Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks

Antonio Fernández Anta

Developing the Science of Networks

Team

- Universidad Carlos III de Madrid
  - Rubén Cuevas
  - Henry Laniado
  - Rosa E. Lillo
  - Juan Romo
  - Carlos Sguera
- IMDEA Networks Institute
  - Arturo Azcorra
  - Luis F. Chiroque
  - A.F.A.

Motivation

- Online Social Networks (OSNs) are used everyday by billions of people
- They are invaluable to extract information and to actuate in advertising, marketing, politics, etc.
- A recurring problem in OSNs analyses is to identify “interesting” or “influential” users
- Usually the characterization of influential users is given a priori, and algorithms to find these characteristics are proposed

Characterizing Influential Users

- Several characterization that have been used for influential OSN users:
  - Large number of followers [Cha HBG 2010][Pastor-Satorras Vespignani 2001] [Cohen EbAH 2001]
  - Capacity of engagement [Domingos Richardson 2001] [D'Agostino ANT 2015]
  - High infection capacity in an epidemic model [Kitsak GHLMSM 2010] [Morone Makse 2015] [Kempe Kleinberg Tardos 2015]
- Each of these characterizations may miss important interesting users
- They disregard many available attributes of the users

Contributions

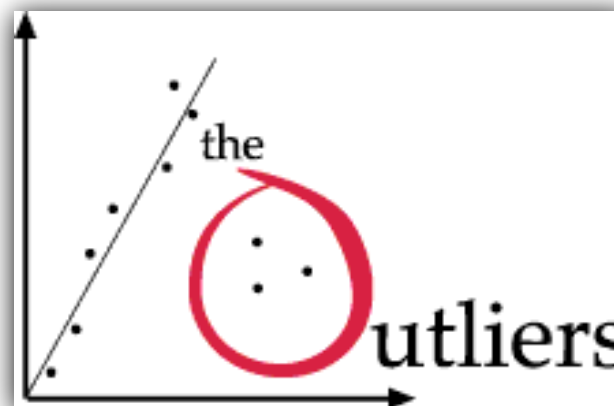
- We propose a new unsupervised method to identify "interesting" users: **Massive Unsupervised Outlier Detection (MUOD)**
- MOUD finds **outliers** in the multidimensional data available from the users
- These outliers can later be explored further to identify their nature: MUOD identifies multiple types of outliers to make this easier
- MUOD scales to millions of users, so it is usable in large OSN
- We successfully tested MUOD in data of Google+ with 170M users over 2 years

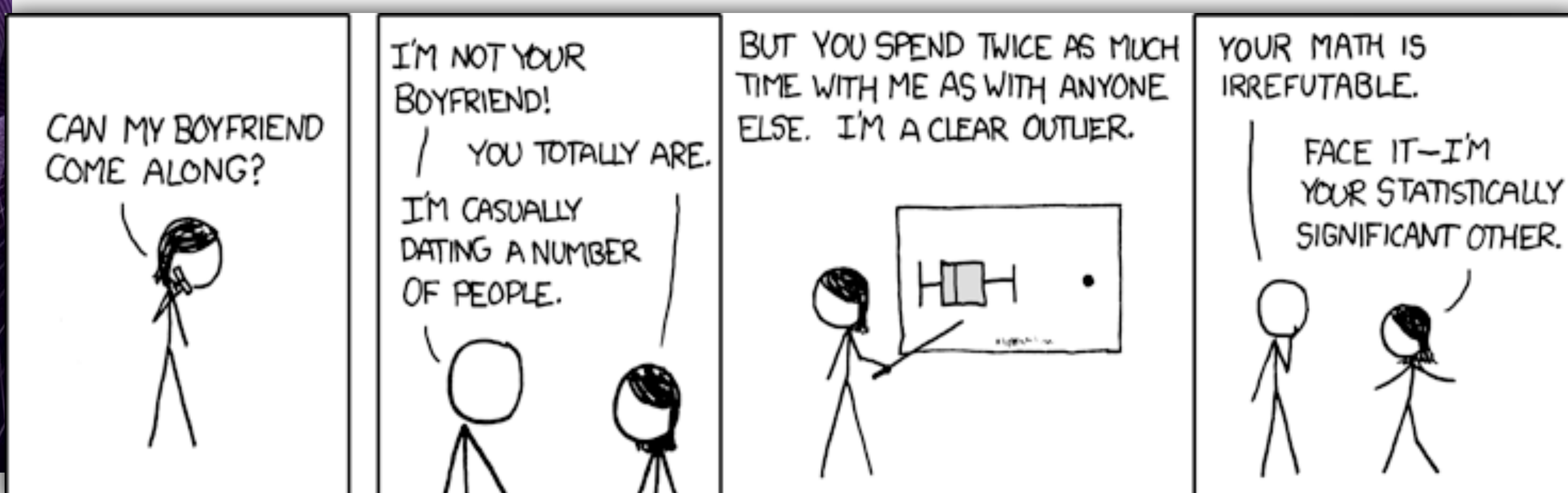
Problem Statement

- We have a set of *n* OSN users
- For every user we have *d* attributes:
  - Connectivity: Number of friends, followers, centrality metrics, etc.
  - Activity: Number of posts, likes, reposts, etc.
  - Profile: user's name, location (e.g., city where she lives), job, education, gender, and related data

Outliers

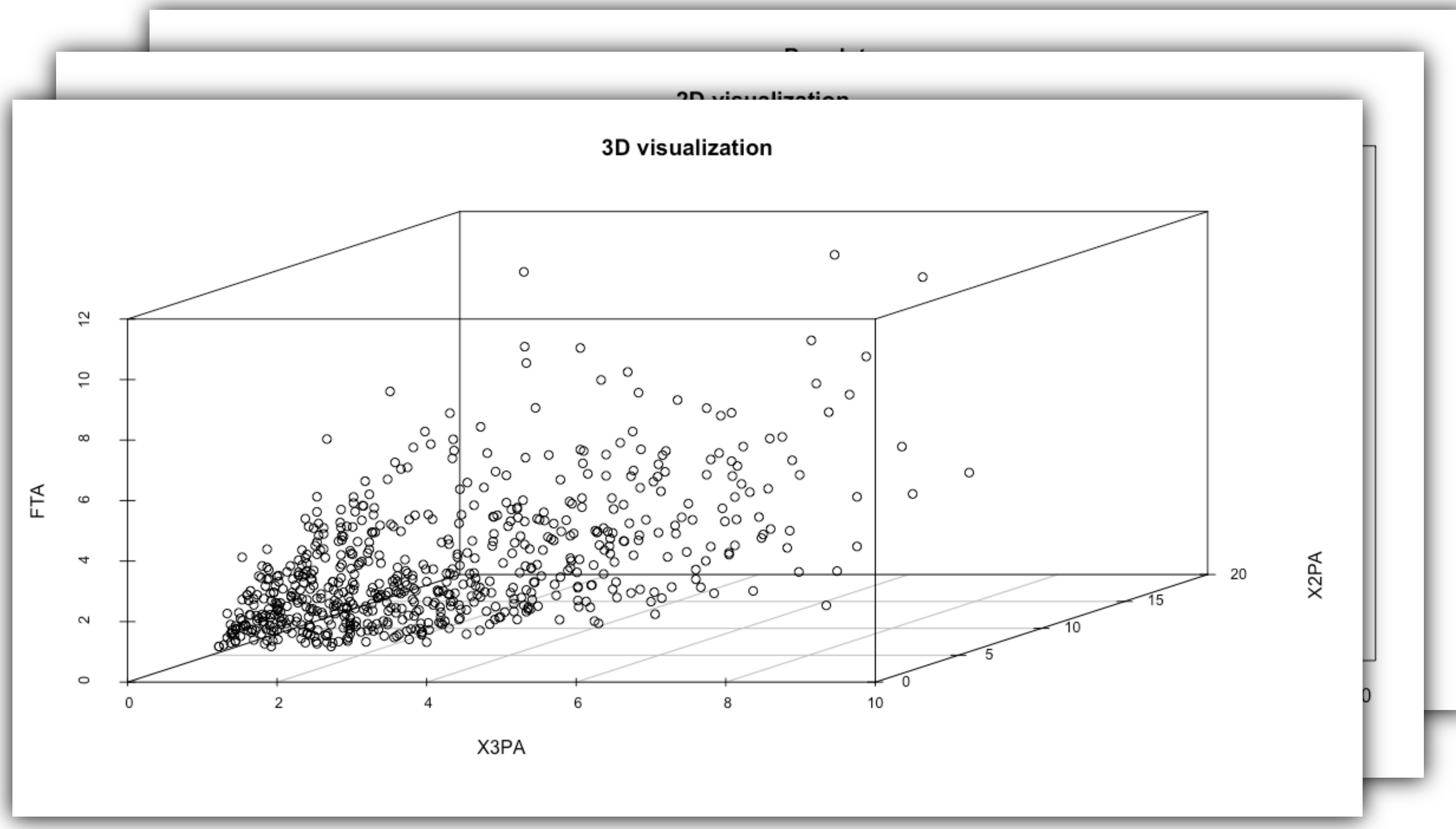
- The objective is to find the outliers in the set of OSN users





