Computer Networks 57 (2013) 3025–3038

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Data clustering based on correlation analysis applied to highly variable domains

Stefania Tosi *, Sara Casolari, Michele Colajanni

Department of Information Engineering, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Article history:
Available online 14 July 2013

Keywords:
Traffic clustering
High variability
Correlation index
Network management

ABSTRACT

Clustering of traffic data based on correlation analysis is an important element of several network management objectives including traffic shaping and quality of service control. Existing correlation-based clustering algorithms are affected by poor results when applied to highly variable time series characterizing most network traffic data. This paper proposes a new similarity measure for computing clusters of highly variable data on the basis of their correlation. Experimental evaluations on several synthetic and real datasets show the accuracy and robustness of the proposed solution that improves existing clustering methods based on statistical correlations.

1. Introduction

Clustering is a widely adopted approach for augmenting the level of knowledge on rough data. The goals of clustering applied to computer and network datasets can be different, going from Web sites characterization [1], classification of users navigation patterns [4], network traffic classification and management [2]. For example, many network management goals such as flow prioritization, traffic shaping and policing, and diagnostic monitoring as well as many network engineering problems, such as workload characterization and modeling, capacity planning, and route provisioning may benefit from traffic clustering [2].

In this paper, we are interested in correlation-based clustering algorithms applied to highly variable time series. This set of algorithms (e.g., Pearson product moment [7], Spearman and Kendall ranks [8,9]) consider that time series are similar if they exhibit some degree of statistical inter-dependency, and differ from other popular approaches using some geometrical distance (e.g., Euclidean distance [6], cosine distance [5]) as their *similarity measure*. The reason of focusing on correlation similarity measures is distance functions are not always adequate in capturing dependencies among the data. In fact, strong dependencies may exist between time series even if their data samples are far apart from each other as measured by distance functions [3]. In the next section, we will support this statement through a network related example.

The choice and the performance of the similarity measure impact the quality of any clustering algorithm. The better the accuracy and robustness of the measure in finding similarity, the better the quality of the clustering model. Existing correlation indexes are accurate and robust in disclosing similarity except when time series exhibit high variability. This is the case of most traffic data that are highly variable in terms of number of connections, request inter-arrivals, flow sizes (e.g., [13,16,14]). In these scenarios, popular correlation indexes, such as the Pearson coefficient [7], the Spearman rank [8], the Kendall rank [9], and the Local Correlation index [10], show poor results because they are unable to capture correlations even when they exist.

We propose a new similarity measure that is able to disclose correlation even when time series are characterized by high variability. The accuracy and robustness of the proposed correlation index is achieved through an

* Corresponding author. Address: Via Vignolese 905/B, 41125 Modena, Italy, Tel.: +39 0592056273; fax: +39 0592056129.

E-mail addresses: stefania.tosi@unimore.it (S. Tosi), sara.casolari@unimore.it (S. Casolari), michele.colajanni@unimore.it (M. Colajanni).

1389-1286/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2013.07.004

original approach that separates trend from perturbation patterns, and evaluates correlation by computing the similarity of trend patterns. On this basis, clustering models can group time series presenting similarity also when characterized by high variability, such as network traffic [16], workloads [15], and data center resource metrics [11]. Such data may have strong correlations that are masked by perturbations. When some correlations exist, the similarity measure we propose is able to identify them and clustering algorithms can group time series accordingly. The improvements with respect to the state of the art are shown on synthetic and real datasets characterized by high variability.

The remainder of this paper is organized as follows. Section 2 defines the problem of correlation clustering for highly variable datasets. Section 3 presents the proposed algorithm. Section 4 compares the performance of different correlation indexes applied to synthetic time series that represent a fully controlled scenario for evaluation. Section 5 evaluates the proposed algorithm on real scenarios. Section 6 concludes the paper with some final remarks.

2. Problem definition

We define the clustering process based on similarity by considering a dataset $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. For example, it contains all time series of a monitored network, where each time series $\mathbf{x}_j = [x_{j1}, \ldots, x_{jn}]$ is a vector containing a time-ordered discrete sequence of traffic data sampled once. We are interested in partitioning the N time series into K clusters $\mathcal{C} \equiv \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ $(K \leqslant N)$, such that:

1. $\mathcal{C}_i \neq \emptyset, i = 1, \ldots, K$;
2. $\bigcup_{i=1}^{K} \mathcal{C}_i = \boldsymbol{X}$;
3. $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset,\ i, j = 1, \ldots, K$ and $i \neq j$.

The clustering algorithm requires the choice of a similarity measure determining groups of time series so that the similarity between time series within a cluster is larger than the similarity between time series belonging to different clusters. As a similarity measure, we adopt the *correlation index* ρ between two time series $\mathbf{x}_i$ and $\mathbf{x}_j \in \boldsymbol{X}$, where the absolute value of ρ ranges between 0 and 1. When $\rho = 0$, there is no correlation between the two time series, while $\rho = 1$ indicates a complete correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. The literature offers several guidelines for the best interpretation of the value of the correlation measure [17,7], but all criteria depend on the context and purposes of the analysis. In this paper, we do not refer to a specific traffic scenario, hence we can adopt the most general interpretation indicating a *strong correlation* when $\rho > 0.5$, and a *weak correlation* for $\rho \leqslant 0.5$ (e.g., [17]). Different choices for the threshold do not impact the main conclusions of this paper.

