# Scalable Statistical Tools for Social Data Analysis


**Rosa E. Lillo**
Department of Statistics
Universidad Carlos III de Madrid


**Engineering group** (A. Azcorra, R. Cuevas, A. Fernández, L. Chiroque)
**Statistical group** (H. Laniado, R. Lillo, J. Romo, C. Sguera)

**Universidad Eafit**
Medellín, Colombia
May 31, 2016

# What is the problem to be solved?

- **Online Social Networks** (OSNs) such as Facebook, Twitter or Google+ have rapidly become the most popular online services. (Hundreds of millions of users intensively interact every day).
- OSNs have an invaluable channel of information for different sectors such as advertising, marketing or politics.
- Important unsolved problem: **the identification of relevant users**.
- **Why?** They will be the users to be addressed in order to advertise a product, propagate a message, improve the image of a company,...

# Background of the topic

- The research community in OSNs is focusing on identifying metrics that best define influential users.
- Most existing works pre–define the properties of the target users to be found, and based on such definition, they establish ad–hoc mechanisms to find the target users. **(Supervised techniques)**
- Two main drawbacks:
  1. They require a considerable manual analysis of the problem and the data.
  2. Their effectiveness is fully tied with the definition of the target users' profile. **(Results would be likewise inaccurate or incorrect)**.
- **General Objective**:Unsupervised methods for the detection of relevant users are required to advance in the state–of–the–art of this important field.

# Our Big Data set

- We have a dataset of 10 million Google+ users and their associated public activity during two years (Jun 2011-July 2013). (González et al. (2015))

- Each user (or agent) is represented by 23 different variables covering connectivity, activity and user profile information including:

  1. **Number of followers**: it characterizes the popularity of a user.
  2. **Number of published posts**: it characterizes the level of activity of a user in the network.
  3. **Number of received likes, reshares and comments to the users' posts**: They characterize the influence capacity of a user to create engagement.

- We have removed all users in our dataset with less than 10 public posts over a period of 2 years. (They are "**consumers**" but not relevant).

- Final size of the dataset after applying the filtering: **5.619.786 users**.

# From the dataset to a statistical challenge

- When multivariate data have more than three dimensions, it is practically impossible to graphically visualize the observations using Cartesian coordinates.
- **Convenient alternative**: parallel coordinates (Wegman (1990)).

**A multivariate point $\equiv$ a series of points in the plane connecting each pair of adjacent points by a line.**

- Once represented by means of parallel coordinates, observations $\mathbf{x} \in \mathbb{R}^d$ can be seen as **real functions** defined on an arbitrary set of equally spaced domain points, e.g., $\{1, \ldots, d\}$, and $\mathbf{x}$ can be expressed as $x = \{x(1), \ldots, x(d)\}$. (López-Pintado and Romo (2009)).

# From the dataset to a statistical challenge

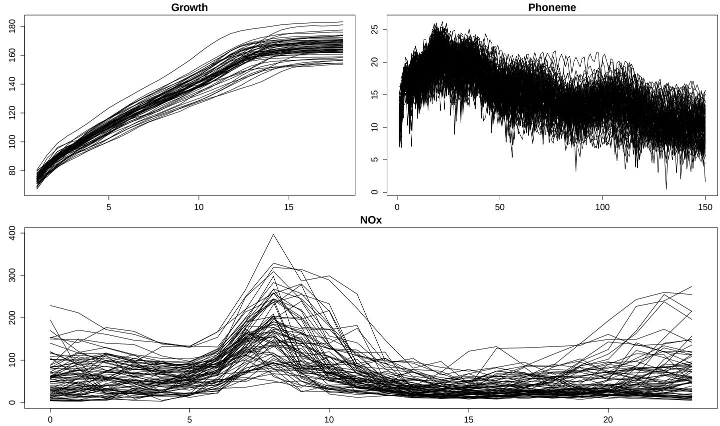
- Observations can be represented as curves $\Longrightarrow$ We can use the tools provided by an area of statistics known as **Functional data analysis** (FDA) (Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) or Cuevas (2014)).

- In the FDA framework, it is common to assume that:
  - Observations are generated by a functional random variable $X \in \mathbb{F}$, where $\mathbb{F}$ is a functional space.
  - Or $X$ is as a stochastic process $\{X(t), t \in I\}$, where $I$ is an interval in $\mathbb{R}$.

- Three functional **real datasets**:
  1. **Growth data (girls)**: growth curves of 54 heights of girls measured at a common discretized set of 31 nonequidistant ages between 1 and 18 years.
  2. **Phoneme data ("aa")**: 100 log-periodograms of length 150 corresponding to recordings of speakers pronouncing the phoneme "aa".
  3. **$NO_x$ data (working days)**: 76 nitrogen oxides ($NO_x$) emission level daily curves measured every hour near to an industrial area in Poblenou (Barcelona).

# Functional data examples

**Figure**: growth data (top left), phoneme data (top right), $NO_x$ data (bottom)



# Who is a relevant user in the dataset?

**An atypical observation $\equiv$ Outlier**

- **Our proposal**: **Relevant users in OSNs can be viewed as outliers in FDA**.

  (They usually show behaviors and patterns that are different from the ones of non–relevant commons users)

- Our methodology can be used to search for potentially relevant Google+ users, whose identification will be based on a statistical criterion but not by directed arguments. **(Unsupervised)** (Cha et al. (2010); Bakshy et al. (2011); Simmie et al. (2014); Basaras et al. (2013)).