Multidimensional Data

- Detecting outliers in multidimensional data is not easy

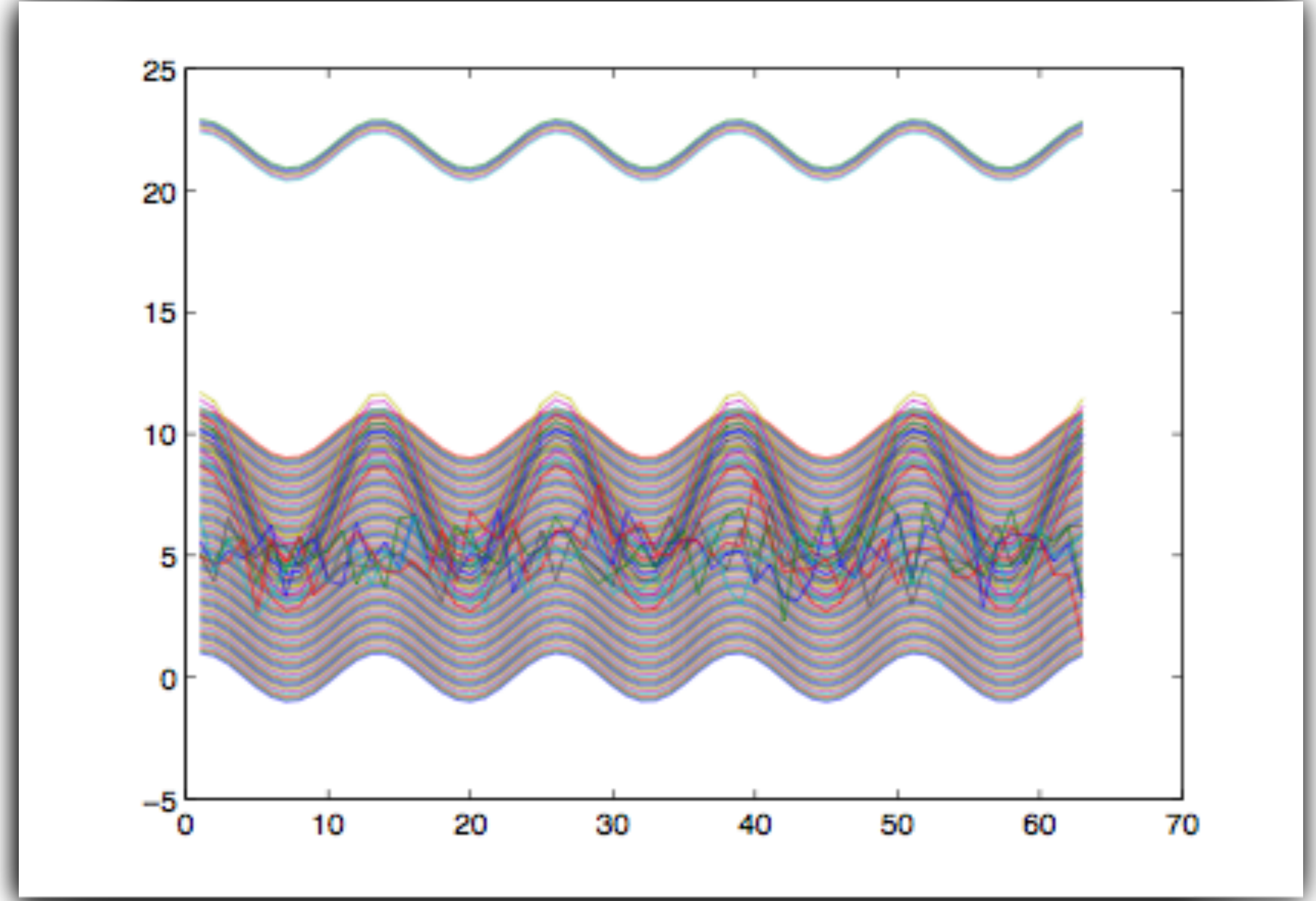


Multidimensional Data

- With more than three dimensions, it is practically impossible to graphically visualize the observations using Cartesian coordinates.
- Convenient alternative: parallel coordinates [Wegman 1990]
- Observation $x \in R^d$ can be seen as real function defined on an arbitrary set of equally spaced domain points, e.g., $\{1, \ldots, d\}$, and x can be expressed as $x = \{x(1), \ldots, x(d)\}$ [López-Pintado Romo 2009]

Functional Data Analysis

- Each observation/user is expressed as a curve, and the outliers are curves that are different from "the mass" [Hubert Rousseeuw Segaert 2015] in
  - Magnitude
  - Amplitude
  - Shape



The Method

- In MOUD we assign to each user an index that gives the outlier intensity of each type:
  - The shape index I_S is based on the correlation coefficient between the functions
  - The amplitude index I_A is based on the slope of linear regression curves between the functions
  - The magnitude index I_M is based on the constant term of linear regression curves between the functions
- The higher the corresponding index, the more likely the user is an outlier