This paper proposes a new similarity measure that is able to determine correlation clustering even in datasets exhibiting a high degree of variability, where existing correlation indexes (e.g., [7,10,8,9]) are not accurate. High variability is a typical phenomenon in network-related time series [16] in which most observations take values around the time series trend (*trend pattern*) and some observations depart from it with appreciable frequency, even by assuming extremely large values with non-negligible probability (*perturbation pattern*). Trend patterns represent the tendency of a time series that may be related to the other time series, while perturbation patterns consist of random observations hiding trends. In this paper, we use the standard deviation as the measure of time series variability because a high standard deviation is the most typical trademark of highly variable network measurements [16]. For our purposes, this feature causes trend patterns that are hard to identify because masked by perturbations.

Fig. 1 illustrates some examples of highly variable time series derived from network monitors measuring the number of active connections, active clients and transferred bytes during a day period. In each time series, we can see the presence of different trend patterns during working hours and during the night. These patterns are masked by perturbation patterns. However, there is an evident dependency between the number of active connections and the amount of transferred bytes. In fact, when the number of active connections increases (decreases), the amount of transferred bytes increases (decreases) as well. Despite the variability, we want that a good similarity measure can detect this dependency so to group the two time series in the same cluster. This goal cannot be achieved through distance-based similarity measures because the distance between the sample values of the two time series is not always close, hence the two time series cannot be clustered together through a traditional distance-based clustering model. For this reason, we prefer to consider correlation as the similarity measure, and we propose a correlation-based clustering model that is able to disclose dependency even in highly variable scenarios where distance-based clustering models do not work.

The ability of a correlation index in detecting similarity among correlated time series is measured in terms of *accuracy*. The ability in guaranteeing a stable correlation index when conditions do not change is measured in terms of *robustness*. In the case of highly variable time series, the most popular correlation indexes are affected by two main problems:

1. low accuracy, since they are unable to detect similarities even among correlated time series;
2. low robustness, since they do not guarantee a stable evaluation of the correlation index, even when the relationships between the time series do not change.

Let us give an example of the above problems by referring to the time series shown in Fig. 1 that reports the number of active connections and transferred bytes monitored during a 12-h period in Fig. 2(a). We evaluate the correlation index between the two time series through the Pearson correlation index [7], and we report the results in Fig. 2(b). (The Pearson index is used as an example, but other existing models do not change the conclusions.) Despite the relationship between the two time series, the Pearson model is affected by both issues reported earlier: its results are characterized by low accuracy because the Pearson correlation index remains lower than 0.5 during

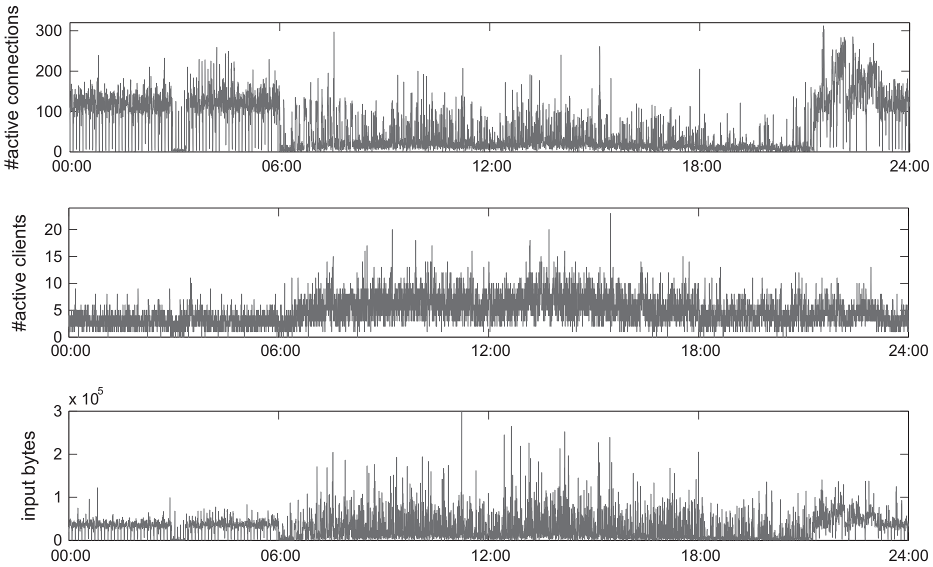
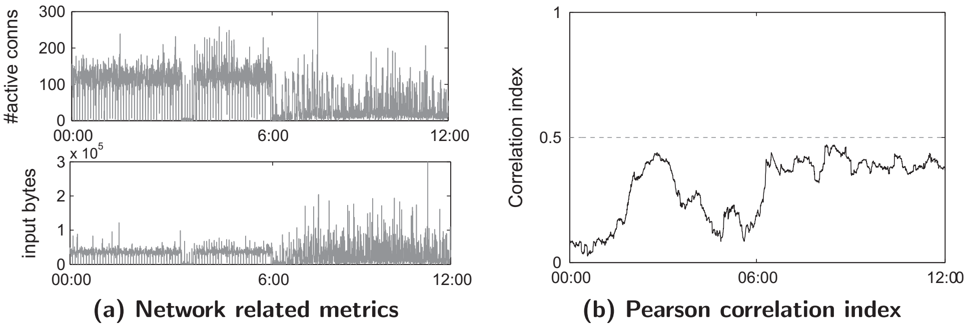


Fig. 1. Examples of network-related time series.