**BUT...**

# What is an outlier in FDA?

- **Formal definition?:** An outlier can be defined as an observation generated by a functional random variable with a different distribution from the one generating the normal observations of a functional sample (Febrero et al. (2008)).

- We focus on the three types of persistent outliers defined by Hubert et al. (2015):
  1. **Shift/magnitude outliers** $\equiv$ those who have the same shape of the majority but are moved away.
  2. **Amplitude outliers** $\equiv$ curves that may have the same shape as the majority but their scale differs.
  3. **Shape outliers** $\equiv$ curves whose shape differs from the majority.

# What is an outlier in FDA?

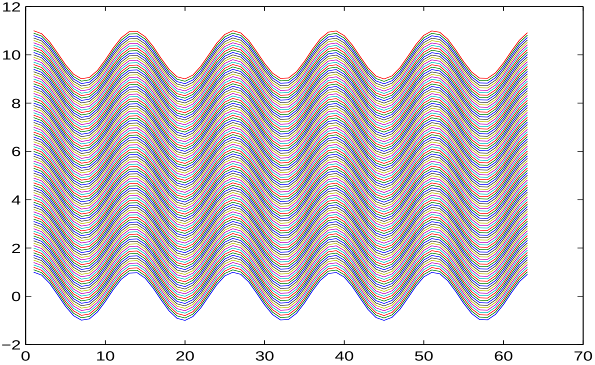



Figure: Functional sample without outliers.

# What is an outlier in FDA?

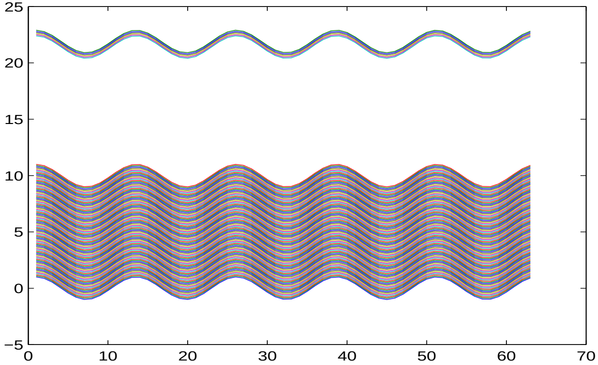



Figure: Functional sample with magnitude outliers.

# What is an outlier in FDA?

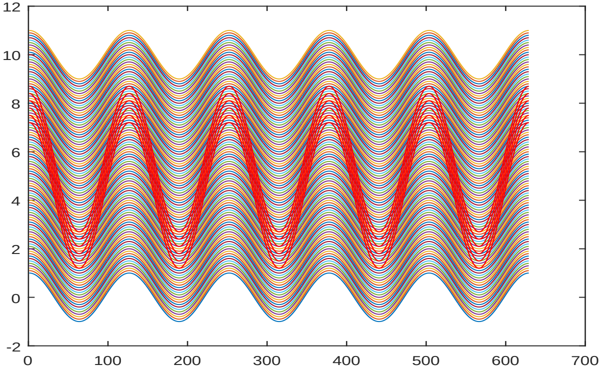



Figure: Functional sample with amplitude outliers.

# What is an outlier in FDA?

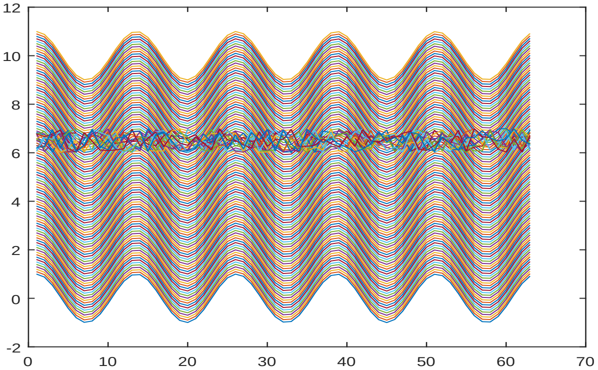



Figure: Functional sample with shape outliers.

# What is an outlier in FDA?

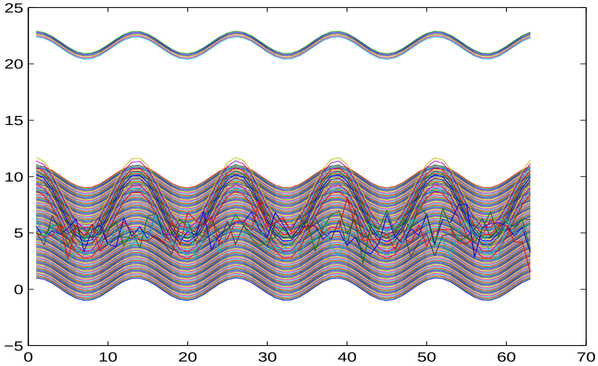


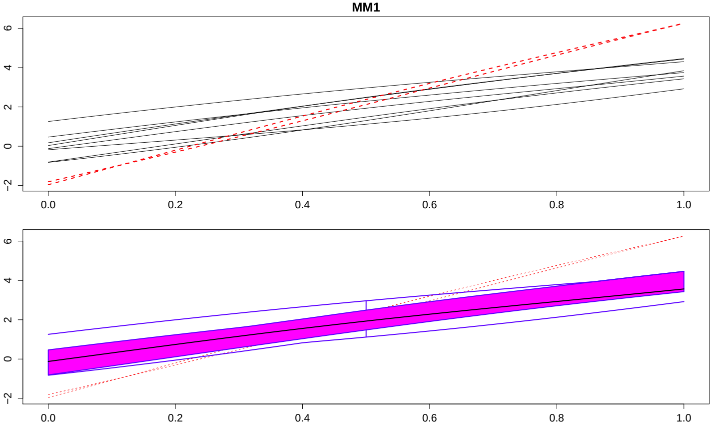

Figure: Functional sample with magnitude, amplitude and shape outliers.

