# Multidimensional Outlier Detection in Interaction Data: Application to Political Communication on Twitter

Audrey Wilmet, Robin Lamarche-Perrin

# Multidimensional Outlier Detection in Interaction Data: Application to Political Communication on Twitter

Audrey Wilmet and Robin Lamarche-Perrin

Sorbonne Université, UMR 7606, LIP6, F-75005 Paris, France
Institut des Systèmes Complexes de Paris Île-de-France, ISC-PIF, Paris, France
firstname.lastname@lip6.fr

**Abstract**

We introduce a method which aims at getting a better understanding of how millions of interactions may result in global events. Given a set of dimensions and a context, we find different types of outliers: a user during a given hour which is abnormal compared to its usual behaviour, a relationship between two users which is abnormal compared to all other relationships, etc. We apply our method on a set of retweets related to the 2017 French presidential election and show that one can build interesting insights regarding political organization on Twitter.

## 1 Introduction

Within Twitter, users can post information via tweets as well as spread information by retweeting tweets of other users. This dissemination of information from a variety of perspectives may lead to global events which affect users' opinions.
In this paper, we introduce a method which aims at getting a better understanding of how these interactions are organised. To this end, we look for outliers in interaction data formed from a set of retweets. For instance, an event in a data stream is an outlier: it can be view as a statistical deviation of the total number of retweets at a given point in time. More generally, outliers, depending on which dimensions define them, highlight instants, users, users during given periods, or interactions for which the retweeting process behaves unusually. Therefore, they constitute important information regarding interactions' organisation.
We consider an interaction to be a triplet $(s, a, t)$ meaning that user $s$, called the spreader, has retweeted a tweet of user $a$, called the author, at time $t$. We model the set of interactions as a data cube with three dimensions: spreaders, authors and time. This representation enables us to access to local information, as the number of retweets between two users during a specific hour, as well as more global and aggregated information, as for instance the total number of retweets during a given hour. In the next step, we combine and compare theses different quantities in order to find outliers according to different contexts. This multidimensional analysis gives us insight into the possible reasons why some events emerge more than others and, in particular, whether they are global phenomena or, whether they originate from specific actors only.
The paper is organized as follows. First, we review the related work about outlier detection within Twitter in Section 2. In Section 3, we introduce the modelling of interactions as a data cube. In the following, we describe our method in Section 4 and apply it to a set of

retweets related to the 2017 French presidential election in Section 5. Finally, Section 6 concludes the paper with future work.

## 2 Related Work

The problem of outlier detection on Twitter has been approached in various ways depending on how outliers are defined. Some researchers consider outliers as real-world events taking place at a given place and at a given moment. For example, Sakaki *et al.* [14] and Bruns *et al.* [2] trace specific keywords attributed to an event and find such outliers by monitoring temporal changes in word usage within tweets. In other approaches, authors infer, from timestamps, geo-localizations and tweet contents, a similarity between each pair of tweet and find event into clusters of similar tweets, see for instance the works of Dong *et al.* [6], Li *et al.* [12] and Walther *et al.* [18]. Other researchers, instead, consider outliers as users with abnormal behaviours according to different criteria. For instance, Varol *et al.* [17] detect bots by means of a supervised machine learning technique. The work of Stieglitz *et al.* [16] focus on influential users by investigating the correlation between the vocabulary they use in tweets and the number of time they are retweeted. Ribeiro *et al.* [13], on the other hand, detect hateful users by means of a lexicon-based method. Finally, other works aim at finding privileged relationships between users. Among those, the work of Wong *et al.* [19] apply it to political leaning by combining an analysis of the number of retweets between two users with a sentiment analysis on the retweeted tweets.
With our approach, we want to treat these different types of outliers in a unified way as well as consider different perspectives in the way outliers are considered abnormal. Hence, not only we consider different entities as abnormal users; abnormal relationships; abnormal behaviours of users during specific hours, *etc.*, but also different contexts in which outliers are defined. Thus, an abnormal user may be abnormal during a given hour compared to the way it usually behaves during other hours, but also compared to the behaviour of all other users during the same hour. In this way, our framework aims to give a more complete picture of how users act, interact, and are organized along time in a way similar to what Grasland *et al.* [10] do in the case of media coverage in newspapers.
In practice, instead of characterizing and detecting outliers using tweets' content, as a lot of current approaches do, included those set out above, we focus on interactions' volume and structure. Indeed, text-mining techniques, although providing meaningful results, face challenges as the ambiguity of the language and the fact that resultant models are language-dependent and topic-dependent. Other authors point into this direction, see for instance the works of Chavoshi *et al.* [4] and Chierichetti *et al.* [5], which focus on volume-based features as the number of tweets and retweets, as well as the works of Song *et al.* [15] and Bild *et al.* [1] which focus instead on graph-based techniques.

