Outskewer: Using Skewness to Spot Outliers in Samples and Time Series

Sébastien Heymann, Matthieu Latapy, Clémence Magnien

Outskewer: Using Skewness to Spot Outliers in Samples and Time Series

Sébastien Heymann, Matthieu Latapy, Clémence Magnien
LIP6 - CNRS - Université Pierre et Marie Curie
4 place Jussieu, 75252 Paris, France
Email: first.lastname@lip6.fr

***Abstract*—Finding outliers in datasets is a classical problem of high interest for (dynamic) social network analysis. However, most methods rely on assumptions which are rarely met in practice, such as prior knowledge of some outliers or about normal behavior. We propose here *Outskewer*, a new approach based on the notion of skewness (a measure of the symmetry of a distribution) and its evolution when extremal values are removed one by one. Our method is easy to set up, it requires no prior knowledge on the system, and it may be used on-line. We illustrate its performance on two data sets representative of many use-cases: evolution of ego-centered views of the internet topology, and logs of queries entered into a search engine.**

I. Introduction

Faced with complex networks that evolve over time, a frequent need is to monitor their evolution and automatically raise alerts on abnormal behaviors of the system, i.e. events which are statistically different from most others. This challenging task is generally called *outlier detection*. In spite of many works addressing this question for decades in various fields, the diversity of cases leading to different outlier definitions makes it hard to create a single universal method. Here we consider the case of a property measured on an evolving network (Figure 1). How can we automatically and reliably identify outliers in it? It is challenging because these data contain both regime changes (i.e. sudden changes of the mean of the time series) due to the evolution of the normal behavior, and outlying values that deviate globally or locally from the main trend. Moreover, we have no prior knowledge on the data; events may occur at different time scales; we want an on-line method for real-time analysis. These settings are known to pose a difficult problem. This paper introduces a new method to automatically detect outliers in sets of numbers and in time series. We also show its relevance for detecting abnormal events in computer and social networks. The source code is available [1].

A. Related Work

Given a data set, outlier detection aims at finding data points which are very different from the remainder. This field has received a large attention in the last decades because outliers often represent critical information about an abnormal behavior of the system described by the data. It covers a broad spectrum of applications such as the identification of mechanical faults, changes in system behavior, human and instrument errors, natural deviations in a population, or data cleaning prior to modelling. Outliers are also called: event, novelty, anomaly, noise, deviation or exception [2].

However there is no formal definition of an outlier because this intuitive notion varies with the context and the desired characteristics of outliers. In a statistical perspective, Grubbs [3] defined that "an outlying observation, or outlier, is one that deviates markedly from other members of the sample in which it occurs". Hawkins [4] defines an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism", while Barnett and Lewis [5] call an outlier "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". The diversity of applications has led to the introduction of various techniques for outlier detection [6]. Areas of research such as statistics, data mining, information theory and process control theory have produced various methods for spotting outliers in stochastic processes. Specific researches also address the question of detecting anomalies in the Internet traffic [7].

Existing methods may be divided between *univariate methods* (i.e. considering one variable), proposed in earlier works in statistics, and *multivariate methods* (i.e. considering multiple variables) which form the main part of the current body of research. Although univariate methods have been studied during a long time, and despite recent focus on multivariate methods due to the increase of computational power, univariate methods remain important to study.

We also distinguish parametric and non-parametric (model-free) procedures [8]. Parametric procedures assume the values to be identically and independently distributed following a known probability distribution (generally a normal distribution), or at least a statistical estimate of the distribution parameters to fit the data. They flag as outliers the values that deviate from the model hypotheses. They are often unsuitable for data sets without prior knowledge of the underlying distribution [9] because the hypotheses (e.g. the independence of values) are not satisfied, and because the statistical models are not reliable for real data and are hard to validate since many data sets do not fit one particular model.

Non-parametric procedures do not assume knowledge of the data distribution, and learn to detect outliers. In some cases (*supervised learning*) labelled data sets are available, from which the program builds a model of normal behavior (and sometimes also a model of outlying behavior). Otherwise

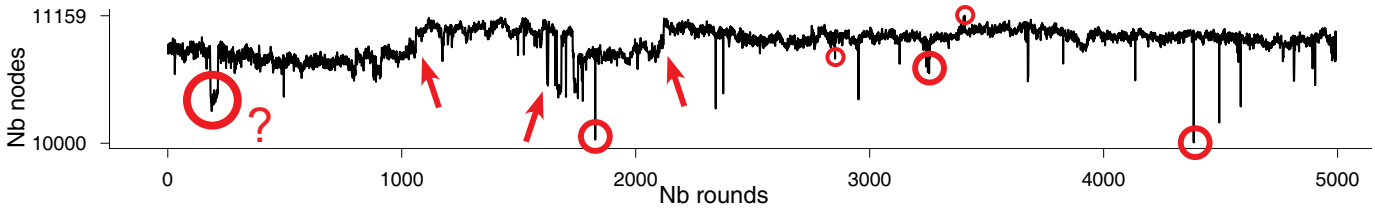

Fig. 1. Evolution of a property measured on a network during time. Some outliers are circled. Regime changes are pointed by arrows.

(*unsupervised learning*), the procedure builds a probabilistic model of the data set, and updates this model as new points appear. These procedures classify as outliers the data points that deviate significantly from the model. These approaches are based on histogram analysis, kernel density, distance measures or clustering analysis.

The output of an outlier detector is a score of "outlierness" assigned to each data point, which represents its probability to be an outlier, or the distance from normal points. Data points are ultimately classified as outliers when their score is above a given threshold which is a parameter of the method.