# Outlier detection in functional data

- There are several methods to detect an outlier in FDA.
- Some of them are based on the use of measures known as **functional depths**: A measure that allows to order and rank the observations in a functional sample from the most to the least central.
  - High values to central observations.
  - Low values to non-central observations.
- Unlike univariate statistics where $\mathbb{R}$ provides a natural order criterion for observations, several criteria have been employed to order functional data $\Longrightarrow$ there exist different implementations of the notion of functional depth (see Sguera et al. (2014)).

# Our competitors: Functional boxplot

**Functional boxplot** (*FBPLOT*, Sun and Genton (2011)): 50%-central region (smallest band containing at least half of the deepest curves) factor non-outlying region = 1.5, functional depth = Modified band depth.



# Our competitors: Two bootstrap-based procedures

- Febrero et al. (2008) proposed two depth-based outlier detection procedures selecting a **threshold** for the h-modal depth (Cuevas et al. (2006)).

- The threshold is obtained through **two alternative robust smoothed bootstrap procedures** whose single bootstrap samples are obtained using:
  1. $B_{tri}$: the resampling is done on a trimmed version of the original sample, that is, after deleting from the sample a given proportion of least deep curves (trimmed resampling).
  2. $B_{wei}$: the resampling is done giving weights to sample observations that are proportional to their depth values (weighted resampling).

- At each bootstrap sample, the 1% percentile $p_{0.01}$ of the empirical distribution of the depth values is obtained.

- Let $B$ be the number of bootstrap samples:

$$\text{threshold} \rightarrow \text{median of the } B\text{-sized collection of } p_{0.01}$$

- Except for the computation of the threshold, both procedures are iterative.

# Our competitors: Robust FPCA-based procedures

- *FBAG*, **Functional Bagplot** (Hyndman and Shang (2010)):

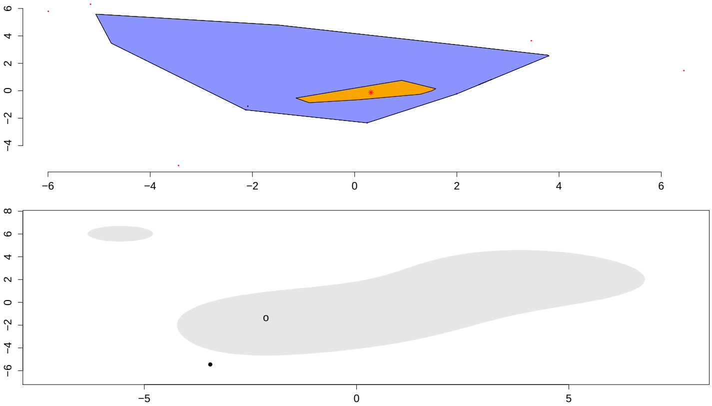
  1. Reduces the outlier detection problem from functional to multivariate by means of the functional principal component analysis technique.
  2. Once obtained the first two functional principal components scores, *FBAG* orders the scores using the multivariate halfspace depth (Tukey (1975)) and builds a non-outlying region.
  3. *FBAG* **detects as outliers those observations whose scores are outside the non-outlying region**.

- *FHDR* **Functional hig densisty region boxplot** (Hyndman and Shang (2010)):

  1. Procedure that differs from *FBAG* after obtaining the first two functional principal components scores.
  2. *FHDR* performs a bivariate kernel density estimation on the scores and defines a high density region.
  3. *FHDR* **detects as outliers those observations whose scores are outside the high density region**.

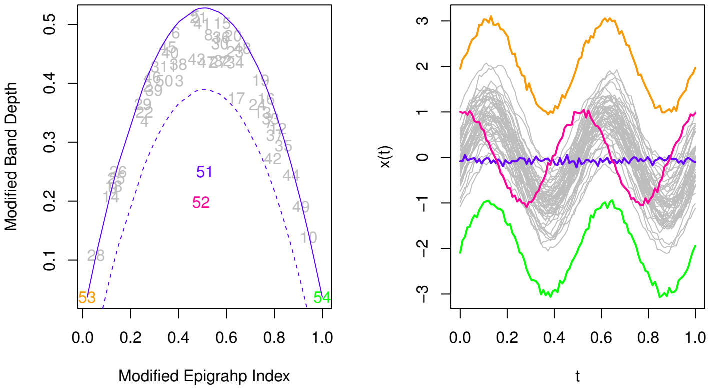
# Our competitors: Robust FPCA-based procedures

**Functional bagplot** (FBAG): 50%-central region, factor non-outlying region = 2.58, bivariate depth = halfspace depth (top);
**Functional high density region boxplot** (FHDR): 90%-high density region (bottom)



**Outliergram** (*OG*, Arribas-Gil and Romo (2014)): depth-based outlier detection method based on a visualization tool known as outliergram.

*OG* exploits the relation between the modified band depth (López-Pintado and Romo (2009)) and the modified epigraph index (López-Pintado and Romo (2011)) to help understanding shape features of observations.