## 3 Formalism

We denote the set of interactions by a set $E$ of triplets such that $(s, a, t) \in E$ indicates that $s$, called the spreader, has retweeted $a$, called the author, at time $t$. We model this set as a data cube. In this section, we formally define this tool as well as the possible operations we can apply on it to explore data in all its dimensions and to have access to more or less aggregated information.

## 3.1 Data Cube Definition

A data cube is a general term used to refer to a multi-dimensional array of values [11]. Given $n$ dimensions characterized by $n$ sets $X_1, ..., X_n$, we can built $N = \sum_{i=0}^{n} \binom{n}{n-i}$ data cubes, each representing a different degree of aggregation of data. The quantity $\binom{n}{n-i}$ corresponds to the number of data cubes of dimension $n - i$ in which $i$ dimensions are aggregated. Within this set of data cubes, we call the base cuboid the cube which has the lowest degree of aggregation. We denote it $\mathcal{C}_n(X, f)$ where $X = X_1 \times ... \times X_n$ is the Cartesian product of the $n$ sets $X_1, ..., X_n$, and $f$ a feature which maps each $n$-uplet to a value:

$$\begin{array}{rccc} f: & X & \longrightarrow & W \\ & (x_1, ..., x_n) & \longmapsto & f(x_1, ..., x_n) \end{array}$$

where $W$ is the value space of the feature. In the following, $n$-uplets are also called cells of the cube and denoted $c$ such that $c = (x_1, ..., x_n) \in X$.

**Dimensions** are the entities with respect to which we want to study data. In this paper, the three dimensions we consider are: the *spreaders*, denoted $S$, the *authors*, denoted $A$, and *time*, denoted $T$. In addition, we can organise elements of a dimension into sub-dimensions. For instance, the temporal dimension can be organised depending on temporal granularity. In our case, we divide it into the two sub-dimensions *days*, denoted $D$, and *hours*, denoted $H$, such that $t = (d, h)$ with $(d, h) \in D \times H$.

**The feature** is a numerical measure which provides the quantities according to which we want to analyse relationships between dimensions. Here we consider the *quantity of interaction*, denoted $v$. It gives the number of retweets for any combination of the three dimensions. In the base cuboid, $v(s, a, (d, h))$ gives the number of times $s$ retweeted $a$ during hour $h$ of day $d$:

$$v: \ S \times A \times D \times H \longrightarrow \mathbb{N}$$

For instance, in Figure 1, the gray cell indicates that $s_3$ retweeted $a_4$ 40 times on day $d_1$ at hour $h_1$. For the sake of clarity, in the following we will refer to $(s, a, (d, h))$ as $(s, a, d, h)$.

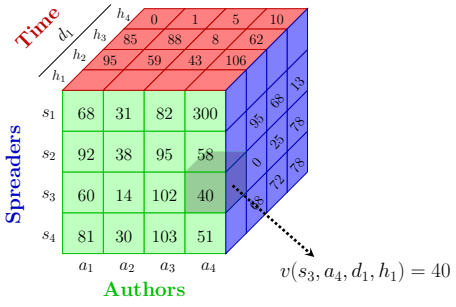


Figure 1: **Base Cuboid** $\mathcal{C}_4(S \times A \times D \times H, v)$

## 3.2 Data Cube Operations

We can explore the data through three operations called aggregation, expansion and filtering.

**Aggregation** is the operation which consists in seeing information at a more global level. Given a data cube $\mathcal{C}_n(X, f)$, the aggregation operation along the dimension $X_i$ leads to a data cube of dimension $n - 1$, $\mathcal{C}_{n-1}(X', f)$ where $X' = X_1 \times ... \times X_{i-1} \times X_{i+1} \times ... \times X_n$. Formally, a dimension $X_i$ is aggregated by adding up feature's values for all elements $x_i \in X_i$. We indicate by a " $\cdot$ " a dimension which is aggregated with respect to $f$. Hence, $\mathcal{C}_{n-1}(X', f)$

is constituted of $n-1$-dimensional cells denoted $c' = (x_1, ..., x_{i-1}, \cdot, x_{i+1}, ..., x_n) \in X'$ where

$$f(c') = \sum_{x_i \in X_i} f(c)\,.$$

For instance, one can aggregate along the hour dimension such that $v(s, a, d, \cdot) = \sum_{h \in H} v(s, a, d, h)$ gives the total number of time $s$ retweeted $a$ during day $d$.