The detection of outliers in temporal data relies mainly on two approaches. In the first one, points which deviate from a temporal model like the autoregressive integrated moving average (ARMA) model [10] or a finite-state automaton model [11] are marked as outliers. In the second one, points very different from other points within a sliding window are marked as outliers. Regime changes (i.e. change points in time series that are observed by sudden changes of the mean) may be considered as anomalies as well [12], [13].

Finally, recent papers address the issue of outlier detection in networks and graph streams [14] by finding surprising motifs [15] [16].

B. Contribution and Organization of the Paper

We propose in this paper a new unsupervised non-parametric univariate method that reliably detects multiple outliers on either static or temporal data sets given the following setting, which is known to be hard: values may not be independent and identically distributed; we have no prior knowledge of the underlying process which generated the data, or of the probability distribution; in time series, regime changes may exist due to the evolution of the normal behavior (non-stationarity), and also outlying values which deviate globally or locally from the main trend. We finally want an on-line method for real-time monitoring. In this context, our method has the following advantages: (a) it uses a novel approach based on the study of the skewness of distributions, and is easy to interpret; (b) it looks for outliers only when the notion of outlier is relevant in the considered data set; (c) it is easy to use, as the only parameter is the size of the time window for time series, and (d) it may be used on-line.

We describe our method in Section II, validate it in Section III, and apply it on real-world data in Section IV; we conclude in Section V.

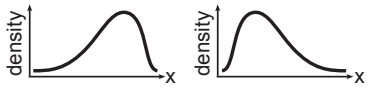

Fig. 2. Example of negative (left) and positive (right) skewed distributions.

II. The Outskewer Method

Our method relies on the notions of *skewness* of distributions and its evolution when extremal values are removed, which we call *skewness signature*; we use this to detect outliers in multisets of numbers and in time series.

A. Skewness

We consider a multiset (i.e. a set in which members may appear more than once) X of n values. The distribution of these values is the fraction P_x, for each x, of values in X which are equal to x. Such distribution samples are basically described by their mean $\bar{x} = \sum_{x \in X}(x/n)$ and standard deviation $\sigma = \sqrt{1/(n-1) \cdot \sum_{x \in X}(x - \bar{x})^2}$. Going further, the sample skewness is a measure of distribution asymmetry, and can be estimated by:

$$\gamma(X) = \frac{n}{(n-1)(n-2)} \sum_{x \in X} \left(\frac{x - \bar{x}}{\sigma} \right)^3 .$$

Intuitively a negative skewness indicates a tail on the left of the distribution more pronounced than the one on the right, while a positive skewness means the converse, see Figure 2. If no tail exists, i.e. all values are equal, $\gamma(X)$ is undefined because $\sigma = 0$. If both tails exist on each side and are equal, $\gamma(X) = 0$. For normal distributions ($P_x = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$), $\gamma(X) = 0$, while for Pareto distributions ($P_x = \frac{ab^a}{x^{a+1}}$ where $0 < b \leq a$), $\gamma(X) > 0$. Examples of unimodal skewed distributions are shown on Figure 2.

The skewness has the interesting feature to be influenced by values which are far from other values, because it is based on the cubed distance from values to the mean. Hence its value changes a lot if they are removed. We show now how to use this feature for outlier detection.

B. Skewness Signature

We consider the evolution of the skewness of a distribution of values in a multiset X while extremal values are removed one by one from X, which we call the *skewness signature* of

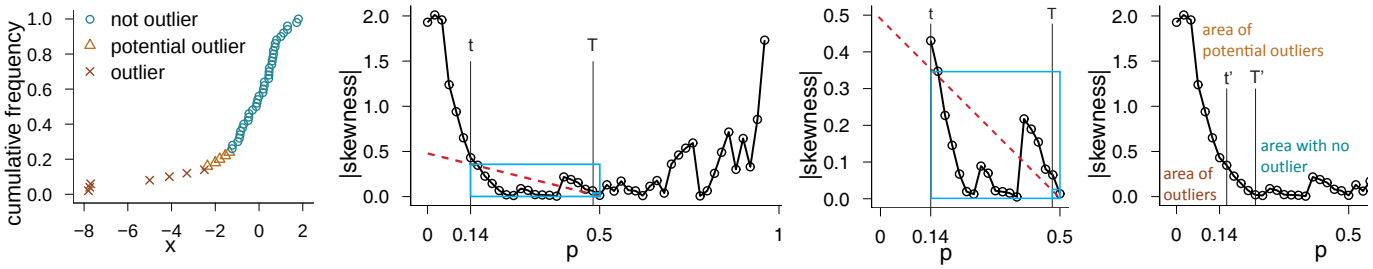

Fig. 3. Example of 50 values with 7 outliers and 5 potential outliers (from left to right): cumulative distribution; absolute values of the skewness signature; zoom on it; absolute values of skewness for which outliers and potential outliers are detected. We obtain $t = 0.14$, $T = 0.48$, $t' = 0.16$, $T' = 0.24$.

X. The extremal value of X, denoted by $e(X)$, is:

$$\mathrm{e}(X) = \begin{cases} \max(X) & \text{if } \gamma(X) > 0, \text{and} \\ \min(X) & \text{otherwise.} \end{cases}$$

In practice, the skewness is almost never equal to zero, hence always choosing $\min(X)$ in the case where $\gamma(X) = 0$ induces a negligible bias.