Shape Index

Let us consider the set of users

$$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$$

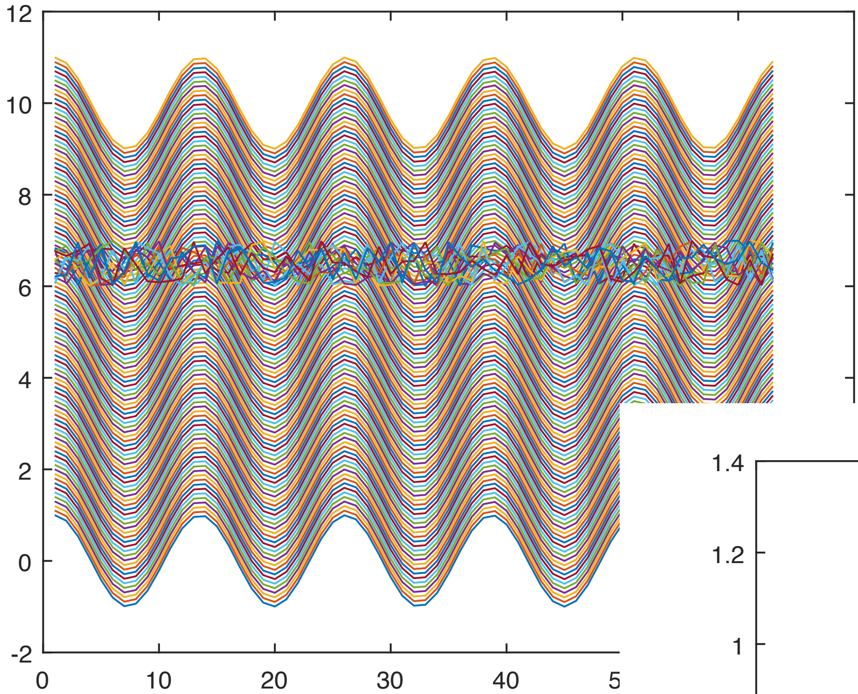
Where each user is a vector of d values

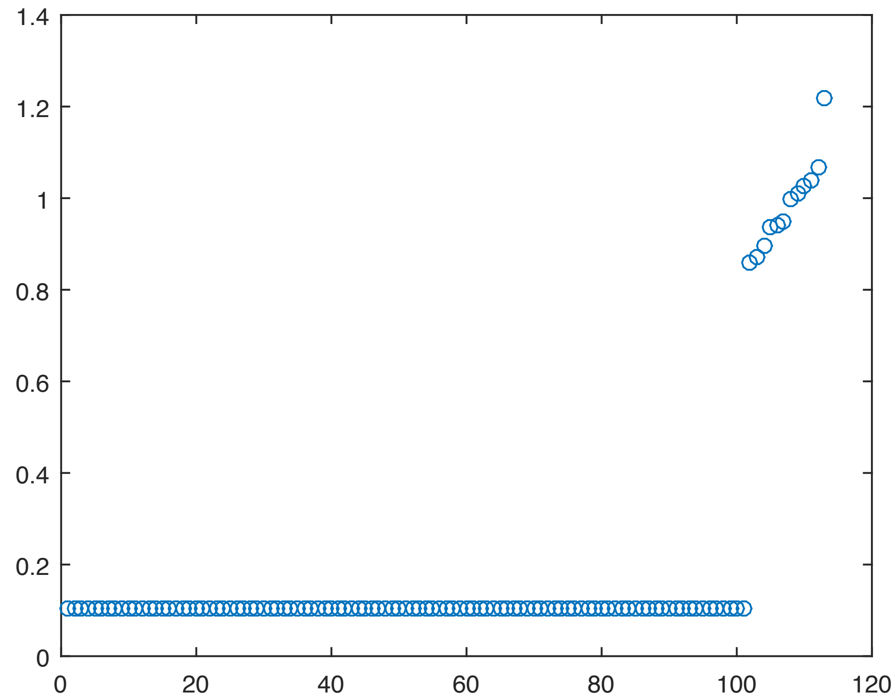
The shape index of a user x is computed as

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

Where $\rho(x, x_j)$ is the Pearson correlation coefficient

Shape Index Example





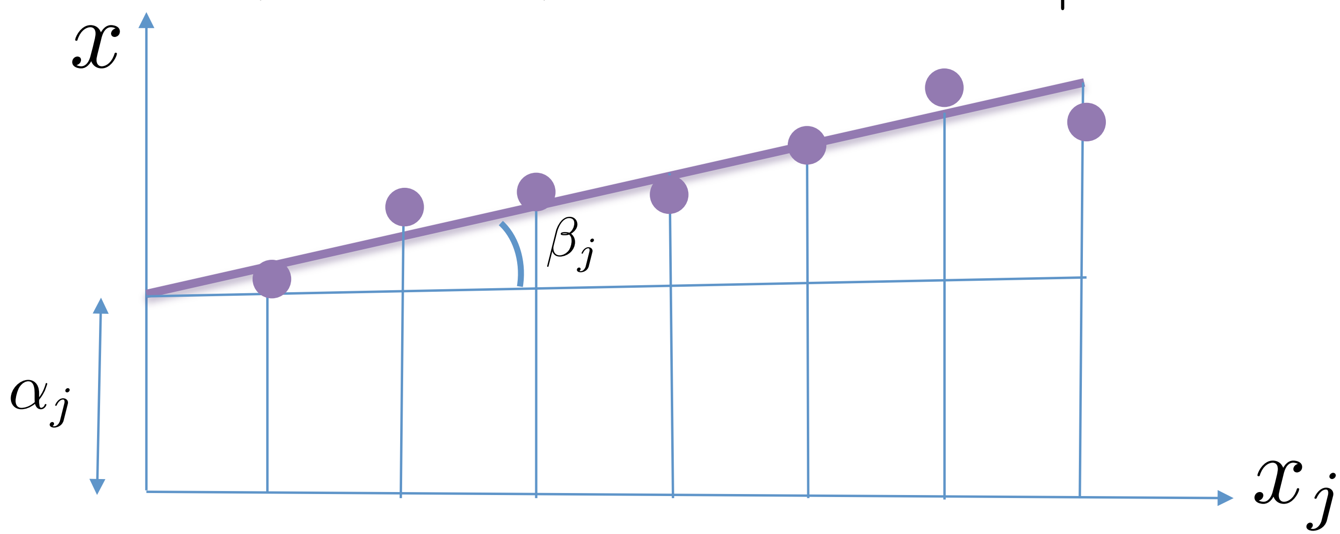
Magnitude and Amplitude Indices

We use linear regression

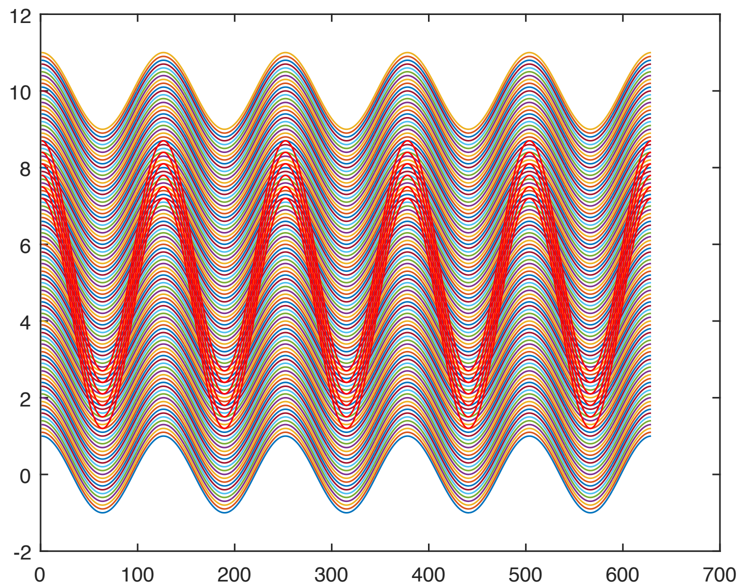
$$\hat{\beta}_j = Cov(x, x_j) / Var(x_j) \qquad \hat{\alpha}_j = \overline{x} - \hat{\beta}_j \overline{x}_j$$

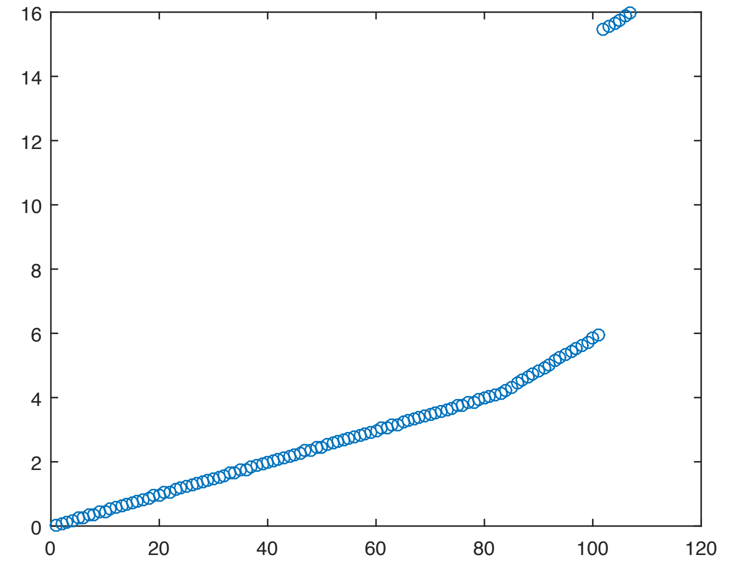
To obtain the magnitude and amplitude indices