# Our competitors: Probabilistic methods

- $KFSD_{smo}$, $KFSD_{tri}$ and $KFSD_{wei}$ (Sguera et al. (2015)): depth-based outlier detection methods which select the threshold of the kernelized functional spatial depth (KFSD, Sguera et al. (2014)) by means of a probabilistic procedure based on three alternative resampling techniques that differ in their resampling steps:

    1. $KFSD_{smo}$: the resampling is simple and smoothed, that is, once an observation is sampled, a small perturbation is added to the observation to avoid repeated observations.
    2. $KFSD_{tri}$: the resampling is trimmed and smoothed.
    3. $KFSD_{wei}$: the resampling is weighted and smoothed.

# What is the problem of these methods for Big Data?

**They are not scalable!!**

- We have tested all these methods with random samples of our dataset in order to observe the time performance.
- Experiments have been carried out in an AMD Opteron 6276 x64 cores @ 2.3GHz with 512GiB of RAM under Debian 7.9.
- We ran our method with one single partition, and using 10 partitions in order to check the scalability and verified that the time performance decreased by one order of magnitude.

# What is the problem of these methods for Big Data?

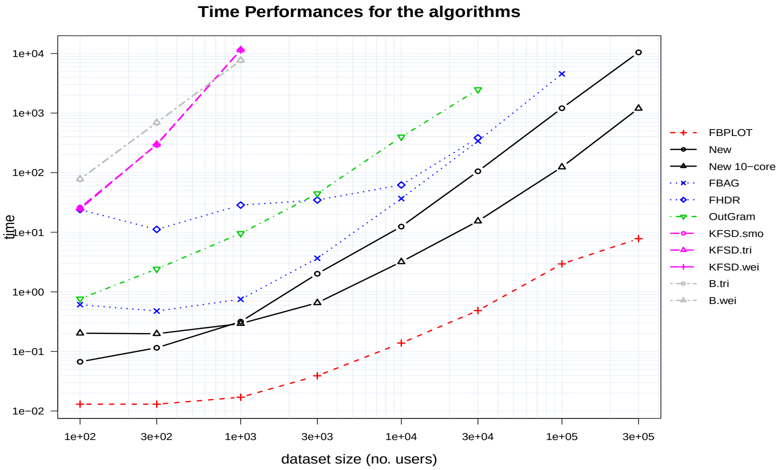



Figure: Time performance for the different algorithms (log-log scale).The new method appears twice, with 1 core and 10 cores

# A new outlier detection method

We introduce **three indexes** that can be interpreted as similarity measures of an observation with respect to a sample, and each one of them focus on a different feature of the data: magnitude, amplitude or shape.

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be a set of $n$ curves whose common discretized form is defined on a given set of $d$ equally spaced domain points, and $x$ be another curve defined on the same set.

- **The shape index** of $x$ with respect to $\mathcal{X}$ is defined as

$$I_S(x, \mathcal{X}) = \left| \frac{1}{n} \sum_{j=1}^{n} \rho(x, x_j) - 1 \right|,$$

where $\rho(x, x_j)$ is the Pearson correlation coefficient between the discretized versions of $x$ and $x_j$.

# The shape index

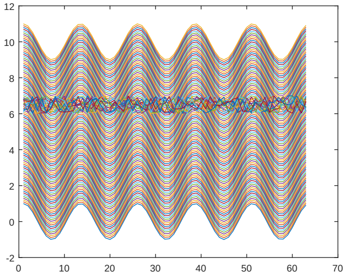



Figure: Functional sample with shape outliers.

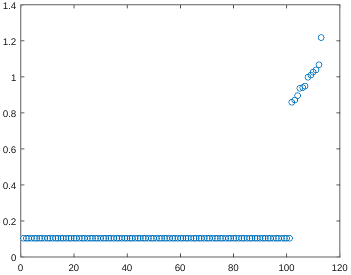



Figure: $I_S(x, \mathcal{X})$-based ranks versus $I_S(x, \mathcal{X})$ values.

# The magnitude and the amplitude index

Let $\alpha_j$ and $\beta_j$ be the estimated intercept and the slope of a linear regression model where the discretized version of $x$ represents the observed values of the dependent variable and the discretized version of $x_j$ represents the observed values of the regressor.

- We define the **magnitude index** of $x$ with respect to $\mathcal{X}$ as

$$I_M(x, \mathcal{X}) = \left| \frac{1}{n} \sum_{j=1}^{n} \alpha_j \right|,$$

- And the **amplitude index** of $x$ with respect to $\mathcal{X}$ as

$$I_A(x, \mathcal{X}) = \left| \frac{1}{n} \sum_{j=1}^{n} \beta_j - 1 \right|.$$

# The magnitude index

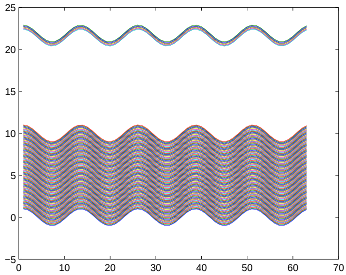



Figure: Functional sample with magnitude outliers.

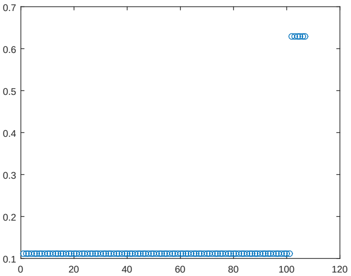



Figure: $I_M(x, \mathcal{X})$-based ranks versus $I_M(x, \mathcal{X})$ values.

# The amplitude index

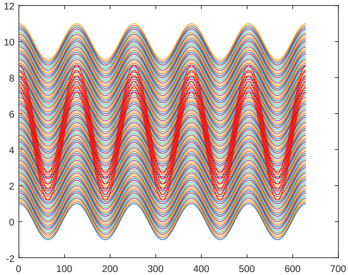



Figure: Functional sample with amplitude outliers.