We define a series of multisets as follows: $X_0 = X$, $X_i = X_{i-1} \backslash \{\mathrm{e}(X_{i-1})\}$, for all $i > 0$. In other words, X_i is the multiset obtained by removing one occurrence of the largest (resp. smallest) value of X_{i-1} if the distribution of values in X_{i-1} has a positive (resp. negative or zero) skewness. Finally, we define the skewness signature as the function $s(p, X) = \gamma(X_{\lfloor p \cdot n \rfloor})$, where n is the size of X and $X_{\lfloor p \cdot n \rfloor}$ is the multiset obtained from X by removing $\lfloor p \cdot n \rfloor$ extremal values, i.e. a fraction p of extremal values.

For example, if $X = \{$-3, -2, -1, -1, 0, 1, 2, 3, 7$\}$, values 7, 3, 2, -3, 1, -2, 0 are removed in this order[1], and the values of the skewness signature are 1.09, 0.22, 0.17, 0, 0.40, 0, 1.73.

The skewness signature may be used to find outliers in unimodal distributions because outliers lie at their extremities, and because skewness is sensitive to the removal of outliers.

C. Outlier Detection

Our method relies on the following hypotheses: outliers are extremal values which cause the skewness to be far from zero; the skewness signature converges to zero (i.e. the distribution becomes more symmetric) when outliers are removed one by one. Therefore, the distance of the skewness to zero can be used to identify outliers. Extremal values which cause this distance to be too large should be classified as outliers. But how is it possible to determine that the distance is too large without making any hypothesis on the data set?

We propose to consider the distance relatively to the proportion of extremal values removed: the more extremal values removed, the closer to zero the skewness is expected to be. For any $p \in [0; 0.5]$ we say that s is ***p*-stable** if and only if $|s(p', X)| \leq 0.5 - p$, for all $p' \in [p, 0.5]$. We do not consider values of p larger than 0.5 because this corresponds to a removal of more than half of all values; in such situations, the skewness has little to do with the original data, and it may vary much if too many values are removed.

Let t be the smallest value such that s is t-stable, and T be the largest value such that s is T-stable. When s is never p-stable for any p, t and T do not exist. This case indicates that it is irrelevant to look for outliers in the given data set, according to our notion of outlier; in this case our method classifies all values in the data set as **unknown**. Otherwise we find outliers as follows.

We denote the smallest and largest numbers in X_i by $min_i = \min(X_i)$ and $max_i = \max(X_i)$. Then, $min_{\lfloor p \cdot n \rfloor}$ (resp. $max_{\lfloor p \cdot n \rfloor}$) is the smallest (resp. largest) remaining value when a fraction p of all values has been removed. Let t' (resp. T') be the smallest value of p such that $|\gamma(X_{\lfloor t' \cdot n \rfloor})| \leq 0.5 - t$ (resp. $|\gamma(X_{\lfloor T' \cdot n \rfloor})| \leq 0.5 - T$). Our method concludes as follows: below $min_{\lfloor t' \cdot n \rfloor}$ and above $max_{\lfloor t' \cdot n \rfloor}$, values are **outliers**; between $min_{\lfloor t' \cdot n \rfloor}$ and $min_{\lfloor T' \cdot n \rfloor}$ included (resp. $max_{\lfloor t' \cdot n \rfloor}$ and $max_{\lfloor T' \cdot n \rfloor}$), values are **potential outliers**; values are **not outliers** otherwise. Notice that when $t' = T'$, $min_{\lfloor t' \cdot n \rfloor} = min_{\lfloor T' \cdot n \rfloor}$ (resp. $max_{\lfloor t' \cdot n \rfloor} = max_{\lfloor T' \cdot n \rfloor}$). In this case, values equal to $min_{\lfloor t' \cdot n \rfloor}$ (resp. $max_{\lfloor t' \cdot n \rfloor}$) are potential outliers. Figure 3 illustrates our method on an example.

D. Dynamic Extension

Our method may be used on time series representing the evolution of a system's property. Let $\{x_0, x_1, ..., x_n\}$ be a time series. We consider the multisets which contain w values: $X^i = \{x_{i-w+1}, ..., x_i\}$. Any value x_i of the series belongs to $X^i, X^{i+1}, ..., X^{i+w-1}$. We use our method on all these w multisets, and consider the final class of x_i to be the one which occurs most often among these w classifications. In case of equality, we give priority of *outlier* upon *potential outlier* upon *not outlier*, because we prefer to detect too much outliers than too few.

III. Experimental Validation

The validation of outlier detection methods is difficult because of the various outlier definitions, hypotheses and use cases [8]. Labelled data sets raise also the issue of prior criteria to label the data. We consider that our method should detect outliers, if any, if the notion of outlier is relevant for the given data set. In particular, we consider for our experimental validation the following cases: (a) distributions like Power laws (e.g. Pareto and Zipf's law) commonly contain

[1] Values are removed until $\gamma(X)$ is not computable: our skewness estimator is only relevant for data sets with at least 3 values.