(a) Network related metrics

(b) Pearson correlation index

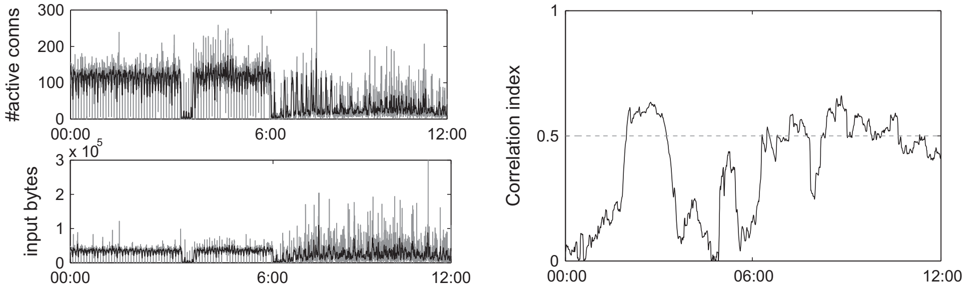
Fig. 2. Correlation index results on highly variable data.

the entire interval of observation; its results are characterized by low robustness because the correlation index presents oscillations even when the correlation between the times series does not change.

A typical approach addressing issues related to high variable time series is to refer to some filtering algorithms (e.g., [18]). In fact, filtering data and then applying some correlation model does not address the above issues, but adds further hard-to-solve problems that are related to the choice of the “right” filter and to the setting of its parameters. Discussing the largely investigated problems related to filters is out of the scope of this paper. We observe the two main conclusions: it is impossible to choose the best filter and its parameters without a preliminary extensive study about data statistical properties [18,11,12]. A similar study is even more complex when data are characterized by high variability. Moreover, the concept of best filter requires an anticipated definition of the context and goals of filtering, that in general is tough or impossible. Basically, we have two alternatives that are equally useless for clustering purposes: to apply a weak or a strong filter. A weak filter removes small variability, hence underlying relationships among time series remain undetectable by existing correlation models; a strong filter may remove important information about trend patterns and thus it prevents the possibility of finding existing correlations.

As an example, let us reconsider the time series represented in Fig. 2(a). We choose a popular filter such as the Exponential Weighted Moving Average (EWMA) [18] as a basis, and we apply different parameters in order to obtain a weak and a strong filter. The resultant time series are shown in Figs. 3(a) and 4(a), respectively. In Fig. 3(b), we report the results of the Pearson index computed on the weakly filtered data. As expected, the results are improved with respect to the not filtered case, but the conclusion remains unchanged: the correlation index has low accuracy since it remains lower than 0.5 for most of the observation interval; it has low robustness since its marked oscillations cause a continuous shift from considering the time series correlated to uncorrelated, then correlated again, and so on.

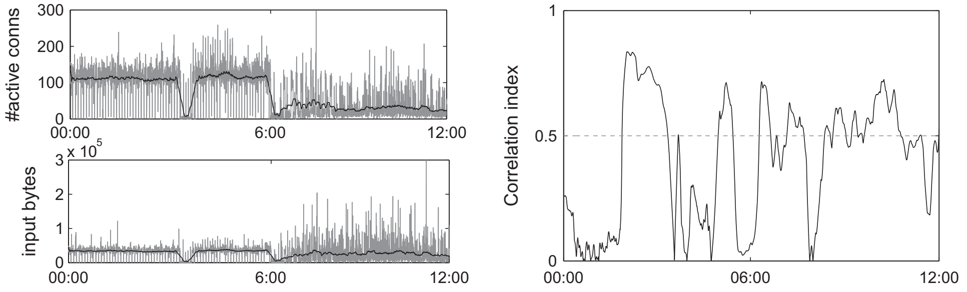
By increasing the filter strength, most perturbations are discarded. The problem is that a strong filter cancels also



(a) Weak filtered network metrics

(b) Pearson correlation index

Fig. 3. Weak filtering applied before correlation analysis.



(a) Strong filtered network metrics

(b) Pearson correlation index

Fig. 4. Strong filtering applied before correlation analysis.

information about trend patterns characterizing the time series. The result is that the Pearson correlation index becomes even less accurate and less robust, as we see in Fig. 4(b).

All these results evidence that high variability represents a limit of existing correlation indexes even when they are integrated with filtering techniques. These limits of existing correlation indexes affect the quality of the results of clustering models based on similarity when applied to highly variable datasets. In these contexts, the risk is that time series characterized by strong correlation may be clustered in different groups, and uncorrelated time series may be assigned to the same cluster. These reasons motivate the need of a new similarity measure that is able to disclose correlations in an accurate and robust way in highly variable contexts. The model proposed in the following section has several benefits: it does not require any assumption about statistical properties, any pre-analysis of time series characteristics, and any integration with pre-filtering techniques. Moreover, it is able to adapt its parameters to data characteristics, to capture linear and non-linear dependences, and to cluster time series basing on their statistical correlation in an accurate and robust way.

3. Clustering of highly variable datasets

In this section, we present a novel correlation index, named *CoHiVa* (*Co*rrelation for *Hi*ghly *Va*riable data), that may be used as the similarity measure of different clustering algorithm(s) applied to highly variable time series. CoHiVa may be viewed as an improvement of the LoCo score [10] that evaluates correlation through the analysis of pattern similarity. CoHiVa extends this idea to the correlation analysis in highly variable domains. Unlike the Pearson model [7] that works well only if time series are linked by a linear relationship, CoHiVa does not assume any data dependency and it is able to capture linear and non-linear dependencies. Moreover, CoHiVa does not assume any data distribution, as required by the Spearman and Kendall ranks [8,9].