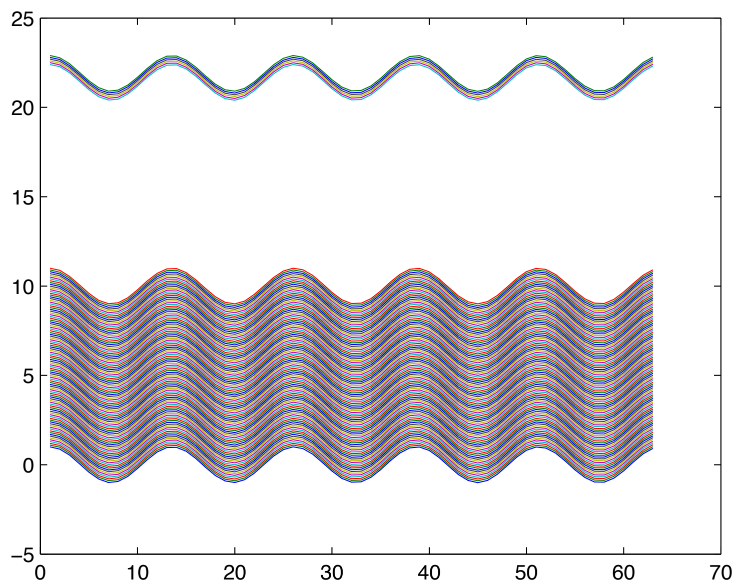
$$I_M(x, \mathcal{X}) = \left| \frac{1}{n} \sum_{j=1}^{n} \hat{\alpha}_j \right| \qquad I_A(x, \mathcal{X}) = \left| \frac{1}{n} \sum_{j=1}^{n} \hat{\beta}_j - 1 \right|$$

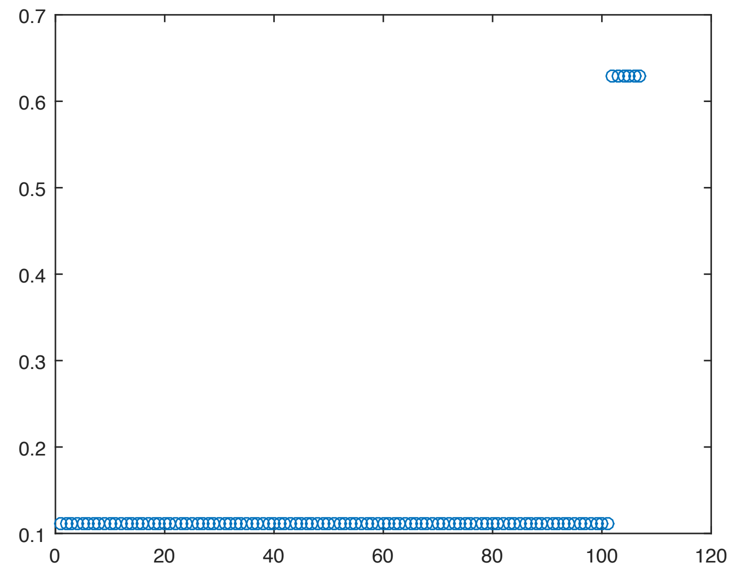


Magnitude and Amplitude Indices



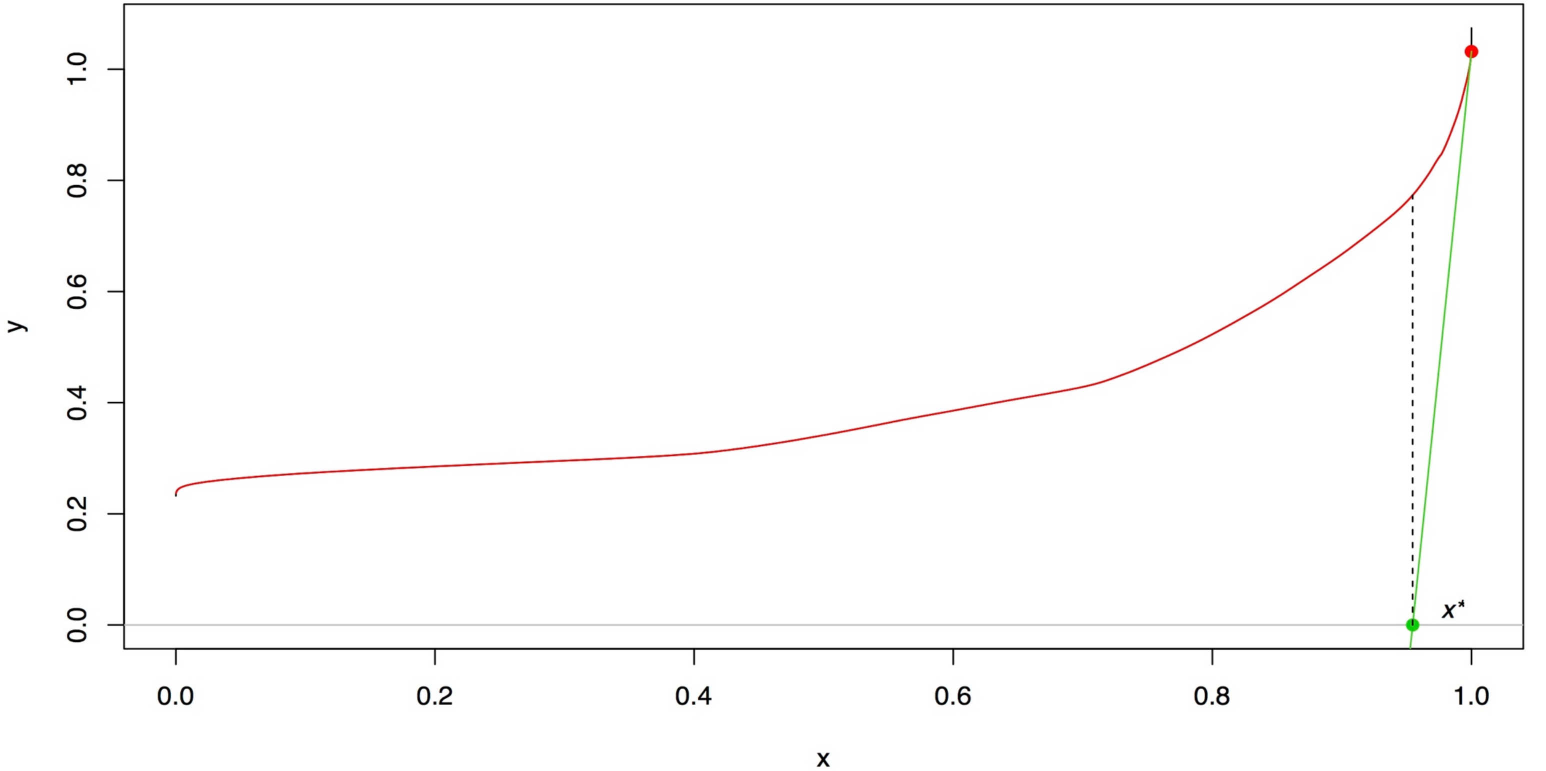






Which are Outliers?