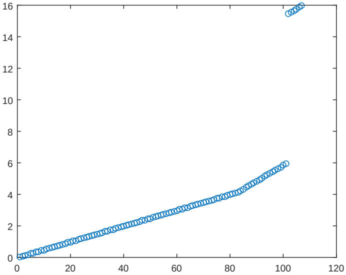



Figure: $I_A(x, \mathcal{X})$-based ranks versus $I_A(x, \mathcal{X})$ values.

# Next problem: Which are the outliers curves?

- Normalize the indexes as follows. Let $I_S(\mathcal{X}) = \{I_S(x_1, \mathcal{X}), \cdots, I_S(x_n, \mathcal{X})\}$ be the vector of the shape indexes and, analogously, let $I_M(\mathcal{X})$ and $I_A(\mathcal{X})$ be the vectors of the magnitude and amplitude indexes respectively. Hereafter we will use $I(\mathcal{X})$ for any of the three vectors of indexes indistinctly. We use the $\infty\text{-norm}$ for vectors and we define

$$\hat{I}_{\mathcal{X}} = \frac{I(\mathcal{X})}{||I(\mathcal{X})||_\infty} = \left\{ \frac{I(x_1, \mathcal{X})}{||I(\mathcal{X})||_\infty}, \cdots, \frac{I(x_n, \mathcal{X})}{||I(\mathcal{X})||_\infty} \right\},$$

  where $\hat{I}_{\mathcal{X}}$ is the normalized vector of indexes and $|| \cdot ||_\infty = \max(\cdot)$.

- Normalization $\Longrightarrow$ using $\hat{I}(\mathcal{X}) \in [0, 1]$.

- Define the following function $f$.

$$\begin{aligned} f \colon \{1..|\hat{I}_{\mathcal{X}}|\} &\rightarrow \hat{I}_{\mathcal{X}} \\ f(i) &\mapsto \hat{I}_{\mathcal{X}}[i], \end{aligned}$$

  where $\hat{I}_{\mathcal{X}}[i]$ is the index ranked in position $i$ in increasing order.

# Next problem: Which are the outliers curves?

- Define the *backward difference* for $f$ as

$$\nabla_h[f](i) := f(i) - f(i-h).$$

  Thus, we can establish the relationship between the derivative definition and the backward difference since

$$f'(i) = \lim_{h \to 0} \frac{f(i) - f(i-h)}{h} \equiv \lim_{h \to 0} \frac{\nabla_h[f](i)}{h}.$$

- Finally, we have computed the derivative function $f'$ for our curve $f$ and we are going to filter those values above a certain threshold value.
  - Given the threshold $\theta$, it represents the maximum slope allowed for the derivative to be considered a "normal" value.
  - Otherwise, the derivative points (onwards) above this threshold are considered outliers.

$$\textbf{Set of outliers} \equiv I_{\mathcal{X}}^{\textbf{out}} = \{I_{\mathcal{X}}[j] : f(j) > f(i_\theta)\}$$

# Simulation study

- We compare our methods with the competitors in functional outlier detection.
- **Important question**: Is our method competitive in the usual framework in FDA?
- For each model, 100 replications of size 100.
- Probability that each curve is an outlier ($\alpha = 0.05$)

# Comparison among methods

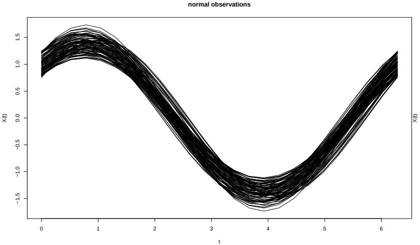
- **c** $\equiv$ Correct outlier detection percentages.
- **f** $\equiv$ False outlier detection percentages.
- **F–measure** $\equiv$ the harmonic mean of precision and recall.

$$F = \frac{2RP}{R + P},$$

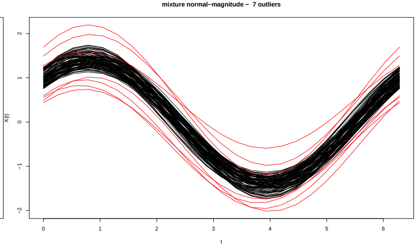
where $R = \frac{TP}{(TP+FN)}$ is known as recall measure, $P = \frac{TP}{TP+FP}$ is known as precision measure and $TP, FN$, and $FP$ are the number of true positive, false negative and false positive, respectively.

- **r** $\equiv$ F-measure-based rankings of the methods in the mixture models.

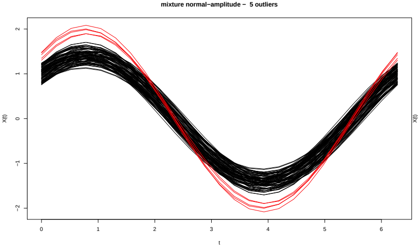
# Models



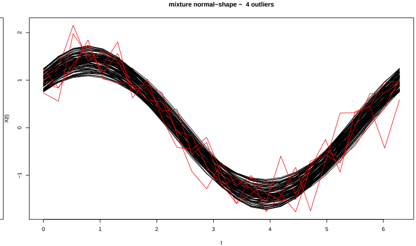

Normal observations




Magnitude outliers




Amplitude outliers



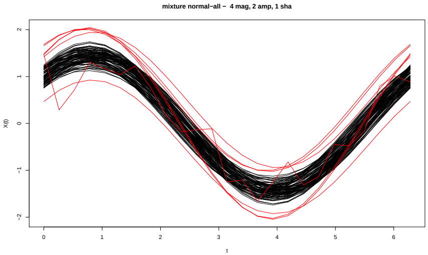

Shape outliers

# ...Summarizing indexes

Table: Correct outlier detection percentages (c), false outlier detection percentages (f), F-measures (F) and F-measure-based rankings of the methods (r) in mixture models 1, 2 and 3 which allow for magnitude (mag), amplitude (amp) and shape (sha) outliers, respectively.