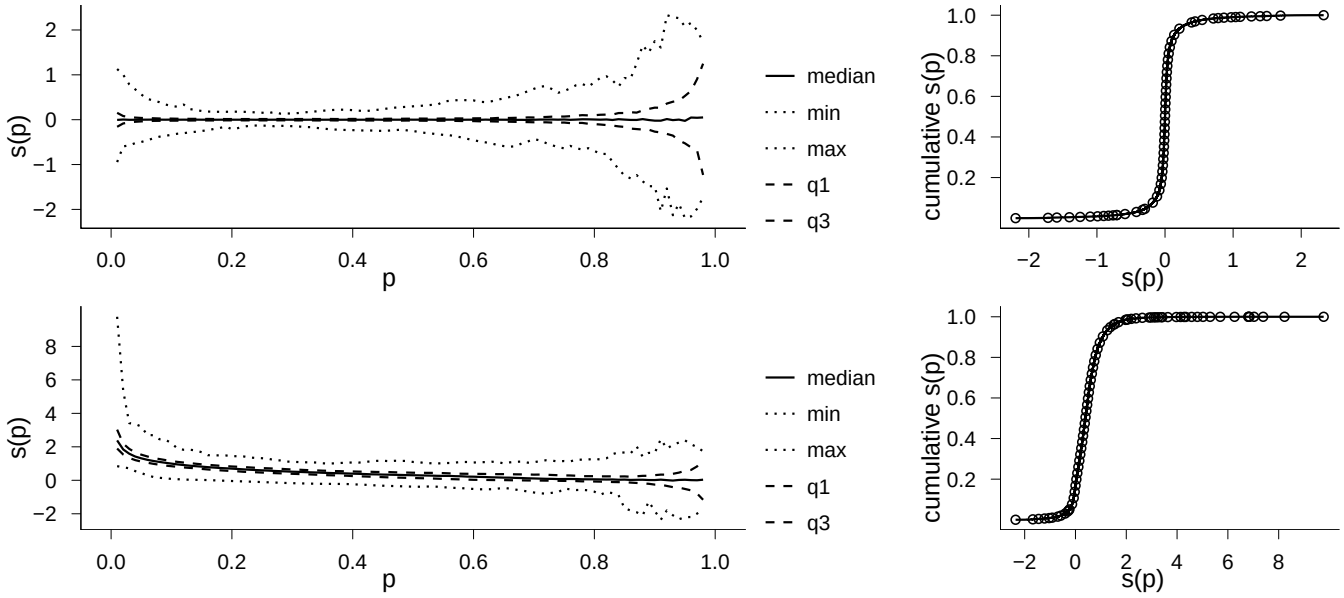

Fig. 4. Quartiles, min and max of $s(p, X)$ on 1,000 normal (top) and Pareto (bottom) samples with the cumulative frequency distributions of $s(p, X)$.

extremal values far from the mean (i.e. *heterogeneous*), so it is erroneous to consider them as outliers, moreover Power law distributions are asymmetrical so our method should conclude that looking for outliers in them is irrelevant; (b) normal distributions are symmetrical and extremal values far from the mean are uncommon (i.e. *homogeneous*), so no outlier should be detected but these extremal values when they occur; (c) half-normal distributions ($P_x = \frac{2a}{\pi} e^{-x^2 a/\pi}$, where $a > 0$), which are basically the absolute of normal distributions with mean equal to 0, are asymmetrical but homogeneous, so this case is ambiguous and should be unclear for our method as well; (d) symmetric Pareto distributions ($P_x = \frac{ab^a}{2}|x|^{-1-a}1_{|x|>b}$, where $0 < a < 2$ and $b > 0$), which are basically the mirror symmetric of Pareto distributions about the vertical axis, are symmetrical but heterogeneous, so we study the behavior of our method in this case.

We first study the relevance of our method on these four distributions, and we study the effect of the sample size (III.A). Then we study the performance of our method to detect outliers, and evaluate the rate of true outliers and false outliers detected (III.B). We finally study the behavior of our method when regime changes occur in temporal data (III.C).

A. Relevance

Our method is applicable if and only if the given data set is p-stable for at least one value of p between 0 and 0.5. A necessary condition for this is that $|s(0.5, X)| < 0.5$. We show in this section that this is true for normal distributions (even with a few outliers) and false for Pareto distributions, which is the expected behavior: normal distributions are symmetrical and homogeneous and Pareto distributions are asymmetrical and heterogeneous.

We study the behavior of s on normal $\mathcal{N}(0, 1)$ and Pareto (shape=6, location=2) probability distributions[2]. For each one, we randomly generate 1,000 samples of 100 numbers to obtain skewness signatures; we compute and plot the skewness signature of each sample in Figure 4. We observe that the values of normal signatures oscillate around zero, whereas the values of Pareto signatures globally decrease and are above zero until $p \approx 0.5$. The cumulative frequency distributions of $s(p, X)$ on Figure 4 confirm these observations. We also computed the skewness signatures of normal and Pareto distributions with various parameters, and also various symmetrical distributions[3] which we do not present here due to space constraints. All of them exhibit patterns similar to normal signatures.

It is clear that the probability for Pareto skewness to be within $[-0.5; 0.5]$ increases with p. We estimate $\mathbb{P}(|s(0.5, X)| < 0.5)$ on 1,000 Pareto and 1,000 normal samples. We obtain that this probability is equal to zero for Pareto samples, and is greater than 0.95 for normal samples. We conclude that our method is able to characterize symmetrical and homogeneous versus asymmetrical and heterogeneous distributions at a confidence level of 0.95. Moreover, the addition of some outliers in these distributions produces almost the same signatures than without outliers, because extremal values are firstly removed. Therefore existing outliers do not notably change the characterization.

Let us study the evolution of $s(0.5, X)$ when the sample size n varies. We generate 1,000 normal and Pareto samples for each value of n between 3 and 1,000, then compute $s(0.5, X)$ for each sample, and we finally obtain the quartiles, min and max of the values of $s(0.5, X)$ at each n. We observe in Figure 5 that the results converge to zero for the normal

[2]Other parameters lead to similar results.

[3]Cauchy, Laplace, some Gamma and Weibull distributions.