The described clustering algorithm adopts CoHiVa to modify the complete-linkage clustering model [19], but other clustering algorithms using similarity measures can be considered as well. The proposed algorithm denotes N singleton clusters $\{\mathcal{C}_1,\ldots,\mathcal{C}_N\}$, each corresponding to a monitored time series ($\mathcal{C}_i \equiv \mathbf{x}_i, i = 1,\ldots,N$). It then computes the $N \times N$ similarity matrix $\mathcal{D}$ containing the CoHiVa correlation indexes $\rho(\mathbf{x}_p, \mathbf{x}_q)$ between all pairs of time series $\mathbf{x}_p$, $\mathbf{x}_q$ in $\boldsymbol{X}$ through the following procedure.

Let us consider one time series $\mathbf{x}_p = [x_{p1},\ldots,x_{pn}]$ of the pair. The first goal is to identify the main patterns in $\mathbf{x}_p$, where the patterns correspond to trends (i.e., periodic and seasonal components) and perturbations. To this end, we apply the Singular Value Decomposition (SVD) [10] to the auto-covariance matrix of the time series. Among the spectral decomposition techniques, SVD is considered as the baseline technique for separating existing patterns without any assumption about the statistical characteristics of data [20,21].

In practice, we estimate the full auto-covariance matrix of the time series $\mathbf{x}_p$, that is defined as:

$$\mathbf{\Phi}(x_p) = \mathbf{x}_p \otimes \mathbf{x}_p, \tag{1}$$

where $\mathbf{\Phi}(x_p)$ is the auto-covariance matrix of $\mathbf{x}_p$.

Then, we compute the SVD of the auto-covariance matrix $\mathbf{\Phi}(x_p)$ as follows:

$$\mathbf{\Phi}(x_p) = \mathbf{U}(x_p)\mathbf{\Sigma}(x_p)\mathbf{V}(x_p)^T, \tag{2}$$

where $\mathbf{U}(x_p)$,$\mathbf{\Sigma}(x_p)$ and $\mathbf{V}(x_p) \in \mathbb{R}^{n\times n}$.

The columns $\mathbf{v}_i$ of $\mathbf{V}(x_p) \equiv [\mathbf{v}_1,\ldots,\mathbf{v}_n]$ are the right singular vectors of $\mathbf{\Phi}(x_p)$. Similarly, the columns $\mathbf{u}_i$ of $\mathbf{U}(x_p) \equiv [\mathbf{u}_1,\ldots,\mathbf{u}_n]$ are the left singular vectors of $\mathbf{\Phi}(x_p)$. Finally, $\mathbf{\Sigma}(x_p) \equiv \mathrm{diag}[s_1,\ldots,s_n]$ is a diagonal matrix with positive values s_i, called the singular values of $\mathbf{\Phi}(x_p)$.

The singular vectors corresponding to small singular values are composed by errors that usually contaminate the measured variables [24]. The contribution of these errors must be discarded by eliminating the singular vectors corresponding to the smallest singular values [25]. By retaining just the principal vectors corresponding to the highest k singular values ($k < n$) we can reconstruct a k-*dimensional approximation* of the correlation matrix:

$$\overline{\mathbf{\Phi}}(x_p) \equiv \overline{\mathbf{U}}(x_p)\overline{\mathbf{\Sigma}}(x_p)\overline{\mathbf{V}}(x_p)^T, \tag{3}$$

where $\overline{\mathbf{U}}(x_p) \equiv [\bar{\mathbf{u}}_1,\ldots,\bar{\mathbf{u}}_k]$, $\overline{\mathbf{V}}(x_p) \equiv [\bar{\mathbf{v}}_1,\ldots,\bar{\mathbf{v}}_k]$ and $\overline{\mathbf{\Sigma}}(x_p) \equiv \mathrm{diag}[\bar{s}_1,\ldots,\bar{s}_k]$.

Literature on SVD gives little importance to the problem of dynamically selecting the appropriate number of principal vectors that capture the patterns (e.g., [10,20]). A common approach is to choose a fixed number of principal vectors independently of data characteristics, but this choice is unsuitable to time varying contexts where the statistical properties of data may change frequently. On the other hand, our algorithm chooses a variable number of principal vectors that takes into account the statistical characteristics of the considered data. To this purpose, we select the principal vectors accounting for 90% of the variation in the analyzed time series as in [23]. In this way, the number of selected principal vectors varies from time series to time series according to data characteristics. The variable number of principal vectors used to capture the main patterns of the time series is denoted by k.