**Expansion** is the reverse operation which consists in seeing information at a more local level by introducing additional dimensions. Given a data cube $\mathcal{C}_n(X, f)$, the expansion operation on the dimension $X_{n+1}$ leads to a data cube of dimension $n+1$, $\mathcal{C}_{n+1}(X', f)$ where $X' = X \times X_{n+1}$.

**Filtering** is the operation which consists in focusing on one specific subset of data. Given a data cube $\mathcal{C}_n(X, f)$, the filtering operation leads to a subcube $\mathcal{C}_n(X', f)$ by selecting subsets of elements within one or more dimensions such that $X' = X'_1 \times ... \times X'_n$ with $X'_1 \subseteq X_1, ..., X'_n \subseteq X_n$.

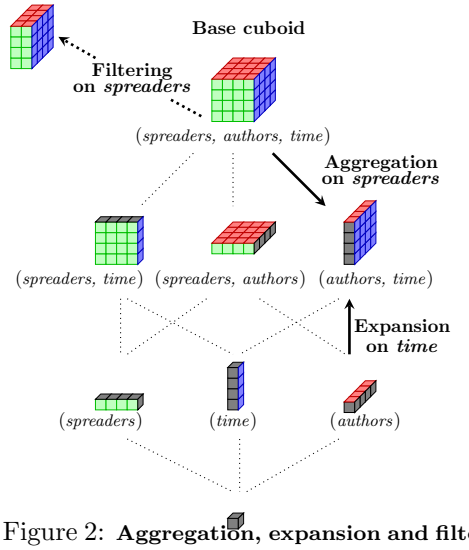


Figure 2: **Aggregation, expansion and filtering on the base cuboid.**

Figure 2 shows the set of more or less aggregated data cubes considering the three dimensions: spreaders, authors and time. It illustrates how to navigate from one to another thanks to the three previously described operations.

# 4 Method

In this paper, our goal is to find abnormal cells, *i.e.*, $n$-uplets $x \in X$ for which the observation $f(x)$ is abnormal. As an observation's abnormality is relative to the elements to which it is compared [3], a given cell may be abnormal or not depending on the *context*. More precisely, the context is the set of observations which are taken into account in order to assess the abnormality of a cell, we denote it $\mathcal{O} = \{o(x) \mid x \in X\}$. In this section, we design a set of steps in order to shape various contexts and show that it leads to a deeper exploration of interactions compared to an elementary outlier detection.

## 4.1 Basic Context

When seeking abnormal cells within a data cube $\mathcal{C}_n(X, f)$, the most elementary context we can consider is the set of raw observations $\mathcal{O} = \{f(x) \mid x \in X\}$. To find outliers, we infer the normal behaviour of observations $o \in \mathcal{O}$ and deduce abnormal behaviours which deviate from it.

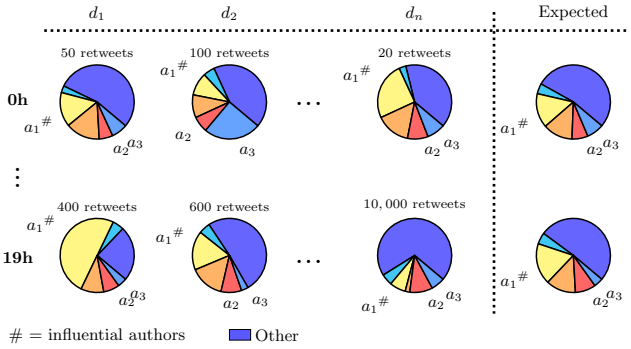


Figure 3: **Different contexts lead to different outliers -** Proportions $p(a, d, h)$ are represented as pie charts. For instance, on $d_1$ at $19h$, the influential author $a_1$ has been retweeted 200 times which represents 50% of all retweets exchanged during this hour.