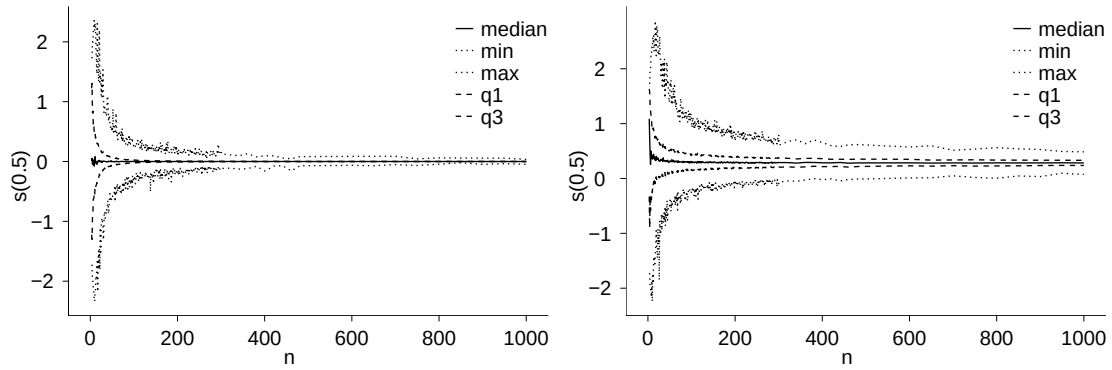

Fig. 5. Quartiles, min and max of $s(0.5, X)$ as a function of n for normal (left) and Pareto (right) distributions.

distribution, and to ≈ 0.3 for the Pareto distribution. Thus, increasing n should lead to a better characterization.

We verify this hypothesis by evaluating the rate of samples where s is never p-stable, for 1,000 normal and Pareto samples for each size n. We observe in Figure 6 that it seems to follow a fast decrease for normal samples. For $n \geq 37$, less than 5% of normal samples are incorrectly characterized, and less than 5‰ for $n \geq 55$. We also observe that it increases with n for $n > 50$ on Pareto samples. The minimum is 79% at $n = 52$, is around 85% at $n = 100$, around 95% at $n = 240$, and above 99.5% for $n > 500$.

We also evaluate this rate for half-normal and symmetric Pareto samples. We observe in Figure 6 that it seems to follow a fast decrease for symmetric Pareto samples, but a slow decrease for half-normal samples. This result is not surprising because the theoretical skewness of half-normal distributions[4] is ≈ 1, and the skewness decreases slowly when extremal values are removed one by one. As expected, our method has unclear results in this case.

We conclude that our methods characterizes samples with size 100 very well, and is excellent on samples of size 1,000. Our method also considers that the symmetric Pareto distribution should contain no outlier.

In addition, we study the skewness range where our method considers s to be p-stable at least once. We vary the shape parameter of a Gamma distribution (thus its skewness) to incrementally generate 1,000 samples of 100 numbers for each skewness value, from Pareto-like samples to normal-like samples, and compute the rate of s that are p-stable at least once for each skewness. We remind that s is p-stable if and only if $|s(p', X)| \leq 0.5 - p$, for all $p' \in [p; 0.5]$. The result in Figure 7 shows that s is always p-stable at least once for samples of skewness below 1.5, and never p-stable for samples of skewness above.

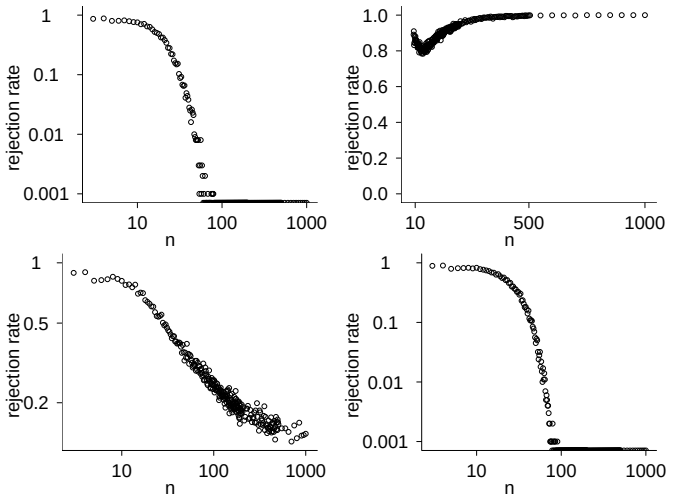

Fig. 6. Fraction of samples for which s is never p-stable as a function of the sample size n, for normal (top left), Pareto (top right), half-normal (bottom left), and symmetric Pareto (bottom right) distributions.

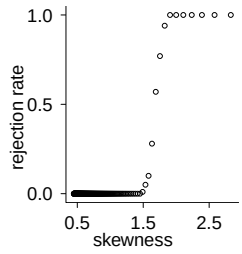

Fig. 7. Fraction of samples for which s is never p-stable as a function of the skewness for Gamma samples (shape varying from 0.3 to 20).

B. Performance

We study the effect of the sample size on outlier detection in normal, Pareto, half-normal and symmetric Pareto distributions. We generate 1,000 samples for each distribution and size n, then we detect outliers on each sample. Normal and Pareto samples contain no outlier by definition, so no outlier should be detected; they are called *false outliers*.

