>
<td>Federal Upper House of the National Congress of Brazil</td>
</tr>
<tr>
<td>CamaraDeputados</td>
<td>Lower Chamber of Deputies</td>
</tr>
<tr>
<td>politica</td>
<td>Hashtag associated with politics in Brazil</td>
</tr>
<tr>
<td>STF_oficial</td>
<td>Twitter account of the Supreme Federal Court of Brazil</td>
</tr>
<tr>
<td>dolar</td>
<td>Portuguese term for “dollar”</td>
</tr>
<tr>
<td>dengue</td>
<td>Term associated with the disease “dengue”</td>
</tr>
<tr>
<td>Zika</td>
<td>Term associated with the disease “zika”</td>
</tr>
<tr>
<td>saude</td>
<td>Portuguese term for “health”</td>
</tr>
</tbody>
</table>

As these hashtags and users have been monitored by the TSViz Robot, the collected texts (for any upper or lower written form) are transformed into time series by the TSViz Core component. Although there exists several producers, in this paper we focused on the Normalized Compression Distance (NCD) and Sentiment Analysis (SA).

Finally, to support experimental reproducibility, it is worth to emphasize that the embedding and time-delay dimensions used to unfold the time series assessed in this work were estimated using the False Nearest Neighbors (FNN) and Average Mutual Information (AMI) methods. FNN analyzes the neighborhood for every observation in the phase space, in an attempt to measure variances after adding a new space axis.\textsuperscript{12} This method starts calculating the distance among observations considering the embedding dimension equals to one. Then, a new dimension is added and distances are again calculated, so if they increase, observations are considered as false neighbors. The stopping criterion is found when the false neighbor rate approaches zero so that such number of dimensions is taken as the embedding parameter. Regarding the time delay, we applied AMI on every time series to assess different time lags. Results were plotted as a curve and the first minimum was selected as the dimension, as proposed by Fraser and Swinney.\textsuperscript{7} Both methods have been widely used in the literature of Dynamical Systems.

In Sec. VI, we discuss the results obtained by using our concept drift approach.

VI. RESULTS

We start with the analysis of the former Brazil’s president, Dilma Rousseff. According to CRQA, whose maximal diagonal lines $L_{\max}$ along the data windows are provided in Fig. 3, we notice very relevant events on November 4 and 5, 2017, for the NCD series (this means that CRQA was...
computed on NCD and \( L_{\text{max}} \) was plotted to illustrate our conclusions). That was motivated by her tweets in which she declared male prejudice against women during her presidential term. In the same period, one of her followers posted a message claiming she is criminal, reinforcing novelties and, consequently, the drift detection found.

Luiz Inácio Lula da Silva, a.k.a. Lula, was president of Brazil from 2003 to 2011. He is a founding member of the Workers’ Party (Partido dos Trabalhadores—PT) and referred to as one of the most popular politicians in the history of Brazil. Lula was convicted of money laundering and passive corruption, defined in Brazilian criminal law as the receipt of a bribe by a civil servant or government official. He was sentenced to nine years and six months in prison by judge Sérgio Moro under the Operation Car Wash (Operation Lava Jato in Portuguese) but remains free pending an appeal of such sentence.

There is a particular date, in Fig. 4, that calls our attention while analyzing her hashtag, which is July 1, 2017, the day that Joesley Batista (one of the owners of JBS—the largest, by sales, meat processing company in the World) stated to have given a bribe of R$ 300 million (around US$ 90 million) to Lula.

We also analyzed publications that directly mentioned the current Brazil’s president Michel Temer by using CRQA on NCD [Fig. 5(a)]. The first higher divergences happened in May 12, 2017, when he completed 1 year as president after the impeachment of Dilma Rousseff. A week later, a new peak was detected when the local media released a secret recording in which Michel Temer is discussing hush money with Wesley Batista, a chief executive from one of JBS.

On May 28, 2017, a few days before the judgment of his impeachment, Michel Temer changed the minister of justice indicating a person who has a good relationship with the Superior Electoral Court and the Supreme Federal Court. People faced this unconventional move, performed by Temer, as a way to defend himself against corruption allegations.