***Example:*** *In data cube* $\mathcal{C}_3(A \times D \times H, v)$*, an abnormal cell* $c^* = (a^*, d^*, h^*)$ *indicates that during the hour* $h^*$ *of day* $d^*$*, the author* $a^*$ *has been retweeted an abnormal number of times compared to the number of times most authors are retweeted during one hour (independently of the hour of the day and of the day under consideration).*

## 4.2 Aggregated Context

The first way in which the context can be shaped is to observe quantities relatively to more aggregated quantities. Resulting contexts are called *aggregated contexts*. Contrarily to the basic context, building an aggregated context requires two data cubes: the cube under study, $\mathcal{C}_n(X, f)$ and the *comparison data cube*, $\mathcal{C}_m(X', f)$, which is used to provide comparative external elements to the first. For the context to be relevant, $\mathcal{C}_m(X', f)$ must derive from the aggregation of $\mathcal{C}_n(X, f)$ on one or more dimensions, hence, $n > m$ and $X = X' \times Y$ where $Y$ is the Cartesian product of the aggregated dimensions. We proceed as follows. For each cell $x = (x', y) \in X$ such that $x' \in X'$ and $y \in Y$, we measure the proportion, $p(x)$, between the quantity $f(x', y)$, within $\mathcal{C}_n(X, f)$, and the quantity $f(x')$, within $\mathcal{C}_m(X', f)$,

$$p(x) = \frac{f(x', y)}{f(x')} \ .$$

Then, as previously, we infer the normal behaviour of the set $\mathcal{O} = \{p(x) \mid x \in X\}$, and deduce abnormal behaviours which deviate from it.

***Example:*** *In data cube* $\mathcal{C}_3(A \times D \times H, v)$*, relatively to data cube* $\mathcal{C}_2(D \times H, v)$*, an abnormal cell* $c^* = (a^*, d^*, h^*)$ *indicates that the proportion of retweets received by author* $a^*$ *among all retweets of hour* $h^*$ *of day* $d^*$*,*

$$p(a^*, d^*, h^*) = \frac{v(\cdot, a^*, d^*, h^*)}{v(\cdot, \cdot, d^*, h^*)}$$

*is abnormal compared to most proportions of retweets received by authors during one hour (independently of the hour of the day and of the day under consideration). Figure 3 illustrates this situation with triplet* $(a_1, d_1, 19h)$*.*

## 4.3 Expected Context

The principle of the *expected context* is similar, except that this time, we compare a value to its expected value. For example, consider an author $a^*$ who presents a morning show:

$a^*$ has a proportion of retweets fluctuating between 10% and 15% every morning from $8h$ to $10h$. Outside this time slot, its proportions do not exceed 1%. On day $d^*$ at $19h$, we observe a proportion of 15%. With the aggregated context, we compare the proportions of retweets $p(a, d, h)$ for all triplets indifferently of the time of day. In this context, $(a^*, d^*, 19h)$ is considered as normal. However, by comparing the proportion of 15% to the expected proportion at $19h$ in the expected context, $(a^*, d^*, 19h)$ is marked as an outlier.

The expected value, denoted $f_{exp}$, is obtained by averaging $f$ on one or more of its variables. Formally, let $Y$ be the Cartesian product of the averaged dimensions such that $X = X' \times Y$. For each $x' \in X'$, we have

$$f_{exp}(x') = \frac{1}{|Y|} \sum_{y \in Y} f(x', y) \, .$$

Subsequently, for each cell $x = (x', y) \in X$ such that $x' \in X'$ and $y \in Y$, we measure a distance $l(x)$ between $f(x', y)$, within $\mathcal{C}_n(X, f)$, and its expected value $f_{exp}(x')$. Then, we infer the normal behaviour of the set $\mathcal{O} = \{l(x) \mid x \in X\}$ and deduce abnormal behaviours which deviate from it. Note that, as discussed in the example above, this context can be combined with the aggregated context.