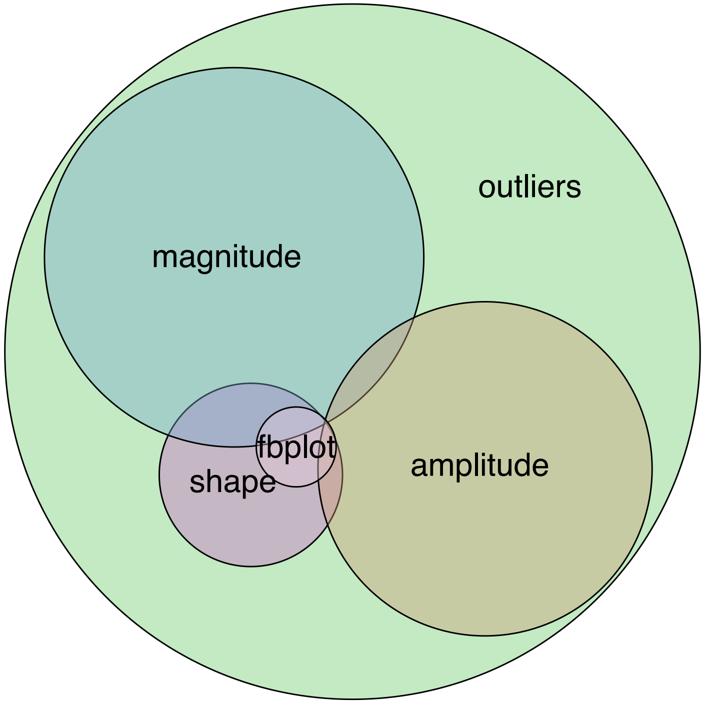
- Given the index I_S of each user we can obtain the set of outliers:
  - Sort by I_S
  - Cut by point given by the tangent method [Louail 2014]



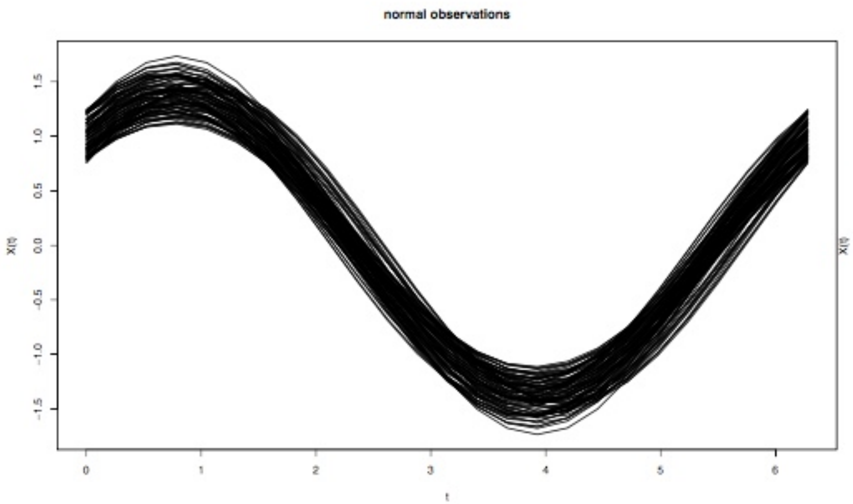
Sets of Outliers

- Given the sets of outliers of shape, magnitude and amplitude, we have up to 7 different outliers subsets to consider, given their possible intersections