| | mag | | | | amp | | | | sha | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | c | f | F | r | c | f | F | r | c | f | F | r |
| $B_{tri}$ | 54.55 | 0.00 | 0.71 | 6 | 16.67 | 0.01 | 0.29 | 9 | 83.82 | 0.00 | 0.91 | 2 |
| $B_{wei}$ | 98.42 | 0.05 | 0.98 | 1 | 25.00 | 0.01 | 0.40 | 8 | 100.00 | 0.00 | 1.00 | 1 |
| *FBAG* | 3.16 | 0.27 | 0.06 | 10 | 91.67 | 0.46 | 0.91 | 2 | 8.29 | 0.24 | 0.14 | 11 |
| *FHDR* | 15.61 | 4.43 | 0.16 | 9 | 75.97 | 1.14 | 0.77 | 6 | 24.08 | 3.96 | 0.24 | 10 |
| *FBPLOT* | 39.13 | 0.00 | 0.56 | 8 | 0.39 | 0.00 | 0.00 | 11 | 64.55 | 0.00 | 0.79 | 9 |
| *OG* | 0.00 | 0.00 | - | - | 0.78 | 0.00 | 0.02 | 10 | 0.00 | 0.00 | - | - |
| $KFSD_{smo}$ | 98.81 | 0.09 | 0.98 | 1 | 82.17 | 0.11 | 0.89 | 3 | 84.39 | 0.13 | 0.90 | 3 |
| $KFSD_{tri}$ | 99.60 | 2.51 | 0.81 | 4 | 96.90 | 2.35 | 0.81 | 5 | 99.23 | 2.45 | 0.81 | 6 |
| $KFSD_{wei}$ | 100.00 | 2.71 | 0.80 | 5 | 97.48 | 2.13 | 0.82 | 4 | 99.81 | 2.66 | 0.80 | 7 |
| *new* | 96.05 | 5.84 | 0.63 | 7 | 96.71 | 6.54 | 0.61 | 7 | 95.18 | 1.60 | 0.84 | 5 |
| $new_{mag}$ | 95.85 | 0.50 | 0.93 | 3 | 0.00 | 2.21 | - | - | 68.98 | 0.16 | 0.80 | 7 |
| $new_{amp}$ | 0.59 | 0.93 | 0.01 | 12 | 96.71 | 0.62 | 0.93 | 1 | 4.62 | 0.98 | 0.08 | 12 |
| $new_{sha}$ | 4.94 | 4.79 | 0.05 | 11 | 0.00 | 5.62 | - | - | 83.04 | 0.50 | 0.86 | 4 |

# Mixing types of outliers




| | all | | | |
|---|---|---|---|---|
| | c | f | F | r |
| $B_{tri}$ | 61.81 | 0.00 | 0.77 | 6 |
| $B_{wei}$ | 96.21 | 0.00 | 0.98 | 1 |
| *FBAG* | 35.32 | 0.26 | 0.50 | 9 |
| *FHDR* | 42.32 | 3.00 | 0.42 | 11 |
| *FBPLOT* | 34.92 | 0.00 | 0.52 | 8 |
| *OG* | 0.52 | 0.00 | 0.02 | 13 |
| $KFSD_{smo}$ | 82.67 | 0.14 | 0.89 | 2 |
| $KFSD_{tri}$ | 99.35 | 2.34 | 0.82 | 4 |
| $KFSD_{wei}$ | 99.80 | 2.51 | 0.81 | 5 |
| *new* | 97.58 | 2.01 | 0.83 | 3 |
| $new_{mag}$ | 41.33 | 0.08 | 0.57 | 7 |
| $new_{amp}$ | 34.01 | 0.37 | 0.48 | 10 |
| $new_{sha}$ | 30.22 | 1.61 | 0.37 | 12 |

# Mixing types of outliers

Table: Decomposed correct outlier detection percentages in mixture model 4 allowing simultaneously for magnitude (mag), amplitude (amp) and shape (sha) outliers.