When the feature consists in counting the number of interactions of cell $x$, as $v(x)$, it can be modelled by a Poisson counting process of intensity $f_{exp}$ [10]. In this case, the distance $l(x)$ can be obtained as follows. If $f(x) \geq f_{exp}(x')$, we calculate the probability of observing a value $f(x)$ or more, knowing that we should have observed $f_{exp}$ on average. We denote this probability $q(\text{Pois}(f_{exp}) \geq f(x))$. By symmetry, we obtain $\mathcal{O} = \{l(x) \mid x \in X\}$ such that

$$l(x) = \begin{cases} - & \log(q(\text{Pois}(f_{exp}) \geq f(x)) \quad \text{if } f(x) \geq f_{exp}, \\ & \log(q(\text{Pois}(f_{exp}) < f(x)) \quad \text{if } f(x) < f_{exp}. \end{cases} \tag{1}$$

The logarithm is calculated for convenience in order to have a better range of value. Defined as such, $l(x)$ allows us to take into account the *significance*, to which a value deviates from its expected value: if it is is very unlikely, namely very high (resp. low) given $f_{exp}$, we will observe high positive (resp. negative) distances. On the contrary, if it is very likely, $l(x)$ will be close to 0. The Poisson counting process is the most simple and frequently used counting process, however, other choices can be made, see for instance the book of Fleming *et al.* [7].

***Example:*** *In data cube $\mathcal{C}_3(A \times D \times H, v)$, relatively to data cube $\mathcal{C}_2(D \times H, v)$, an abnormal cell $c^* = (a^*, d^*, h^*)$ indicates that the distance $l(a^*, d^*, h^*)$ between the proportion of retweets received by author $a^*$ among all retweets of hour $h^*$ of day $d^*$,*

$$p(a^*, d^*, h^*) = \frac{v(\cdot, a^*, d^*, h^*)}{v(\cdot, \cdot, d^*, h^*)}$$

*and its expected proportion $p_{exp}$ during this specific hour of the day $h^*$,*

$$p_{exp}(a^*, h^*) = \frac{1}{|D|} \sum_{d \in D} p(a^*, d, h^*) \, ,$$

*is abnormal compared to most distances observed for other triplets $(a, d, h) \in A \times D \times H$. Figure 3 illustrates this situation with triplet $(a_3, d_2, 0h)$.*

### 4.4 Restrained Context

Restrained context is another way to focus on local patterns. It consists in filtering observations to only compare a subset of the cells. First, we gather observations on the restrained set of cells $X' \subset X$; next, we infer their normal behaviour; finally, we deduce abnormal behaviours which deviate from it.

***Example:*** *In data cube $\mathcal{C}_3(A \times D \times H, v)$, relatively to data cube $\mathcal{C}_2(D \times H, v)$, and considering the set of triplets $(a, d, h) \in {I_a}^{\mathrm{C}} \times D \times H$ where ${I_a}^{\mathrm{C}}$ is the set of non-influential authors, an abnormal cell $c^* = (a^*, d^*, h^*)$ indicates that the proportion of retweets received by author $a^*$ among all retweets of hour $h^*$ of day $d^*$,*

$$p(a^*, d^*, h^*) = \frac{v(\cdot, a^*, d^*, h^*)}{v(\cdot, \cdot, d^*, h^*)}$$

*is abnormal compared to most proportions of retweets received by non-influential authors during one hour (independently of the hour of the day and of the day under consideration). This is what we observe with triplet $(a_3, d_2, 0h)$ in Figure 3.*

Taken separately, each of these contexts allows to study interactions under a different perspective. In our method, we combine together different contexts which leads to numerous kinds of outliers.

## 5 Experiments

In this section, we apply our method on retweets related to political communication during the 2017 French presidential elections. We use a subset of the dataset collected by Gaumont *et al.* as part of the project *Politoscope* [8]. It contains politics-related retweets during the month of August 2016. Formally, our dataset consists in the set of retweets $E$, such that $(s, a, t) \in E$ means that $s$ retweeted $a$ at time $t$, where either the corresponding tweet contains politics-related keywords, or $a$ belongs to a set of $3,700$ French political actors listed by the Politoscope project. It contains $1,142,004$ retweets and involves $211,155$ different users. We present a case study, which, based on events found in the temporal dimension, proposes possible causes of their emergence by exploring the author dimension.

### 5.1 Events

We define an event to be an abnormal hour $(d^*, h^*) \in D \times H$. Figure 4 shows the evolution of the number of retweets per hour[1]. We can distinguish three distinct behaviours:
– nocturnal hours, characterized by a number of retweets fluctuating around 350,
– daytime from the $1^{st}$ of August to the $24^{th}$, characterized by a higher number of retweets fluctuating around $1,600$,
– daytime from $24^{th}$ of August to the $31^{th}$, characterized by a global increase in the number of retweets which fluctuates around $2,900$.