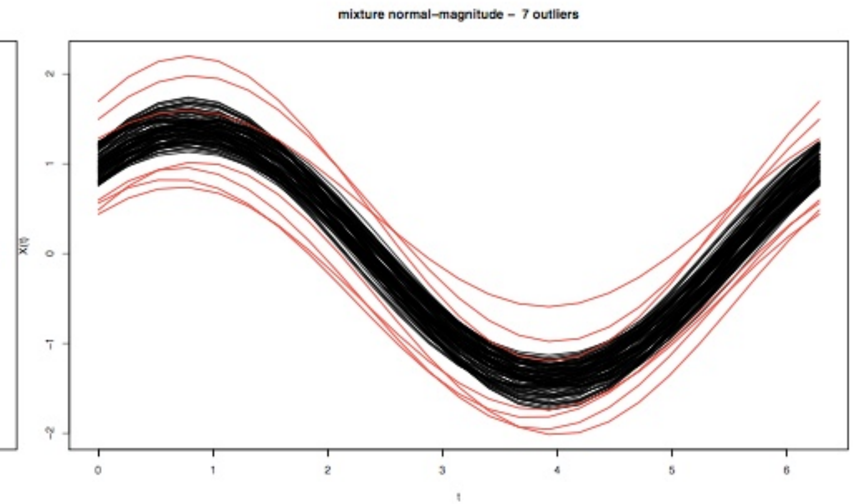
Outliers groups. Simulation



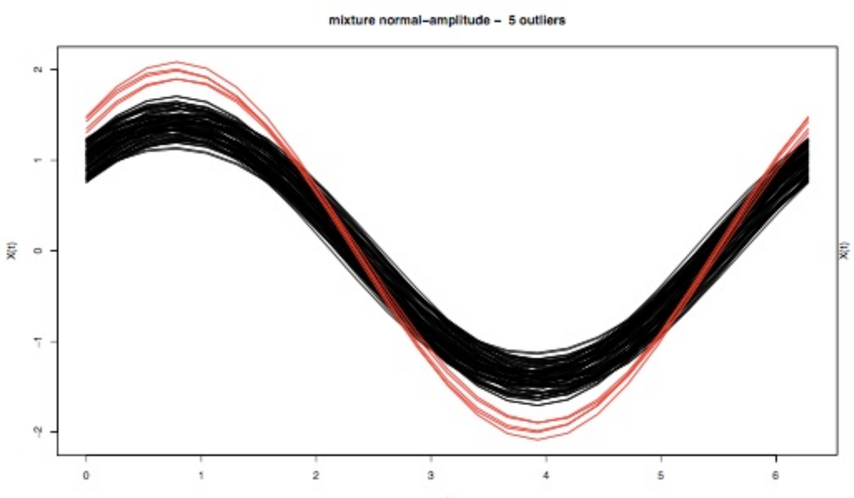
Performance Evaluation



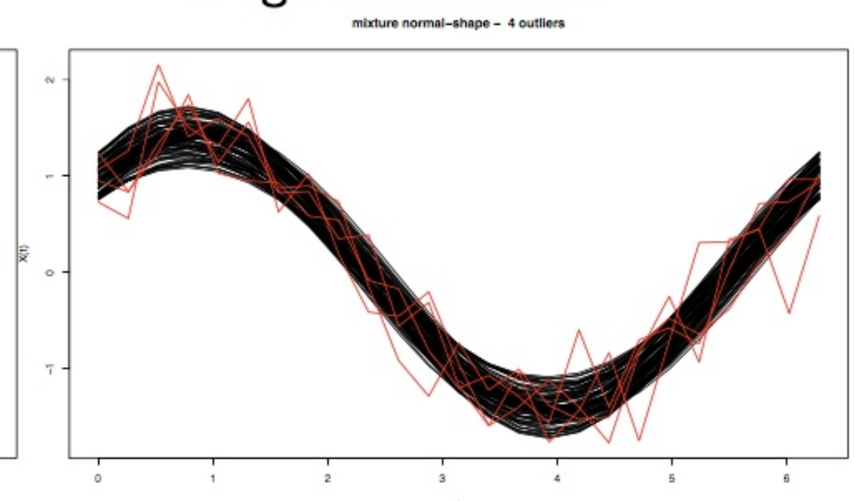
Normal observations



Magnitude outliers



Amplitude outliers



Shape outliers

uc3m

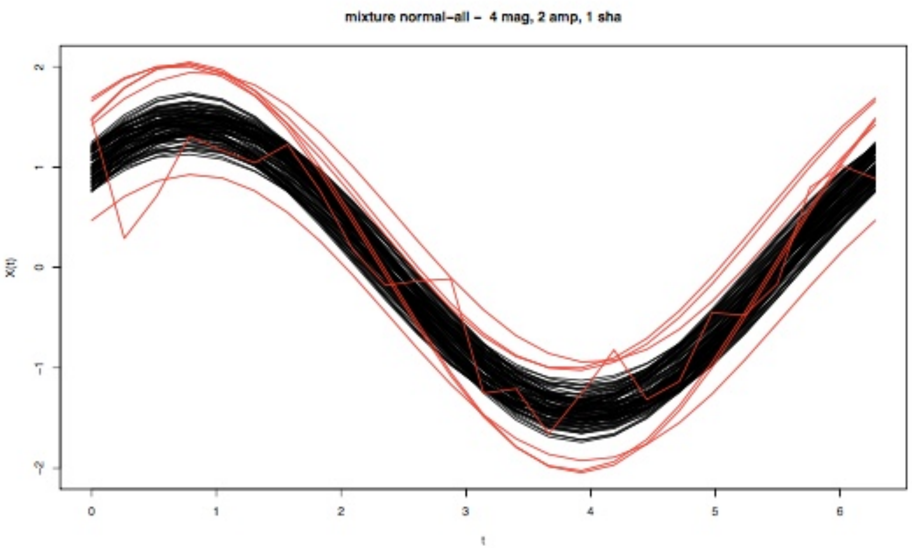
Performance Results

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

institute imdea networks

Mixed 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

Decomposed Results

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

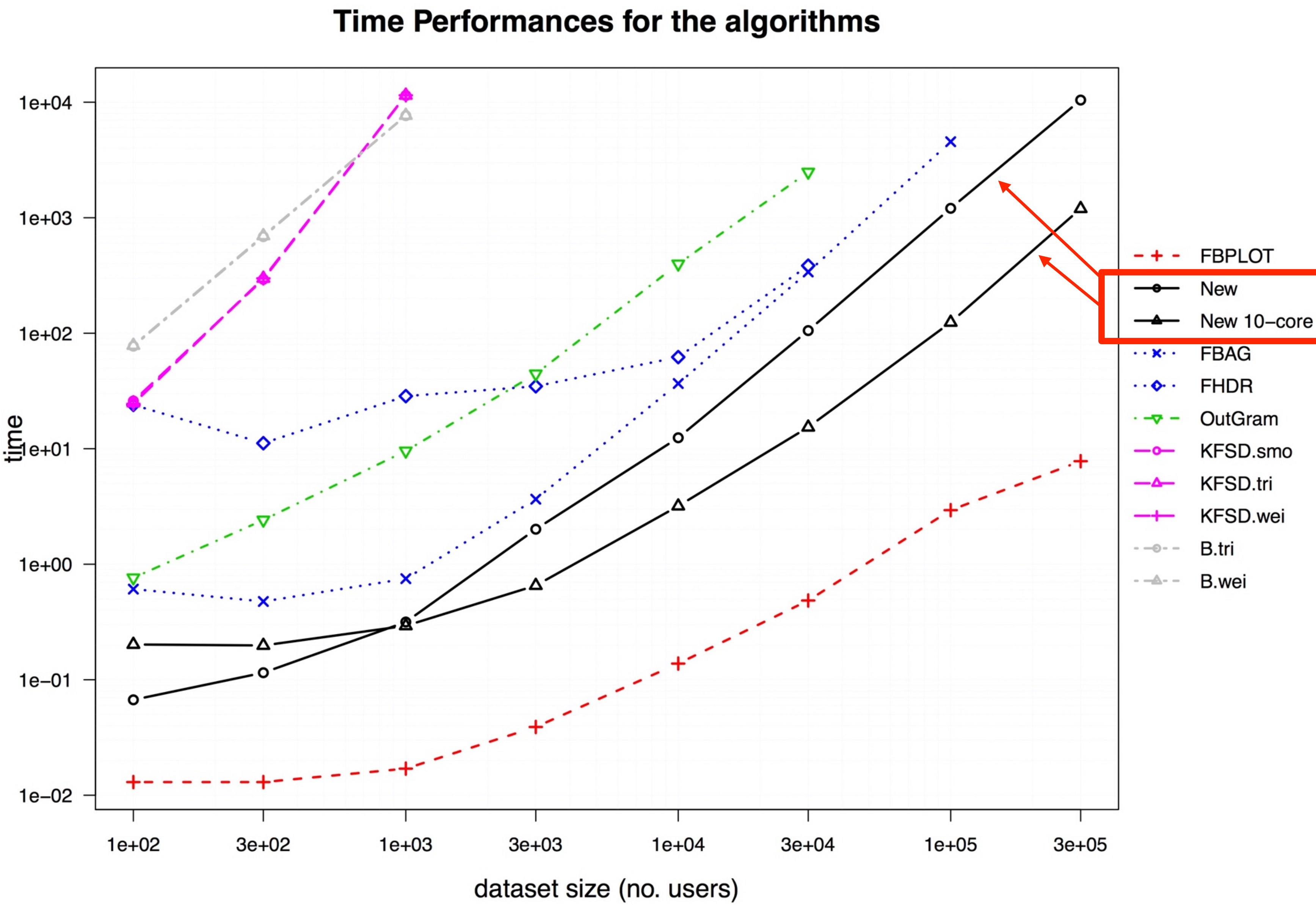
	mag				amp				shape			
	c	f	F	r	c	f	F	r	c	f	F	r
B_{tri}	73.08	1.92	0.52	3	21.40	2.84	0.14	9	89.98	1.65	0.63	1
B_{wei}	100.00	3.23	0.52	3	88.60	3.49	0.45	4	99.80	3.27	0.52	4
FBAG	0.77	2.07	0.01	10	98.40	0.42	0.88	2	8.64	1.94	0.08	11
FHDR	8.65	4.94	0.04	9	98.00	3.42	0.49	3	22.00	4.71	0.11	10
FBPLOT	40.19	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

Implementation

- We have implemented the outlier detection algorithm MUOD in R
- We had to implement it in C++ and add it to the R system, since R functions did not allow the required memory control
- The implementation allows parallel execution in p cores, with time complexity $O(n^2d/p)$
- It has been made available in a public repository:

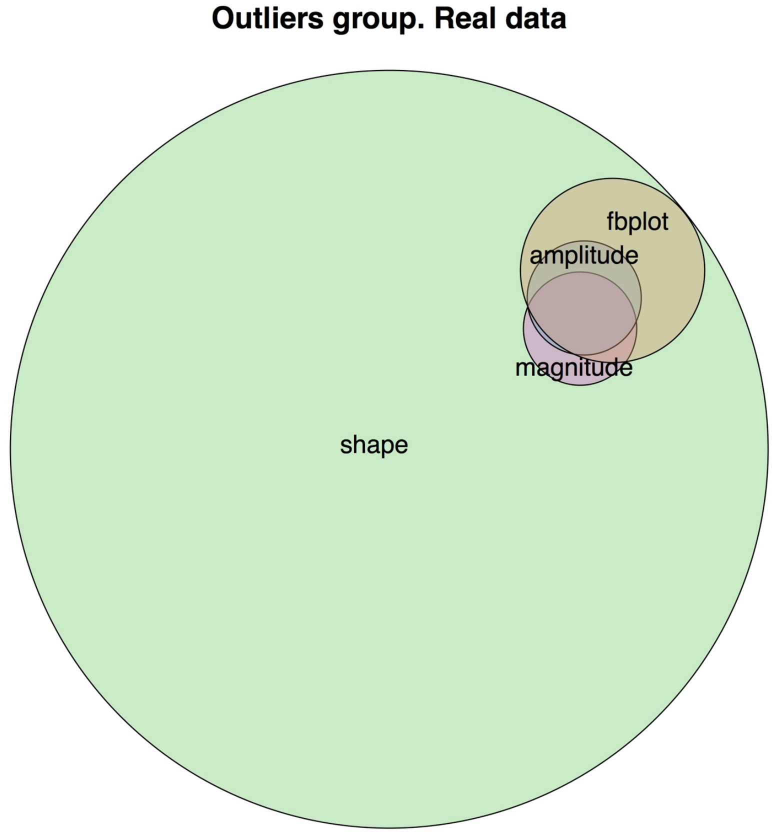
`https://github.com/luisfo/muod.outliers`