We observe in Figure 8 that the rate of false outliers is low, with at most 3% for the normal distribution and at most 5% for Pareto. This rate decreases when n increases to be less than 1‰ above $n \approx 100$ for the normal distribution, and above $n \approx 500$ for the Pareto distribution. We also evaluate the rate of outliers detected for the symmetric Pareto distribution:

[4] $\gamma = (\sqrt{2} \cdot (4 - \pi))/(\pi - 2)^{3/2}$

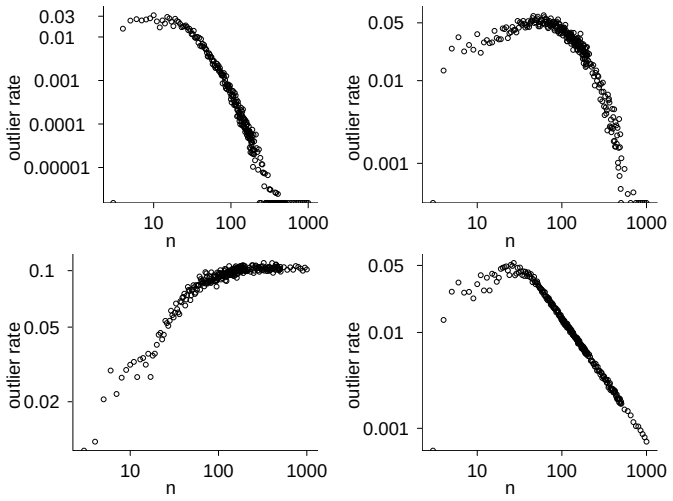

Fig. 8. Fraction of sample points classified as outlier as a function of n for normal (top left), Pareto (top right), half-normal (bottom left), and symmetric Pareto (bottom right) distributions.

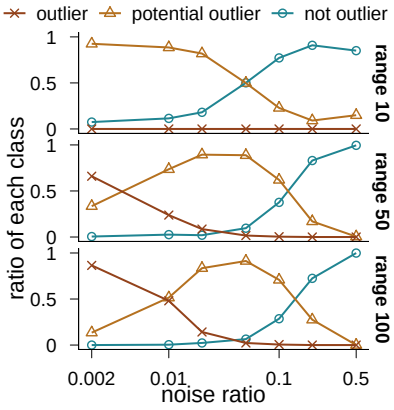

Fig. 9. Ratio of noise points detected as outliers, potential outliers and not outliers as a function of the proportion of noise, for different noise ranges.

reaching 5% at most, it seems to follow a fast decrease when n increases, to reach 1‰ at $n \approx 1000$. For the half-normal distribution, this rate is between 8% and 12% for $n > 100$, and is consistent with the fraction of samples for which s is never p-stable. We conclude that our method detects few false outliers on samples of size 100, and almost none on samples of size 1,000, which is an excellent performance; it rarely detects outliers on symmetric Pareto samples, which is the expected behavior regarding the characterization.

Now we estimate the ability to detect true outliers by generating a sample of size 1,000 composed of a normal sample of variance equal to 1 and a uniform sample (called the *noise*) of size varying from 0.2% to 50% of the total number of values. We then count the number of noise points which are classified as outliers and potential outliers. It is the worst case because the initial skewness is close to zero and outliers are uniformly distributed around the mean with no gap between them and the rest of the distribution. This is also a way to evaluate the robustness of our method against a problem known as the *masking* effect [17], occurring when some outliers are not detected because of the presence of other outliers close to them.

We generate uniform samples of various ranges (i.e. largest minus smallest value). The range of normal samples of size 1,000 is roughly 6 and the range of samples of size 10^6 is roughly 10, so we select noise ranges larger than this: 10, 50 and 100. We observe in Figure 9 that noise points very close to the signal points (range 10) are classified as potential outlier. Larger ranges increase the number of detected outliers. We also see that the less noise, the higher the power to detect true outliers. However almost no outlier can be detected with more than 10% of uniform noise.

C. Regime Changes

Regime changes are change points in time series that are observed by sudden changes of the mean. When they occur we are faced with non-trivial distributions. We study now how our method deals with them. We simulate a stream of values by generating two normal samples of size 110 with mean equal to 0 and 3 respectively. t indicates the order of appearance of the values. Figure 10 shows our method applied dynamically with a sliding window of size $w = 100$. The outlier status of values is unknown at the beginning. At the end, none of them are outliers but one potential outlier. Our method is hence robust against regime changes. Notice that 72 values are classified as potential outliers when our method is applied on the whole data set at once.

IV. Real-World Applications

A. Dynamics of Internet Topology

We applied our method to data collected with the radar for the internet [18], which makes possible to observe the dynamics of the internet's topology at the scale of a few minutes. It consists in focusing on the part of the internet's topology viewed from a single computer called the *monitor*. Periodical measurements of this map, called *ego-centered view*, were performed every 15 minutes during several months, leading to a series of graphs.

The most natural idea to detect events in the dynamics captured by a radar measurement from a given monitor certainly is to study the number N_i of nodes observed at each round i. We plot it for a typical case in Figure 1. Clear outliers appear under the form of sharp decreases of N_i for some values of i, but this brings little information because they may be due to losses of connectivity by the monitor. Except from these statistical outliers, which are detected by our method, the number N_i of nodes observed at each round i in Figure 11 is very stable.

We thus compute the number of distinct nodes seen in five consecutive rounds to avoid the outliers which only reveal losses of connectivity in one round of measurement. We observe events in the dynamics shown in Figure 12, where many decreases existing in Figure 11 have disappeared. Figure 12 is well centered around a typical value, but still exhibits sharp increases and decreases. This means that these outliers, which were also detected by our method, may reveal real events