Based on such observations, if we look for abnormal hours in the basic context, *i.e.* for abnormal observations $o^* \in O$ such that $\mathcal{O} = \{v(\cdot, \cdot, d, h) \mid (d, h) \in D \times H\}$, extreme values

[1]Note that due to a server failure from Tuesday the $9^{th}$ to Thursday the $11^{th}$, no activity is observed during this period.

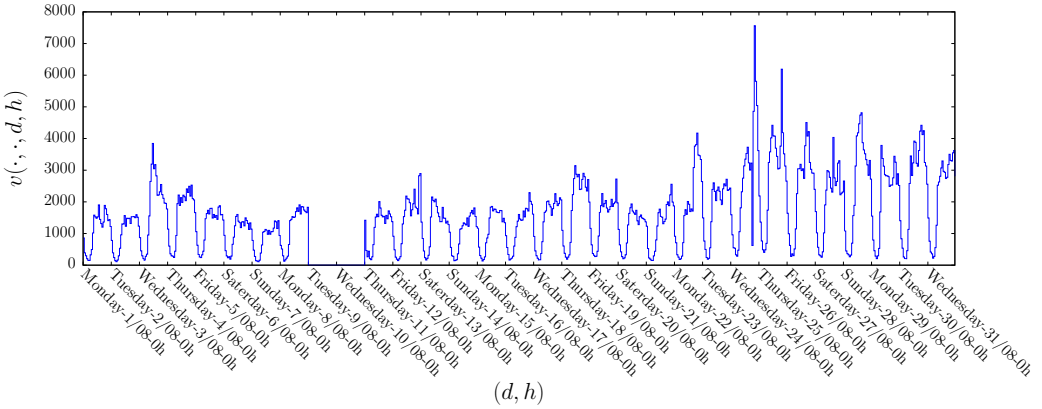


Figure 4: **Number of retweets per hour along the month of August 2016**

would only highlights trivial abnormalities which might only be related to the circadian rhythm as well as the overall trend of the month.

To detect more subtle and local events, we can consider the aggregated and expected context. Indeed, with the aggregated context we normalize each quantity of interaction per hour by the total number of retweet of the day. With the expected context, on the other hand, we consider each of these proportions with respect to its expected value at a given hour. Then, the resulting abnormal hours are independent of daily variations as well as the time of the day. Formally, in data cube $\mathcal{C}_2(D \times H, v)$, relatively to data cube $\mathcal{C}_1(D, v)$, we consider a cell $c^* = (d^*, h^*)$ to be abnormal if the distance, $l(d^*, h^*)$ (see eq. 1), between the proportion of retweets observed during hour $h^*$ among all retweets of day $d^*$,

$$p(d,h) = \frac{v(\cdot,\cdot,d,h)}{v(\cdot,\cdot,d,\cdot)},$$

and its expected proportion $p_{exp}$ during this specific hour of the day $h^*$,

$$p_{exp}(h^*) = \frac{1}{|D|} \sum_{d \in D} p(d,h^*) \ ,$$

is abnormal compared to most distances observed for other hours $(d,h) \in D \times H$.

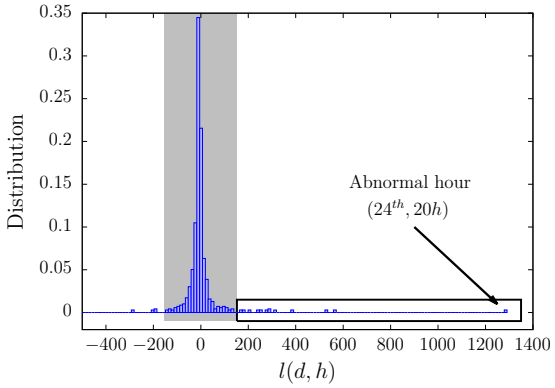


Figure 5: **Abnormal hours in the aggregated and expected context.**

Figure 5 shows the distribution of the set of observations $\mathcal{O} = \{l(d,h) \mid (d,h) \in D \times H\}$. As expected, most observations $o \in \mathcal{O}$ follow a normal distribution centred on $\mu = 0$ (gray zone), whereas some significantly deviates from it. This means that most proportions are likely to be generated by a Poisson counting process of intensity $p_{exp}(d,h)$ while others are not. Given a normal distribution with outliers, we use here the classical assumption that a value is anomalous if its distance to the mean $\mu$ exceeds three

times the standard deviation $\sigma$ [3], [11]. In this paper, we are only interested in hours during which the distance is higher than expected, thus, we mark an observation $o^* \in O$ as an outlier if $o^* > \mu + 3\sigma$.