On July 11, 2017, two important events happened involving Temer. In the first one, Brazil’s senate approved a significant overhaul of labor rules. The second one happened after the rapporteur of the lower house committee, that was examining a corruption charge against Temer, recommended to vote putting him on impeachment trial. However, On July 15, 2017, the lower house committee rejected the corruption charge, triggering another divergence spike. In September 2017, NCD emphasized another divergence when Temer visited China and stated the new labor rules had reduced the number of unemployed people. Finally, other spikes were...
detected in December 2017 when Temer announced his next priority was to approve pension reform. As the labor, the pension reform faced several polemical issues, causing strong reactions on Twitter.

The sentiment analysis performed on tweets mentioning the current Brazil’s president Michel Temer has presented similar divergences when compared to NCD. In summary, our concept drift analysis detected spikes related to people reaction when Temer completed 1 year in presidency, also during his judgement in the lower house, and while announcing the pension reform as shown in Fig. 5(b).

Several detections were produced for the hastag “foratemer” (get out Michel Temer) on the following dates April 5–6, July 6–9, August 2, and September 10, 2017, while analyzing NCD [Fig. 6(a)]. On April 5–6, 2017, a worker painted the message “Fora Temer” on the walls of the National Museum in Brasília, the capital of Brazil. On July 6–9, 2017, several artists started a campaign in an attempt to impeach Temer. On August 2, 2017, a group of deputies get into the lower chamber with a band and posters asking for the impeachment. On September 10, 2017, a famous Brazilian musician, Gilberto Gil, stated Michel Temer had been already...
impeached after a gig that commemorated 40 years of his classic album “Refavela.” The same dates are confirmed while analyzing SA [Fig. 6(b)].

A different behavior is observed while analyzing the detections for Jair Bolsonaro, a Brazilian politician and former military officer who was elected into the Chamber of Deputies. He is seen as a controversial figure in Brazil, due to his positions against the left wing. He has been visiting several cities around Brazil, being acclaimed by some and hated by others. That is in fact observed while analyzing NCD and SA (Fig. 7), given no evident and single detection, but instead several changes along time. That perception is the same collected along news and Youtube about this public person.

Petrobras (Petróleo Brasileiro SA) is a semi-public Brazilian multinational corporation in the petroleum industry. In 2014, the largest corruption scandal in the history of Brazil was centered around Petrobras. Here, we analyze the concept drift of NCD (Fig. 8) whose main dates are October 20–21, 2017. On those particular dates, the operation Car Wash launched ten court orders resulting from the payment of bribes from Odebrecht (Brazilian conglomerate consisting of diversified businesses in the fields of engineering, construction, chemicals, and petrochemicals).

Next, we analyzed the Federal Senate, which is the upper house of the National Congress of Brazil. In terms of NCD drifts [Fig. 9(a)], we found the following important dates: July 4–5 and December 25, 2017. The motivation for drifts on July 4–5, 2017, is due to the fact that senate had voted the labor amendment to change employment contracts in Brazil and to make them move a little more towards a liberal format. On December 25, 2017, several news were published about involvements of the main politicians from the Senate with the Operation Car Wash. In terms of SA [Fig. 9(b)], the main dates were December 19–26, 2017, which are related to the same subjects but also including a public petition for the impeachment of Gilmar Mendes, one of the members of the Supreme Court of Brazil, who has released important politicians from prison.

The Chamber of Deputies composes the federal legislative lower house of the National Congress of Brazil. Such a chamber comprises 513 deputies elected in a proportional fashion to represent each state in a four-year term. Figures 10(a) and 10(b) show the concept drift detection results computed on two respective time series: the Normalized Compression Distance (NCD) and the Sentiment Analysis (SA).
In the case of NCD [Fig. 10(a)], the most relevant dates are July 30, November 2, and December 19, 2017. On July 30, 2017, the Chamber started discussing about the allegations against the current Brazil’s president Michel Temer. On November 2, 2017, Edson Fachin, one of the members of the Supreme Federal Court of Brazil made the decision to carry on with a judicial process against Michel Temer and two of his ministers (Eliseu Padilha and Moreira Franco), besides the Chamber of Deputies had denied the authorization to file a suit against the president. On December 19, 2017, the Chamber approved the lobbying activity in the entities of the federal public administration.

In the case of SA [Fig. 10(b)], two of the same dates were also detected: July 30, November 2, and December 19, 2017, confirming the same facts. Although a new date appears as important too: August 25, 2017, due to the public opinion against corruption, the Chamber of Deputies is pressured to propose new forms of political organizations. On that particular date, they discussed about the public funding of elections, the party coalitions, and barrier clauses to avoid the evergrowing number of parties.