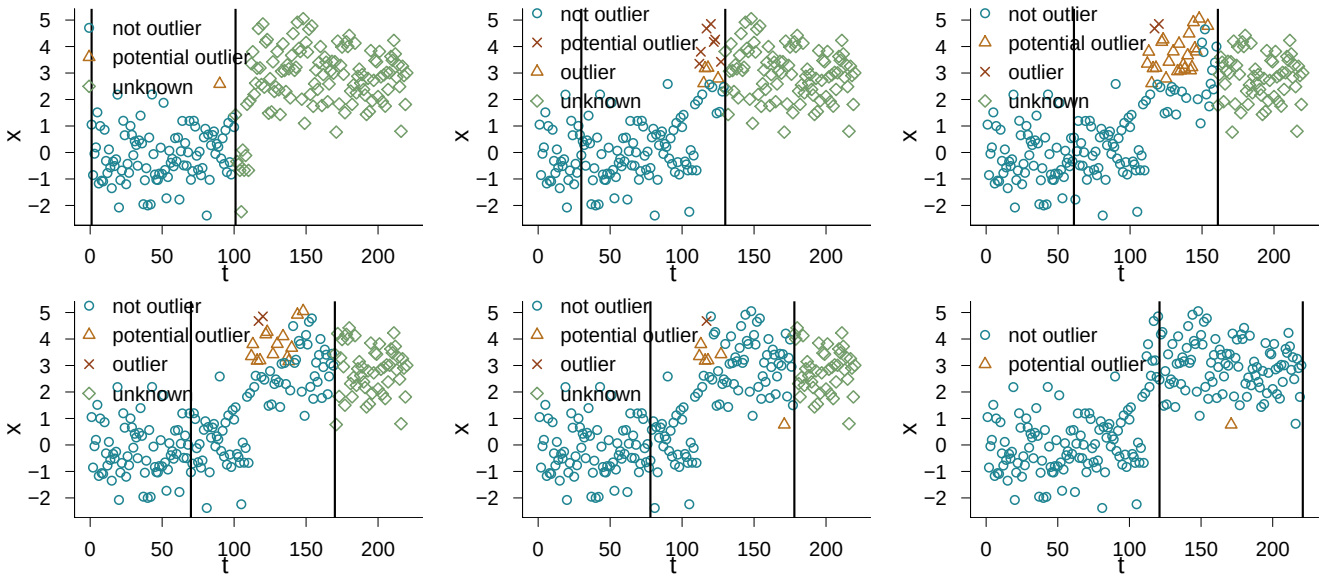

Fig. 10. From top-left to bottom-right, evolution of the outlier status of values in a time series of size $n = 220$, and having one regime change (mean value changing from 0 to 3). Vertical lines indicate the time window boundaries between what outliers are detected.

in the dynamics of this network. Outliers above the typical values indicate a sudden appearance of many new nodes in the network, while outliers below the typical values may indicate longer losses of connectivity or a sudden disappearance of many nodes.

Our approach is hence relevant for studying the evolution of ego-centered views of the internet topology, and for raising automatic alerts in real-time when significant changes of connectivity occur.

B. Search Engine Queries

We applied our method on the data set of search queries captured from a eDonkey server [19]. It consists in textual queries made by users for lists of files matching certain keywords. The measurement lasted for 28 weeks. The data set contains 205,228,820 queries entered from 24,413,195 IP addresses. Samples and procedure descriptions are publicly available [19].

In order to study the number of queries related to the film *Harry Potter and the half blood prince*, we filtered the queries to get only those which contain the words "half blood prince". Then for every 10 minutes we counted the number of queries made during the last hour of measurement. Outliers were finally detected using a sliding window of size $w = 1,008$ (7 days) to capture meaningful events at the scale of one week. We plot in Figure 13 the number of outliers and potential outliers observed each day and each week. The scale of a day seems better for observing fast increases of user queries.

We identify three main events: we observe many values marked as potential outliers during the week after July 15, 2009, when the film was out in theatres. Then an unknown event appears from August 23 to 25, when almost all values are outliers. The last automatically detected event, from October 10 to 12, coincides with the release of a pirated version of the film on October 10 on BitTorrent, another P2P network, as discovered by searching on https://thepiratebay.se. We suppose that this release was made from a promotional DVD, because the commercial DVD was released on December 7 only; we observe no noticeable event on this day.

Our approach is hence relevant for studying logs of search queries, and for detecting bursts of queries related to a same topic.

V. Conclusion and Future Work

We proposed the *Outskewer* method, to detect statistically significant outliers in samples and time series. It uses a novel approach based on the study of the distribution skewness. Our method is easy to interpret because values are classified as *outliers*, *potential outliers* or *not outliers*. The class of all values is unknown when the notion of outlier is not relevant in the considered data set. Our method is also easy to use because it requires no prior knowledge on the data, and the only parameter is the size of the time window for time series. Moreover, it may be used on-line.

We applied it on two data sets representative of many use-cases: evolution of ego-centered views of the internet topology, and logs of queries entered into a search engine. We clearly identify events in the evolution of ego-centered views of the internet topology as shown in Figure 11 and Figure 12. We also automatically detect the release of a pirated version of a film in a P2P system, through the queries entered by users in the search engine, as show in Figure 13.

This paper opens the way to further investigation of the use of the skewness to detect multiple outliers in samples, and to detect events at different time scales in time series. Further studies may also extend our method to detect regime changes.

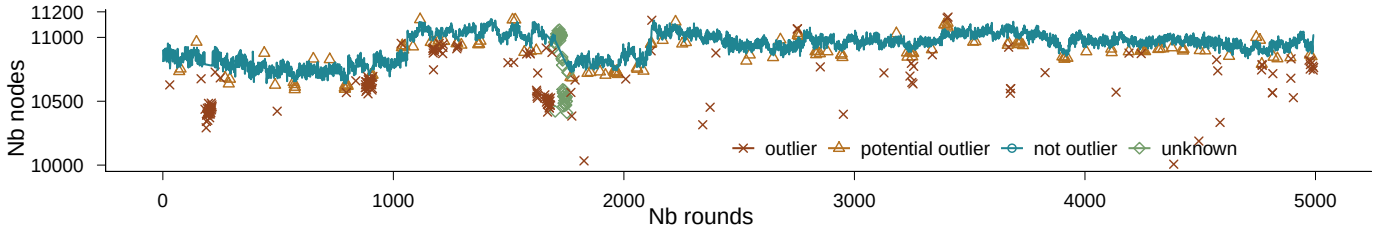

Fig. 11. Number of nodes at each round of radar measurement; outliers are detected using a sliding window of 100 rounds (25 hours).