In order to find correlations also in the case of high variability, we analyze the main patterns of $\mathbf{x}_p$ and retain just trend patterns because they represent the tendency of a time series that may be related to the other time series. The problem is that in contexts characterized by high variability it is expected that the extraction of main patterns includes also some perturbation patterns among the k principal vectors [26]. As these last patterns prevent the identification of trends and their possible correlations, our algorithm discards them. The idea is to build a new matrix based on a subspace of $\hat{k} \leqslant k$ principal vectors that capture just trend patterns. To remove perturbation patterns from $\overline{\mathbf{U}}(x_p)$ in (3), our algorithm computes the Hurst exponent H of the k principal vectors by means of the rescaled range analysis of Hurst [27]. The Hurst exponent measures whether the data have pure random variability or some underlying trends [28]. For each principal vector $\bar{\mathbf{u}}_i \in \overline{\mathbf{U}}(x_p)$ of length n, we first define its cumulative deviate series:

$$Z_t = \sum_{j=1}^{t}(\bar{u}_j - m), \tag{4}$$

where $t = 1,2,\ldots,n$, $\bar{u}_j$ is the jth element of the ith principal vector, and m is the mean of $\bar{\mathbf{u}}_i$.

To compute the rescaled range $\frac{R(n)}{S(n)}$, we calculate the range:

$$R(n) = \max(Z_1, Z_2, \ldots, Z_n) - \min(Z_1, Z_2, \ldots, Z_n), \tag{5}$$

and the standard deviation:

$$S(n) = \sqrt{\frac{1}{n}\sum\nolimits_{j=1}^{n}(\bar{u}_j - m)^2}. \tag{6}$$

The *Hurst exponent H* is defined in terms of the asymptotic behavior of the rescaled range as a function of the time span of a time series as follows [27]:

$$\mathrm{E}\left[\frac{R(n)}{S(n)}\right] = Cn^H \quad \text{as } n \to \infty, \tag{7}$$

where $E\left[\frac{R(n)}{S(n)}\right]$ is the expected value of the rescaled range, n is the number of observations in a time series, and C is a constant.

If the estimated Hurst exponent H_i of a principal vector $\bar{\mathbf{u}}_i \in \overline{\mathbf{U}}(x_p)$ is close to 0.5, then we can conclude that $\bar{\mathbf{u}}_i$ contains perturbations. On the other hand, if the Hurst exponent remains far from 0.5, then we can assume that $\bar{\mathbf{u}}_i$ is a trend principal vector.

Our model builds up a new $\widehat{\mathbf{U}}(x_p)$ matrix containing only the $\hat{k}$ trend principal vectors $\hat{\mathbf{u}}_l, l = 1, \ldots, \hat{k}$ as follows:

$$\forall \bar{\mathbf{u}}_i \in \overline{\mathbf{U}}(x_p), \quad i = 1, \ldots, k:$$

$$\hat{\mathbf{u}}_l \equiv \bar{\mathbf{u}}_i \quad \text{if } H_i < 0.5 - \frac{\delta}{2} \quad \text{or} \quad H_i > 0.5 + \frac{\delta}{2}, \tag{8}$$

where δ is a two-sided 95% confidence interval computed on the number of samples [28].

This separation between perturbation and trend patterns allows us to remove perturbation patterns in the time series and to focus only on trends. By focusing on the $\hat{k}$ trend principal vectors, we are able to construct a new approximation of the $\mathbf{\Phi}(x_p)$ matrix that we name *trend approximation*. Given $\widehat{\mathbf{U}}(x_p) \equiv [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{\hat{k}}]$, the corresponding singular values and the right singular vectors form the matrices $\widehat{\mathbf{\Sigma}}(x_p) \equiv \mathrm{diag}[\hat{s}_1, \ldots, \hat{s}_{\hat{k}}]$ and $\widehat{\mathbf{V}}(x_p) \equiv [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_{\hat{k}}]$, respectively. Through these matrices, the trend approximation of the correlation matrix using only trend patterns is given by:

$$\widehat{\mathbf{\Phi}}(x_p) \equiv \widehat{\mathbf{U}}(x_p)\widehat{\mathbf{\Sigma}}(x_p)\widehat{\mathbf{V}}(x_p)^T. \tag{9}$$

The matrix $\widehat{\mathbf{\Phi}}(x_p)$ approximates the trend behavior of $\mathbf{\Phi}(x_p)$ by removing error information and perturbation patterns that affect the identification of the trends of the time series.

After the extraction of the main trends from the time series $\mathbf{x}_p$, we evaluate whether it is correlated or not with the other time series $\mathbf{x}_q$ of the pair by computing how close their trends are. When the trend approximation matrices

$\widehat{\mathbf{U}}(x_p)$ and $\widehat{\mathbf{U}}(x_q)$ are similar, the time series $\mathbf{x}_p$ and $\mathbf{x}_q$ follow similar (linear or non-linear) trends, and we can guess that the two time series are correlated. In geometric terms, if two time series are correlated, then the trend principal vectors of one time series should lie within the subspace spanned by the trend principal vectors of the other time series. For this reason, we compute the CoHiVa correlation index between $\mathbf{x}_p$ and $\mathbf{x}_q$ by projecting the trend principal vectors of the time series $\mathbf{x}_p$ into the trend principal vectors of $\mathbf{x}_q$, as following:

$$\rho(\mathbf{x}_p,\mathbf{x}_q) = \frac{1}{\hat{k}(x_p)+\hat{k}(x_q)}(\|\widehat{\mathbf{U}}(x_p)^T\widehat{\mathbf{U}}(x_q)\| + \|\widehat{\mathbf{U}}(x_q)^T\widehat{\mathbf{U}}(x_p)\|), \tag{10}$$

where $\hat{k}(x_p)$ and $\hat{k}(x_q)$ are the amounts of trend principal vectors of $\mathbf{\Phi}(x_p)$ and $\mathbf{\Phi}(x_q)$, respectively, while $\widehat{\mathbf{U}}(x_p)$ and $\widehat{\mathbf{U}}(x_q)$ are the trend principal vectors matrices collecting them.