Performance

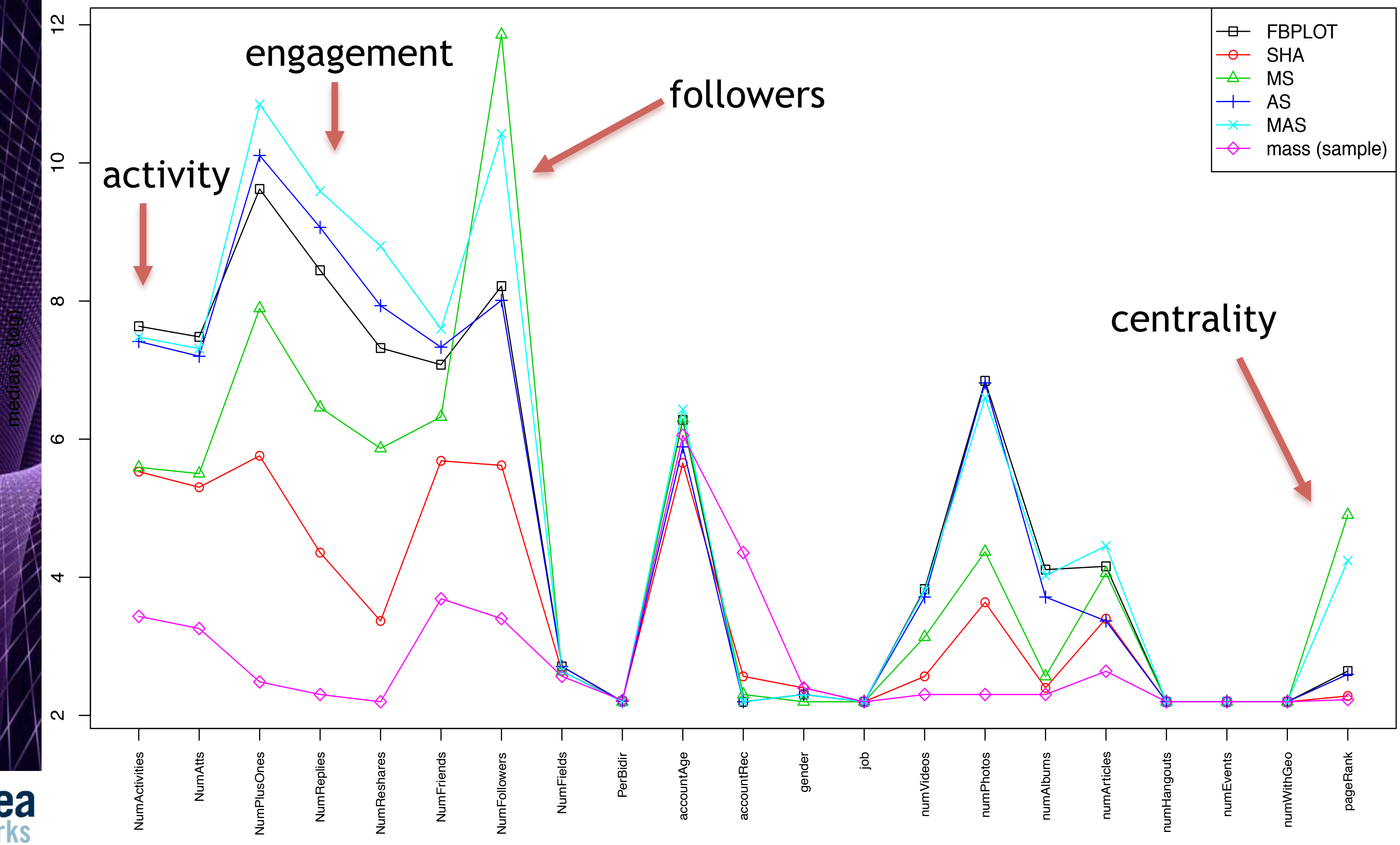


MOUD in Google+

- We have data of n=170M Google+ users and 2 years of activity (2011-2013), with d=21 features for each (of profile, activity, and connectivity)
- We use the 5.6M active
- We find:
  - 4K outliers of MAS
  - 2K outliers of MS
  - 2K outliers of AS
  - 294K outliers of only SHA