The Supreme Federal Court of Brazil is the highest court of law in Brazil for constitutional issues and its rulings cannot be appealed. While analyzing NCD and SA (Fig. 11),
one important date was detected: December 22, 2017. On that
day, a public petition for the impeachment of Gilmar
Mendes, member of such court, was filed with 1.6 million sig-
natures from all around Brazil. It is important to mention that
he released some politicians from prison and many Brazil-
ians believe that is due to his interests with other influential
companies and politicians.

After analyzing specific users and terms that have been
playing an important role in the current political scenario
in Brazil, we decided to study the general term “politic”
(Fig. 12). As a consequence, it was possible to understand the
relationships among such users and terms and discover other
themes that have been calling the attention of the population
in general.

The first observed spike happened on June 4, 2017, when
the local media published analyses on how the political crisis
has affected other areas such as education and economy. On
August 12, 2017, a set of news was published pointing out
the need of carrying out the political reform, before perform-
ing labor and pension reforms. On August 23, 2017, a NCD
divergence emphasized people reactions when the Chamber
of Deputies authorized the use of almost 10 billion dollars as
budget for the 2018 election, in spite of the severe economical
and political crisis in Brazil. In December 2017, new spikes
were detected when Temer submitted a surgical procedure.
Finally, a new divergence, not detected by NCD, has showed
people reaction when the new Labor Minister of Temer’s gov-
ernment was banned from the position after being found guilty
of not paying private workers.

The Brazilian political crisis has directly and strongly
affected the local economy. This can be verified by perform-
ing a sentiment analysis on tweets mentioning the hashtag
“dolar” as shown in Fig. 13. A relevant spike happened on
June 19, 2017, when the current exchange rate between US
Dollar (USD) and BRAZIL REAL (BRL) has significantly
increased. However, this rate has reduced on August 18, 2017,
as Temer and allies approved some reforms in the congress.
On the other hand, on October 16, 2017, as new corruption
cases involving Temer were released, the rate also reacted,
increasing its value.

Finally, we also used NCD to analyze tweets published by
using the following hashtags “dengue,” “zika,” and “saude”
(health). Such analyses were useful to understand people’s
reaction by facing new symptoms and consequences of dis-
eases transmitted by the Aedes aegypti mosquito.

Figure 14 shows the NCD results after analyzing the
tweets published from May 2017 to January 2017 with
the hashtag “dengue.” By observing this figure, we corre-
late the main spikes to important historical facts that hap-
pened around the World. The spikes between May 28 and
29, 2017, reflected high epidemic levels in different parts of
the World as, for instance, Northeast region of Brazil, Delhi
(India), Myanmar, and Sri Lanka. Although dengue fever is
a common virus in such regions, an increasing number of
the dengue hemorrhagic fever cases, also known as severe
dengue, have called the attention of the population, espe-
cially survivors of the devastating floods and landslides of Sri Lanka
caused by heavy southwest monsoon during the final week of
May 2017.

We also noticed another important spike on August 9,
2017, when an article published in Nature confirmed that the
same mosquito species could also transmit the zika virus.
On the same day, a campaign was started in Brazil to con-
trol dengue as well as to provide public awareness about the
habitat and life cycle of the aedes aegypti.

On September 29, 2017, and November 5, 2017, new
epidemics were registered in Brazil caused by the same
mosquito. However, on these dates, the media reported sev-
eral patients diagnosed with another virus called chikungunya.
Finally, new spikes were also detected during November

![FIG. 12. Maximal Diagonal Line for “politic” computed using the Normal-
ized Compression Distance.](image1)

![FIG. 13. Maximal Diagonal Line for “dolar” computed using the Sentiment
Analysis.](image2)
19–20, 2017, when the number of notified cases of dengue has considerably increased in Brazil. This spring period in Brazil is characterized by a greater volume of rain, creating favorable conditions to multiply the mosquito population. For that reason, people reacted in Twitter as the number of reported patients increased, aiming at spreading information on health education during the National Dengue Day.