We find 15 abnormal hours for which the proportion of retweets behave unusually:

$$\begin{aligned}\mathcal{O}^* = \quad & \{(3^{th}, 11h),\ (12^{th}, 23h),\ (21^{th}, 21h),\ (22^{th}, 17h),\ (22^{th}, 18h),\ (22^{th}, 19h),\\ & (24^{th}, 20h),\ (24^{th}, 21h),\ (24^{th}, 22h),\ (25^{th}, 19h),\ (26^{th}, 16h),\ (27^{th}, 15h),\\ & (28^{th}, 14h),\ (28^{th}, 15h),\ (29^{th}, 8h)\}.\end{aligned}$$

We can notice that events on the $22^{th}$, $24^{th}$ and $28^{th}$ of August span over multiple hours. More importantly, contrary to an analysis based on the basic context, we find hours within the first three weeks of August as well as hours of low activity such as $(29^{th}, 8h)$ and $(12^{th}, 23h)$.

## 5.2 Abnormal authors during events

Now, we focus in determining whether an hour's abnormality is due to specific authors, which have been retweeted predominantly, or, on the contrary, results from a more global phenomenon. To do so, we study interactions in a restrained context by considering the entities $(a, d, h) \in A \times T^*$, where $T^* \subseteq \mathcal{O}^*$. For the same reasons as above, we use the aggregated and expected context.

In this aggregated, expected and restricted context, in data cube $\mathcal{C}_3(A \times T^*, v)$, relatively to data cube $\mathcal{C}_2(T^*, v)$, we consider a cell $c^* = (a^*, d^*, h^*)$ to be abnormal if the distance, $l(a^*, d^*, h^*)$ (see eq. 1), between the proportion of retweets received by author $a^*$ during hour $(d^*, h^*)$ among all retweets of hour $(d^*, h^*)$,

$$p(a^*, d^*, h^*) = \frac{v(\cdot, a^*, d^*, h^*)}{v(\cdot, \cdot, d^*, h^*)}$$

and its expected proportion $p_{exp}$ during this specific hour of the day $h^*$,

$$p_{exp}(a^*, h^*) = \frac{1}{|D|} \sum_{d \in D} p(a^*, d, h^*)\ ,$$

is abnormal compared to most distances of other triplets $(a, d, h) \in A \times T^*$.

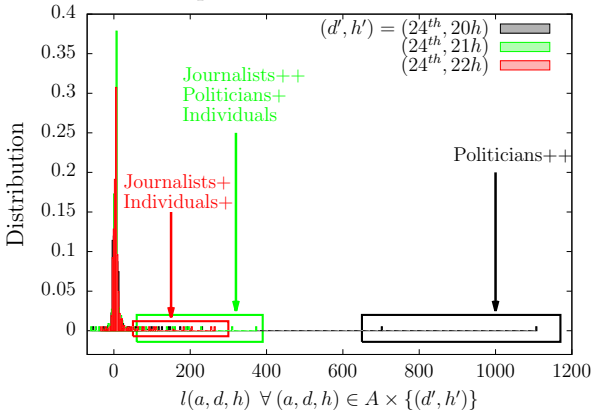


Figure 6: **Evolution of abnormal authors on the $24^{th}$ of August from $20h$ to $22h$.**

Figure 6 displays the distribution of the set $\mathcal{O} = \{l(a, d, h) | (a, d, h) \in A \times \{(24^{th}, h')\}\}$, where $h'$ successively takes the values $20h$, $21h$ and $22h$. At $20h$, all observations $o \in \mathcal{O}$ follow a Gaussian distribution centred on $\mu = 0$, except two authors which stand out strongly from other: *NicolasSarkozy*, a candidate to the election and *TTpourlaFrance*, his campaign slogan. This means that on the $24^{th}$ at $20h$, most authors behave the way they are expected to at $20h$ on other days. Abnormal au-

thors, in contrast, are much more retweeted than they usually are at $20h$. At $21h$, the set of values is more homogeneous, there are more outliers, but less significant. In opposition to the previous hour, this hour's abnormality is not solely due to a few authors, considerably retweeted, but to numerous authors retweeted on a smaller scale. Among these, we see many journalists, politicians supporting Nicolas Sarkozy are still very present, and individuals start to appear. At 22h, values spread over an even smaller range. This time, outliers only consist in journalists and individuals. On the news, this event corresponds to an interview of Nicolas Sarkozy on television news at $20h$.