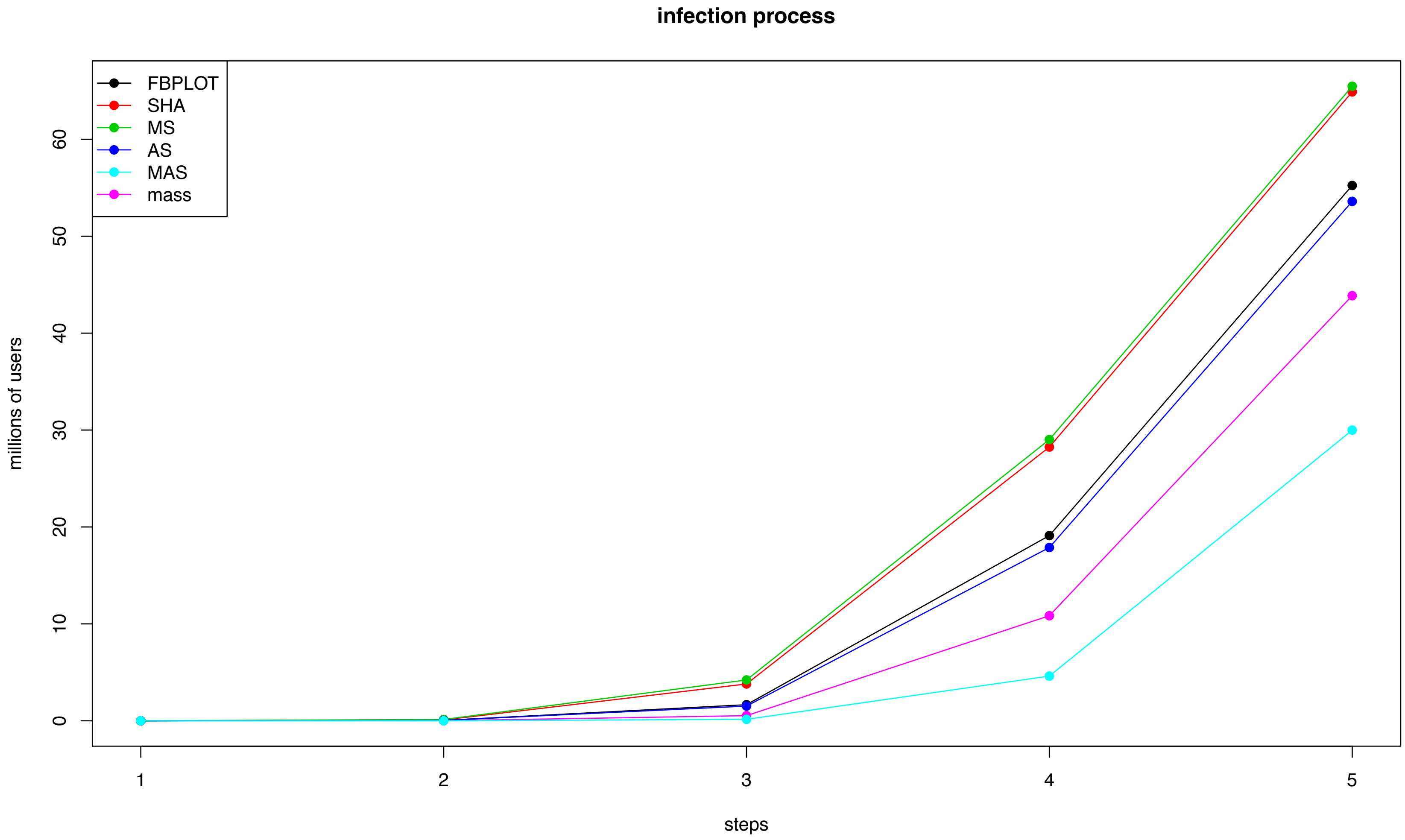


Exploration of the Outlier Sets



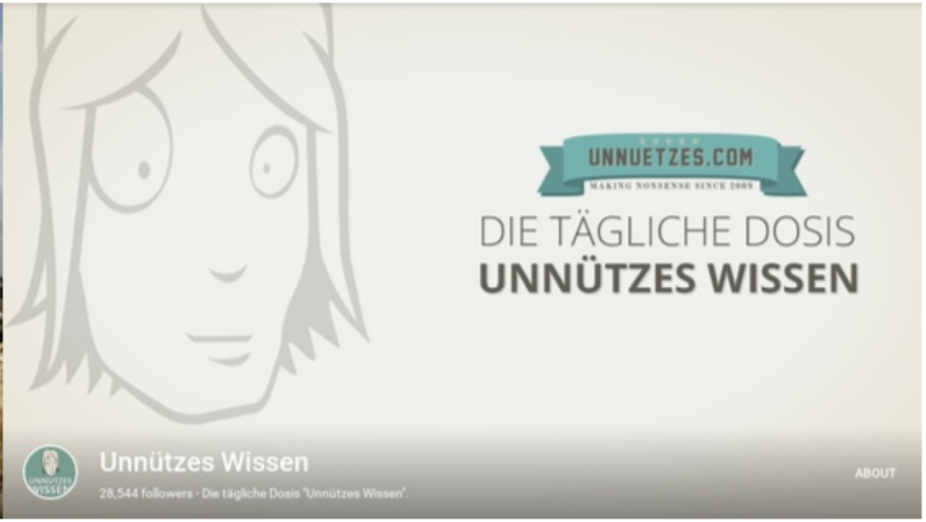
Epidemic Behavior

- We run 10 SI (susceptible-infected) simulations in the connected component (170M users)



uc3m

Examples of Outlier Users



Median amplitude. Al Jazeera

Median magnitude. German Humour.

Amplitude-magnitude outliers. Musical group

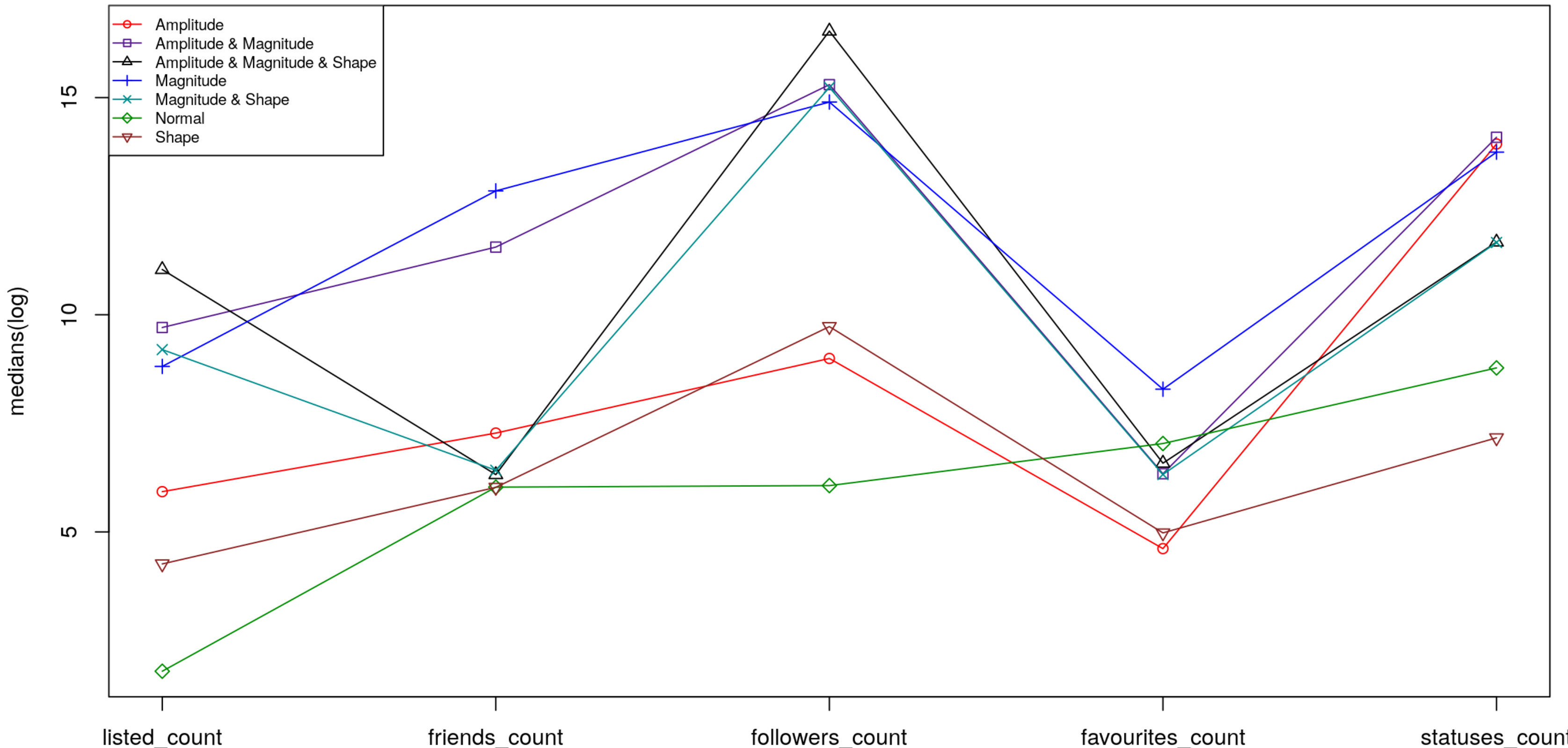
institute imdea networks

Conclusions and Future Work

- We propose to use an unsupervised outlier detection method to identify “interesting” users in OSN
- Then, explore what are the outliers
- We propose a new method that scales to millions of users and test it with a real data set
- In the future we plan to use the method in multiple contexts where identify outliers in multidimensional data is useful (fraud detection, faulty images, etc.)

Ongoing Work

- Data from Twitter (MAG 2, AMP 226, SHA 6871, MA 5, MS 165, MAS 25, rest 138280)



Thank you!!

Azcorra, A., Chiroque, L. F., Cuevas, R., Fernández Anta, A., Laniado, H., Lillo, R. E., Romo, J., and Sguera, C. (2018), "Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks" Scientific Reports (2018).

`https://github.com/luisfo/muod.outliers`