The zika virus is also transmitted by the same mosquito of the dengue fever. However, this virus is especially aggressive when women are infected in the first trimester of pregnancy, affecting the fetus’ central nervous system and leading to malformed babies with microcephaly. By analyzing the NCD results on tweets published using the hashtag “zika” (Fig. 15), we noticed four main spikes. The first one happened on July 7, 2017, when Spanish scientists discovered a molecule that could be used as a potential drug to cure patients with the zika virus. The second peak summarizes people’s reactions on September 17, 2017, when American researchers published results suggesting that the zika virus could kill brain cancer stem cells. On October 10, 2017, the Brazilian ministry of health showed that the uncontrolled mosquito vectors have increased not only the transmission of zika but also malaria, surpassing the previous year by almost 30%. Finally, we noticed a great NCD divergence during October 17–25, 2017 when tests to detect the zika virus in blood donors were approved by the US Food and Drug Administration. During the same period, a Brazilian foundation for tropical medicine started a selection to identify patients infected with the zika virus that were able to participate in a study to discover how long the virus could survive in human body.

As a last result (Fig. 16), we analyzed the NCD divergence computed on tweets published from May 2017 to January 2017, containing a Portuguese word meaning health. We decided to monitor such a word to identify whether the previous hashtags had directly influenced the general publications on health and to understand which are the other concerns related to public health. The first spike was associated with the day on which news reported that the public health care system in Brazil has received medicines to treat babies with microcephaly, whose number of cases increased due to the zika virus. The second spike was related to people’s reactions by getting to know that new medicines to treat cancer were available in government network of distribution (free of charge). Finally, the last spike occurred during December 22–24, 2017, when the local media published the private health insurance
companies had significantly reimbursed the Brazil’s public health system (46% greater than in 2016).

All references supporting our conclusions are listed in the Appendix (segmented by hashtag). Any person knowing the Brazil’s situation can conclude references are broad enough in terms of political opinions.

VII. CONCLUSION

This paper presents an innovative application of Cross-Recurrence Quantification Analysis (CRQA) to understand people’s reactions in Twitter while facing real-world issues. In this context, CRQA confirmed to be an important tool to identify implicit information published in social networks. The first information summarized the similarity rate among different texts. By using NCD (Normalized Compression Distance) and CRQA, we assessed whether most of the posts expressed individual opinions or a large sharing of common points of view. According to our results, we noticed that different posts are mostly published as new important facts happened in real-world and CRQA helped us to identify such divergences. The second information analyzed in this work was extracted by using a Sentiment Analysis (SA) approach in conjunction with CRQA. After classifying messages according to their positiveness and negativeness, CRQA was used to detect concept drift, in which the public opinion on a specific topic has changed over time. Based on our results, we noticed that CRQA and SA allowed us to identify important spikes that represented a reaction of Twitter users. Experiments analyzed relevant subjects associated with Brazil, including politics, health, and economics. As future work, we intend to adopt other CRQA measures to extract new information in the context of social network analysis as, for instance, the determinism level to understand how predictable users or hashtags are along time.

ACKNOWLEDGMENTS

This paper is based upon projects sponsored by CNPq, Brazil, under Grant Nos. 441583/2014-8, 303051/2014-0, and 302077/2017-0, and FAPESP, Brazil, under Grant Nos. 2013/07375-0 and 2017/16548-6. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of CNPq and FAPESP. The TSViz project has been under registration process according to the Brazilian innovation laws. Anyone interested in using TSViz should contact the authors of this paper.

APPENDIX: NEWS ON THE ANALYZED HASHTAGS AND USERS

This appendix presents a list with all news sources (Table II) that were used in the experiments to confirm the concept drift detection. These websites were last visited on March 19, 2018.

<table>
<thead>
<tr>
<th>TABLE II. Websites used to confirm the concept drift detection.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>MichelTemer</strong></td>
</tr>
<tr>
<td><strong>ForaTemer</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>


https://www12.senado.leg.br/noticias/audios/2017/12/senadores-comentam-liminar-de-lewandowski-que-garante-aumento-dos-servidores
http://www.valor.com.br/politica/5162928/pf-apura-pagamentos-de-vantagens-indevidas-petrobras
http://agenciabrasil.ebc.com.br/politica/noticia/2017-08/entenda-o-que-e-o-modelo-distritao
http://www.eb.mil.br/web/resenha/display/-/asset_publisher/9B8IpAnDp1we/content/pedido-de-maluf-e-negado-no-stf
https://www12.senado.leg.br/noticias/audios/2017/12/senadores-comentam-liminar-de-lewandowski-que-garante-aumento-dos-servidores