Through this procedure, our algorithm finds out the CoHiVa correlation indexes $\rho(\mathbf{x}_p, \mathbf{x}_q)$ between all pairs of time series in $\mathbf{X}$ and uses them to fill the $N \times N$ similarity matrix $\mathcal{D}$. These correlation values represent the measures of similarity between all the clusters.

We then proceed through the following steps:

1. Find the most correlated pair of clusters in the similarity matrix:

$$\rho(\mathbf{x}_r,\mathbf{x}_s) = \max_{1\leqslant p,q\leqslant N, p\neq q} \rho(\mathbf{x}_p,\mathbf{x}_q). \tag{11}$$

2. Combine cluster $\mathcal{C}_r$ and cluster $\mathcal{C}_s$ to form a new cluster, denoted as $\mathcal{C}_{(r,s)}$.
3. Update the similarity matrix $\mathcal{D}$ by deleting the rows and columns corresponding to clusters $\mathcal{C}_r$ and $\mathcal{C}_s$ and by adding a row and a column corresponding to the early created cluster. The similarity between the new cluster $\mathcal{C}_{(r,s)}$ and a generic old cluster $\mathcal{C}_i$ is defined as:

$$\rho(\mathbf{x}_i,\mathbf{x}_{(r,s)}) = \max(\rho(\mathbf{x}_i,\mathbf{x}_r),\rho(\mathbf{x}_i,\mathbf{x}_s)). \tag{12}$$

4. Repeat steps (1)–(3) until all objects are in a cluster.

This procedure results as an organized tree built up according to the similarity matrix computed through CoHiVa. Cutting the tree at a given height gives a partition clustering at a selected precision ϕ. For example, if we discriminate between weak and strong correlation through the value of ϕ = 0.5, we cut the tree when the highest correlation value found in the similarity matrix goes below ϕ = 0.5. The time series forming a sub-tree after this cut are considered part of the same cluster [22].

4. Performance evaluation

The quality of the proposed clustering algorithm strongly depends on the performance of the similarity measure. Hence, in this section we evaluate the performance of the CoHiVa index in finding correlation, and we compare it against the results of the following state-of-the-art alternatives: the Pearson product moment (*Pearson*) [7], the Spearman rank (*Spearman*) [8], the Kendall rank (*Kendall*) [9], and the Local Correlation (*LoCo*) index [10]. In order to stress the importance of managing the time series variability for disclosing correlation, we also report the results that we would obtain by avoiding the trend patterns selection step based on the Hurst rescaled range analysis in our model (*CoHiVa without Hurst analysis*). Moreover, we make our evaluation even more exhaustive by comparing the performance of CoHiVa to that of a state-of-the-art model that is integrated with a pre-filtering technique (*Pearson with filtering*).

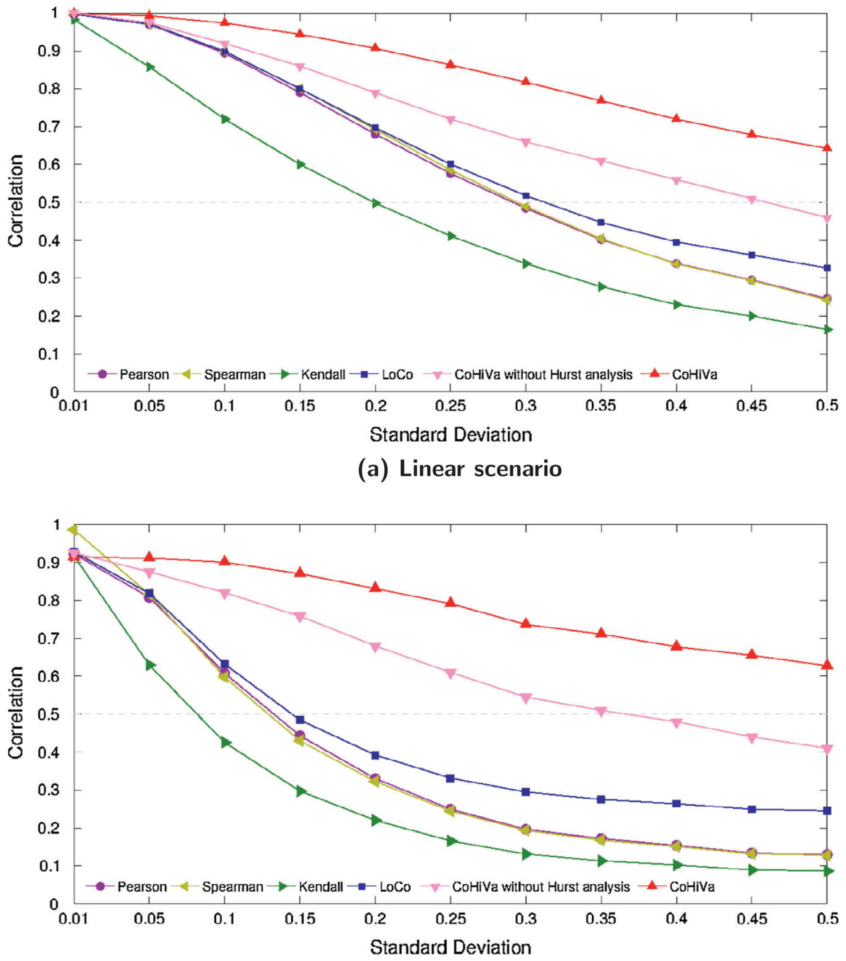
To evaluate the accuracy and robustness of the considered indexes we initially refer to synthetic time series that allow us to have full control on their actual degree of correlation. The considered data refer to three types of time series: (1) correlated with linear dependence, (2) correlated with non-linear dependence, and (3) not correlated.