| | mag | | | | amp | | | | shape | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | c | f | F | r | c | f | F | r | c | f | F | r |
| $B_{tri}$ | 73.08 | 1.92 | 0.52 | 3 | 21.40 | 2.84 | 0.14 | 9 | 89.98 | 1.65 | 0.63 | 1 |
| $B_{wei}$ | 100.00 | 3.23 | 0.52 | 3 | 88.60 | 3.49 | 0.45 | 4 | 99.80 | 3.27 | 0.52 | 4 |
| *FBAG* | 0.77 | 2.07 | 0.01 | 10 | 98.40 | 0.42 | 0.88 | 2 | 8.64 | 1.94 | 0.08 | 11 |
| *FHDR* | 8.65 | 4.94 | 0.04 | 9 | 98.00 | 3.42 | 0.49 | 3 | 22.00 | 4.71 | 0.11 | 10 |
| *FBPLOT* | 40.19 | 1.10 | 0.39 | 6 | 1.00 | 1.79 | 0.01 | 11 | 62.87 | 0.73 | 0.61 | 3 |
| *OG* | 0.00 | 0.03 | - | - | 1.60 | 0.00 | 0.04 | 10 | 0.00 | 0.03 | - | - |
| $KFSD_{smo}$ | 100.00 | 2.66 | 0.57 | 2 | 69.80 | 3.24 | 0.39 | 5 | 77.60 | 3.09 | 0.43 | 5 |
| $KFSD_{tri}$ | 100.00 | 5.64 | 0.39 | 6 | 98.40 | 5.74 | 0.37 | 7 | 99.61 | 5.69 | 0.37 | 6 |
| $KFSD_{wei}$ | 100.00 | 5.84 | 0.37 | 8 | 99.80 | 5.91 | 0.36 | 8 | 99.61 | 5.88 | 0.37 | 6 |
| *new* | 100.00 | 5.23 | 0.40 | 5 | 100.00 | 5.30 | 0.39 | 5 | 92.73 | 5.39 | 0.37 | 6 |
| $new_{mag}$ | 100.00 | 0.46 | 0.88 | 1 | 1.80 | 2.19 | 0.01 | 11 | 20.24 | 1.87 | 0.18 | 9 |
| $new_{amp}$ | 0.00 | 2.12 | - | - | 100.00 | 0.43 | 0.89 | 1 | 3.93 | 2.05 | 0.03 | 12 |
| $new_{sha}$ | 1.35 | 3.10 | 0.01 | 10 | 0.00 | 3.12 | - | - | 89.39 | 1.58 | 0.63 | 1 |

# Focusing on shape outliers

Models used in Arribas-Gil and Romo (2014) to evaluate *OG*.

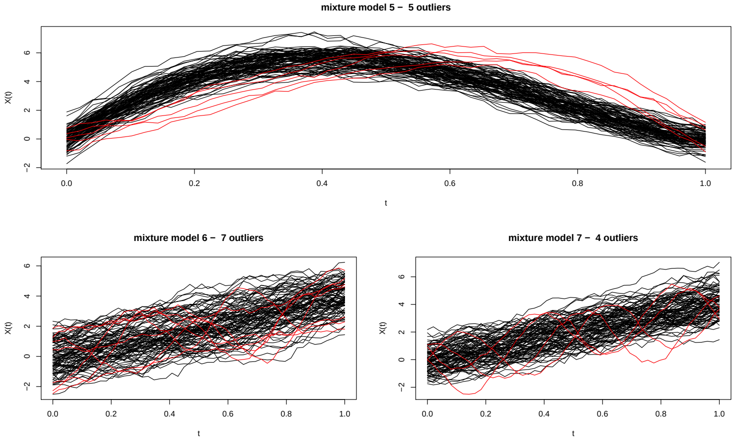



Figure: Mixture models 5 (top), 6 (bottom left) and 7 (bottom right).

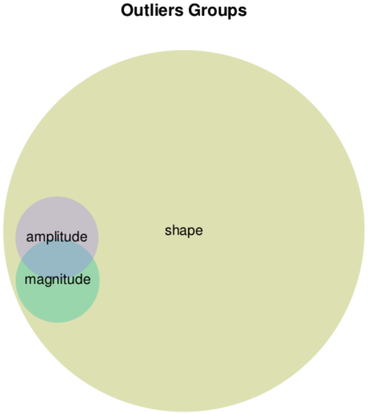
# Focusing on shape outliers

Table: Correct outlier detection percentages (c), false outlier detection percentages (f), F-measures (F) and F-measure-based rankings of the methods (r) in mixture models 5, 6 and 7.

| | mix mod 5 | | | | mix mod 6 | | | | mix mod 7 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | c | f | F | r | c | f | F | r | c | f | F | r |
| $B_{tri}$ | 48.23 | 0.03 | 0.65 | 10 | 26.20 | 0.59 | 0.38 | 8 | 23.55 | 0.79 | 0.34 | 8 |
| $B_{wei}$ | 88.08 | 0.41 | 0.90 | 2 | 30.59 | 0.71 | 0.43 | 7 | 23.55 | 0.76 | 0.35 | 7 |
| *FBAG* | 99.63 | 6.70 | 0.63 | 11 | 36.14 | 7.00 | 0.27 | 9 | 8.58 | 7.86 | 0.06 | 10 |
| *FHDR* | 65.74 | 1.55 | 0.68 | 8 | 23.71 | 3.97 | 0.24 | 10 | 5.59 | 4.97 | 0.06 | 10 |
| *FBPLOT* | 26.44 | 0.01 | 0.41 | 13 | 0.19 | 0.00 | 0.00 | 13 | 0.40 | 0.02 | 0.00 | 13 |
| *OG* | 97.95 | 2.31 | 0.82 | 4 | 98.85 | 3.36 | 0.76 | 1 | 100.00 | 3.92 | 0.73 | 1 |
| $KFSD_{smo}$ | 84.92 | 0.55 | 0.87 | 3 | 49.52 | 3.05 | 0.48 | 5 | 45.71 | 3.67 | 0.43 | 5 |
| $KFSD_{tri}$ | 98.14 | 4.36 | 0.71 | 7 | 79.54 | 6.29 | 0.54 | 4 | 82.04 | 6.33 | 0.55 | 3 |
| $KFSD_{wei}$ | 99.26 | 5.27 | 0.68 | 8 | 86.81 | 7.02 | 0.56 | 3 | 88.22 | 7.22 | 0.54 | 4 |
| *new* | 95.53 | 3.42 | 0.75 | 5 | 67.30 | 6.91 | 0.46 | 6 | 67.66 | 7.47 | 0.44 | 5 |
| $new_{mag}$ | 47.49 | 1.65 | 0.53 | 12 | 9.37 | 3.17 | 0.11 | 12 | 1.00 | 3.92 | 0.01 | 12 |
| $new_{amp}$ | 72.63 | 1.60 | 0.72 | 6 | 14.53 | 3.49 | 0.17 | 11 | 6.19 | 3.75 | 0.07 | 9 |
| $new_{sha}$ | 92.74 | 0.53 | 0.92 | 1 | 60.04 | 2.24 | 0.60 | 2 | 66.07 | 2.15 | 0.64 | 2 |

# Going back to the OSN problem

After applying the outlier detection method, we obtain 285.804, 4.270 and 4.434 "relevant" users based on the shape, amplitude, magnitude metrics.