The study of this event with our method enable us to illustrate political communication via Twitter. The more time passes, the more distributions are homogeneous. This shows that the event becomes a global phenomenon as information spreads: first, politicians tweet and are retweeted during the interview; one hour later, journalists propagate their analyses and are rewteeted; information reach individuals which start to tweet and being retweeted as well; finally, at $22h$, information reaches a larger scale and more and more individuals react and get retweeted.

# 6 Conclusion and Future Work

In this paper, we provided a method to explore temporal interactions and find outliers in a multitude of different situations. We applied it on a set of politics-related retweets and showed that it successfully highlights events as well as abnormally retweeted users. Section 5 only presents a small part of the extent of possibilities offered by our method. For instance, we could continue our study and look into the spreaders dimension to explore the cause of an author's emergence. More generally, we could split the authors or spreaders dimension into sub-dimensions according to their political leaning. Using restrained contexts, this would allow us to study the behaviour of each community separately as well as communities interactions. Also, we could include additional semantic dimensions. For instance, we could consider 5-uplets $(s, a, d, h, k)$ meaning that $s$ retweeted a tweet written by $a$ and containing the hashtag $k$ at time $(d, h)$. Applying similar contexts along the hashtag dimension would give us a lot more details on events' content. Moreover, it would include tweets' content needed by numerous work in this domain. Finally, the reaction to a television show through Twitter, as with Nicolas Sarkozy's interview, shows that this study could be interesting in researches aiming at characterizing the use of a second web-connected screen while watching television, see for instance the work of Gil de Zúñiga *et al.* [9].

# Acknowledgement

This work is funded in part by the European Commission H2020 FETPROACT 2016-2017 program under grant 732942 (ODYCCEUS), by the ANR (French National Agency of Research) under grants ANR-15- E38-0001 (AlgoDiv), by the Ile-de-France Region and its program FUI21 under grant 16010629 (iTRAC).

# References


[1] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1):4, 2015.

[2] A. Bruns, J. E. Burgess, K. Crawford, and F. Shaw. # qldfloods and@ QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods. 2012.

[3] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[4] N. Chavoshi, H. Hamooni, and A. Mueen. Temporal patterns in bot activities. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1601–1606. International World Wide Web Conferences Steering Committee, 2017.

[5] F. Chierichetti, J. M. Kleinberg, R. Kumar, M. Mahdian, and S. Pandey. Event Detection via Communication Pattern Analysis. In *ICWSM*, 2014.

[6] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5):1374–1405, 2015.

[7] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

[8] N. Gaumont, M. Panahi, and D. Chavalarias. Reconstruction of the socio-semantic dynamics of political activist Twitter networks-Method and application to the 2017 French presidential election. *PLoS ONE*, 13(9), 2018.

[9] H. Gil de Zúñiga, V. Garcia-Perdomo, and S. C. McGregor. What is second screening? exploring motivations of second screen use and its effect on online political participation. *Journal of Communication*, 65(5):793–815, 2015.

[10] C. Grasland, R. Lamarche-Perrin, B. Loveluck, and H. Pecout. International agenda-setting, the media and geography: A multi-dimensional analysis of news flows. *L'Espace géographique*, 45(1):25–43, 2016.

[11] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[12] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. Tedas: A twitter-based event detection and analysis system. In *Data engineering (icde), 2012 ieee 28th international conference on*, pages 1273–1276. IEEE, 2012.

[13] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr. Characterizing and Detecting Hateful Users on Twitter. *arXiv preprint arXiv:1803.08977*, 2018.

[14] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[15] J. Song, S. Lee, and J. Kim. Spam filtering in twitter using sender-receiver relationship. In *International workshop on recent advances in intrusion detection*, pages 301–317. Springer, 2011.

[16] S. Stieglitz and L. Dang-Xuan. Political communication and influence through microblogging–An empirical analysis of sentiment in Twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 3500–3509. IEEE, 2012.

[17] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*, 2017.

[18] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *European conference on information retrieval*, pages 356–367. Springer, 2013.

[19] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, 28(8):2158–2172, 2016.