The time series of each scenario take values in the range [0,1]. In order to evaluate the ability of the correlation indexes in capturing different types of dependency for different levels of variability, we introduce perturbations from $N(0,\ \sigma)$, where $\sigma \in \{0.01, 0.05, 0.1, \ldots, 0.5\}$ is the standard deviation that quantifies the intensity of perturbations added to data [29,30]. We remind that we consider a strong correlation when $\rho > 0.5$, and a weak correlation for $\rho \leqslant 0.5$ [17] although this choice does not affect the main conclusions of this paper. The performance of the correlation index is evaluated in terms of accuracy and robustness over 1000 independent generations of time series for each scenario.

4.1. Accuracy

We define the *accuracy* of a correlation index as its ability in capturing correlation when data present some linear or non-linear relationships, and in categorizing as not correlated time series having no dependence. For example, an accurate index should obtain a correlation value close to 1 in the two correlated scenarios, and a value close to 0 in the not correlated scenario. The first set of experiments evaluates the accuracy of the correlation indexes when the time series are characterized by different intensities of perturbations in the three scenarios.

The performance of the considered correlation indexes in the linear scenario is reported in Fig. 5(a). As expected, we observe a decrease of the accuracy of each index for increasing values of σ, but the impact of perturbations differs substantially for different indexes. When the dispersion is very low ($\sigma \leqslant 0.1$), all indexes are able to capture the strong correlation between data. When the dispersion increases ($\sigma > 0.1$), the Kendall rank is the first losing its ability of detecting time series correlation. In higher variable contexts ($\sigma > 0.15$), only the CoHiVa index captures the strong data correlation, thanks to a value always higher than 0.6. We can also appreciate the benefits achievable by performing the Hurst rescaled range analysis for retaining only trend patterns in the time series. If we skip this step, CoHiVa without Hurst analysis still improves state-of-the-art models performance but it cannot capture the strong correlation between time series when the data variability reaches critical levels ($\sigma > 0.45$).



(a) Linear scenario

(b) Non-linear scenario

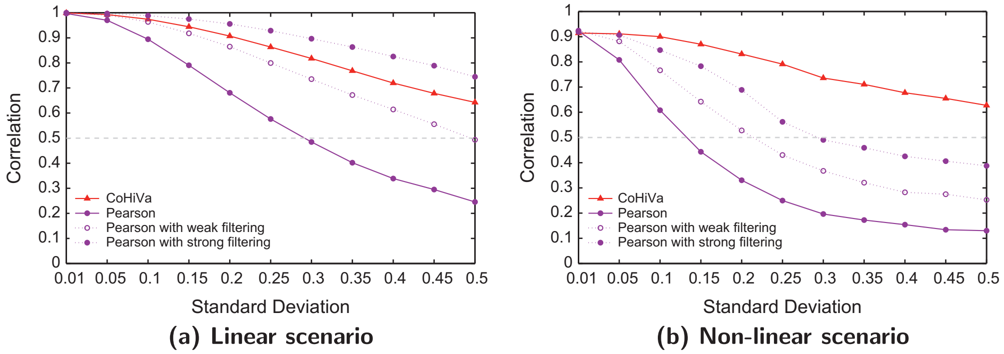
Fig. 5. Accuracy of the correlation indexes.

The accuracy of the indexes deteriorates when we pass to a scenario where the correlation between time series is non-linear. A comparison between Fig. 5(a) and (b) gives a first idea about the overall results. Only the CoHiVa index is able to detect a strong correlation for any σ when the relationship between data is non-linear, also thanks to the Hurst-based trend pattern selection that guarantees high correlation results for each perturbation level. On the other hand, all existing indexes are affected by a low accuracy for increasing values of σ. (They estimate a weak correlation even when time series are perturbed by very low levels of dispersion, such as $\sigma = 0.15$.) It is also interesting to observe that the Spearman rank, which is specifically oriented to capture non-linear dependencies [8], exhibits the best accuracy when the dispersion is very low (that is, $\sigma < 0.05$), but it loses its capacity as soon as the time series are characterized by higher perturbations.

To address issues related to high variability, the state-of-the-art models may increase their accuracy by working on a filtered representation of the original time series. We anticipated in Section 2 that this approach does not work well, but for the sake of an exhaustive comparison we compare the performance of CoHiVa against a Pearson model combined with a pre-filtering technique. We have to specify that the choice of the best filtering model and of its parameters is a serious issue by itself, and is out of the scope of this paper.

We integrate the Pearson correlation model with an EWMA filter that we have experimentally evaluated as giving good results. We do not claim that we are applying an optimal filter with optimal parameter setting, even because the definition of optimum is improper in this context.