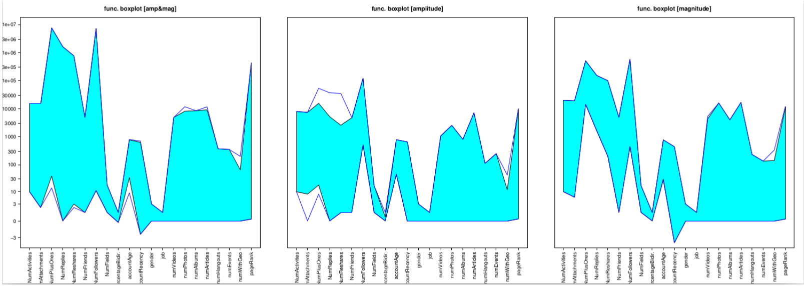



Figure: Functional boxplots

# Are the users detected as outliers different?

- In order to discuss the "relevance" of outliers, we rely on metrics measuring the ratio of number of reactions (likes, comments and shares) per activity (post) $\Longrightarrow$ capture the ability of a user to generate engagement.

  1. Our methodology is efficient since users identified in the three groups present 1 or 2 order of magnitude more reaction per activity than regular users. **(More engagement)**
  2. **The** *amp&mag group* **shows roughly one order of magnitude more reactions per activity than** *amp* **outliers.**
  3. **The difference is smaller when comparing** *amp&mag group* **vs.** *mag*.

# Are the users detected as outliers different?

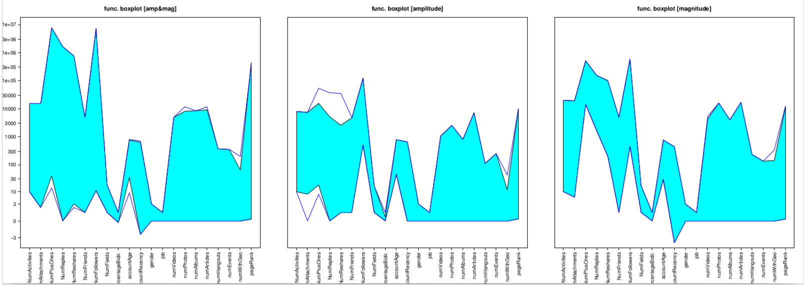
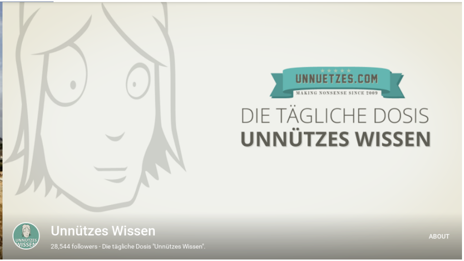



Figure: Functional boxplots

# Type of relevant users


Median amplitude

Median magnitude


Amplitude-magnitude outliers

# Conclusions

- We have converted a **Big Problem** in OSN to a **Statistical Problem with Big Data**.
- Relevant users are considered as outlier in Functional Data.
- We have introduced a new method to detect outliers that distinguishes amplitude, shape and magnitude outliers and besides; it is:
  1. **Competitive** respect to performance.
  2. **Scalable** for big data.
- The evaluation of our method in a real OSN dataset provides solid evidences about its ability to identify relevant agents in real cases.
- We obtain interesting results with semantic interpretation.

# References I


Arribas-Gil, A. and Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15:603–619.

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Identifying influencers on twitter. In *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*.

Basaras, P., Katsaros, D., and Tassiulas, L. (2013). Detecting influential spreaders in complex, dynamic networks. *Computer*, 46(4):24–29.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Cuevas, A., Febrero, M., and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51:1063–1074.

Febrero, M., Galeano, P., and González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, 19:331–345.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis : Theory and Practice*. Springer, New York.

González, R., Cuevas, R., Motamedi, R., Rejaie, R., and Cuevas, A. (2015). Assessing the evolution of google+ in its first two year. *IEEE/ACM Transactions on Networking*, Forthcoming:1–16.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data With Applications*. Springer, New York.

Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications*, 24:177–202.

# References II


Hyndman, R. J. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19:29–45.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734.

López-Pintado, S. and Romo, J. (2011). A half-region depth for functional data. *Computational Statistics and Data Analysis*, 55:1679–1695.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *TEST*, 23:725–750.

Sguera, C., Galeano, P., and Lillo, R. (2015). Functional outlier detection by a local depth with application to nox levels. *Stochastic Environmental Research and Risk Assessment*, Forthcoming:1–16.

Simmie, D., Vigliotti, M. G., and Hankin, C. (2014). Ranking twitter influence by combining network centrality and influence observables in an evolutionary model. *Journal of Complex Networks*, 2(4):495–517.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20:316–334.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of American Statistical Association*, 85:664–675.