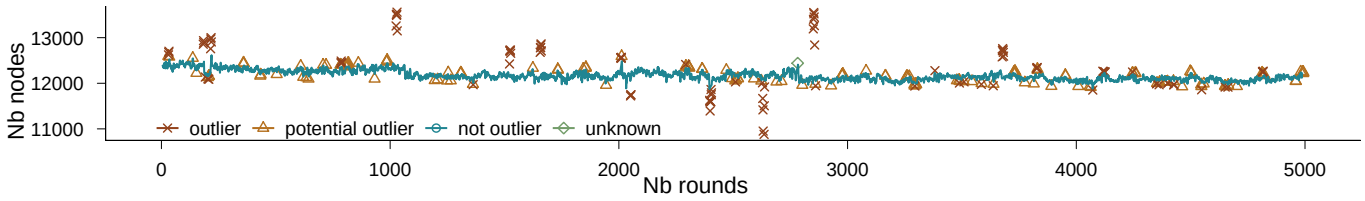

Fig. 12. Number of nodes in the union of 5 consecutive rounds of radar measurement; outliers are detected using a sliding window of 100 rounds.

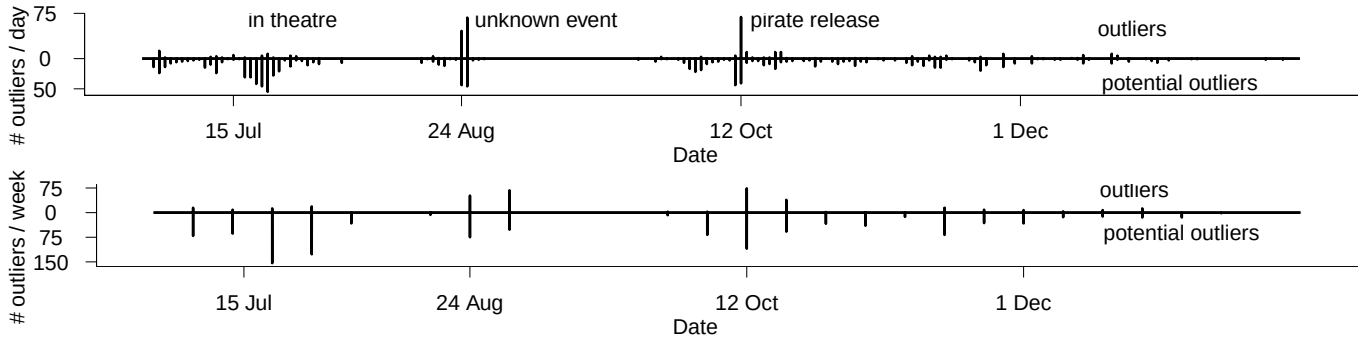

Fig. 13. Number of outliers and potential outliers each day (top) and each week (bottom) in the number of search queries containing "half blood prince".

ACKNOWLEDGMENT

This work is partly supported by the European Commission through the EULER project (grant 258307), part of the Future Internet Research and Experimentation (FIRE) objective of the Seventh Framework Programme (FP7). It is also partly supported by the DynGraph grant from the Agence Nationale de la Recherche with reference ANR-10-JCJC-0202.

REFERENCES

[1] Outskewer source code. [Online]. Available: http://outskewer.sebastien.pro

[2] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, 2004.

[3] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, 1969.

[4] D. M. Hawkins, *Identification of Outliers*. Chapman and Hall, 1980.

[5] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. John Wiley & Sons Ltd., 1994.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection : A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, July 2009.

[7] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, "A survey of outlier detection methods in network anomaly identification," *The Computer Journal*, vol. 54, no. 4, 2011.

[8] I. Ben-gal, "Outlier detection," in *The Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners*, O. Maimon and L. Rockach, Eds. Kluwer Academic Publishers, 2005.

[9] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: fast outlier detection using the local correlation integral," in *Proc. 19th International Conference on Data Engineering (ICDE)*. IEEE, 2003.

[10] *Time series analysis: Forecasting and control*. Holden-Day, 1976.

[11] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proc. 8th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 91–101.

[12] E. Carlstein, "Non-parametric change point estimation," *Annals of Statistics*, pp. 188–197, 1988.

[13] J.-i. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from time series," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 4, pp. 482–492, Apr. 2006.

[14] C. C. Aggarwal, Y. Zhao, and P. S. Yu, "Outlier detection in graph streams," in *Proc. IEEE 27th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, 2011.

[15] L. Akoglu, M. Mcglohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD10)*, 2010.

[16] L. Kovanen, M. Karsai, K. Kaski, J. Kertsz, and J. Saramki, "Temporal motifs in time-dependent networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 11, 2011.

[17] S. M. Bendre and B. K. Kale, "Masking effect on tests for outliers in exponential models," *Journal of the American Statistical Association*, vol. 80, no. 392, 1985.

[18] M. Latapy, C. Magnien, and F. Ouédraogo, "A radar for the internet," *Complex Systems*, vol. 20, no. 1, 2011.

[19] M. Latapy, C. Magnien, and R. Fournier, "Quantifying paedophile activity in a large P2P system," *Information Processing and Management*, to appear. [Online]. Available: http://www-rp.lip6.fr/~latapy/antipaedo/