Fig. 6 shows the results obtained by applying the Pearson model to data filtered through a weak and a strong filter. If we compare the results of Pearson without filtering to the results of Pearson with filtering, we can appreciate that the filter in fact improves accuracy: the correlation value is higher for every σ and for linear and non-linear scenarios. In the linear scenario shown in Fig. 6(a), the Pearson model with filtering is able to detect a correlation index higher than 0.5 for any σ. On the other hand, if we consider the non-linear scenario shown in Fig. 6(b), there is a drastic decrease of the correlation index. Both weak and strong fil-



(a) Linear scenario

(b) Non-linear scenario

Fig. 6. Accuracy of the correlation indexes using data filtering.

tering are useless because they estimate a $\rho \leqslant 0.5$ when $\sigma > 0.2$ and $\sigma > 0.3$, respectively.

These results demonstrate that filters do not guarantee accurate results, besides the further problems related to the choice of the best filter and of its parameters in highly variable contexts.

We finally report some results of the considered correlation indexes applied to a scenario characterized by synthetic time series characterized by no dependence. These results complete the accuracy evaluation of the indexes, because we expect that an accurate index can also detect the absence of correlation. Despite the level of variability, we see in Fig. 7 that all the indexes are accurate and detect a weak correlation between time series having no dependence. These results confirm that, even though CoHiVa assumes high values when applied to correlated time series, its values does not remain high when applied to time series having no dependence. As well as existing correlation indexes, our index is able to avoid to detect correlation when applied to uncorrelated datasets.

4.2. Robustness

The accuracy of a correlation index must be combined with information about its *robustness*, that assesses the reliability of correlation results across different evaluations. We quantify the robustness in terms of *coefficient of variation* (CoV) for different evaluations. The coefficient of variation is defined as the ratio of the standard deviation to the mean of the correlation values over all the experiments. A lower CoV denotes a better robustness of the correlation index.

We evaluate the robustness of the results obtained in Section 4.1. Table 1 reports the CoV of each considered correlation index applied to time series in a linear scenario. The columns refer to the increasing values of perturbations intensity σ, while the rows report the correlation indexes. The CoV of all correlation indexes increases when σ increases. Compared to existing solutions, the CoHiVa index is able to keep the lowest CoV for any σ value. Thanks to a CoV always lower than 0.15, the proposed correlation index guarantees high robustness in capturing linear correlations also among highly variable data. Moreover, we see that the robustness of our model benefits from performing the rescaled range analysis of Hurst.

As expected, a non-linear context worsens the robustness of all the indexes. This main conclusion is confirmed by the CoV values reported in Table 2. These results demonstrate that only the CoHiVa model is able to guarantee a CoV lower than 0.2 for any perturbation intensity. On the other hand, state-of-the-art models show poor results even for medium–low values of σ ($\sigma \leqslant 0.2$). With the

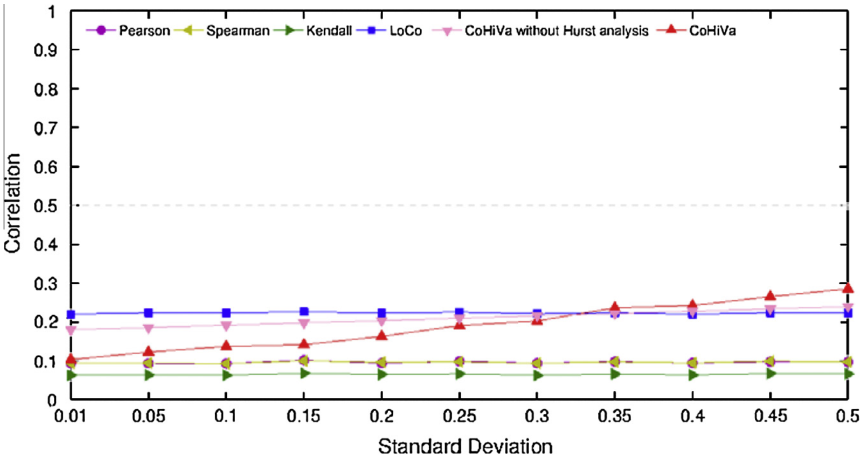


Fig. 7. Analysis of accuracy in a not correlated scenario.

Table 1
Coefficient of variation in the linear scenario.

	σ					
	0.01	0.1	0.2	0.3	0.4	0.5
Pearson	0.0232	0.0299	0.0914	0.1992	0.3437	0.4817
Spearman	0.0227	0.0304	0.0905	0.2036	0.3496	0.4874
Kendall	0.0371	0.0486	0.1098	0.2170	0.3606	0.4936
LoCo	0.0220	0.0284	0.0835	0.1653	0.2452	0.2735
Pearson with weak filtering	0.0001	0.0069	0.0324	0.0785	0.1206	0.1787
Pearson with strong filtering	0.0001	0.0123	0.0497	0.0917	0.1318	0.1736
CoHiVa without Hurst analysis	0.0127	0.0190	0.0342	0.0688	0.1301	0.2405
CoHiVa	0.0073	0.0086	0.0217	0.0498	0.0888	0.1343

Table 2
Coefficient of variation in the non-linear scenario.

	σ					
	0.01	0.1	0.2	0.3	0.4	0.5
Pearson	0.0274	0.1296	0.3704	0.5843	0.6838	0.7052
Spearman	0.0266	0.1457	0.3894	0.5919	0.6892	0.7104
Kendall	0.0391	0.1632	0.4014	0.5997	0.6960	0.7194
LoCo	0.0258	0.1136	0.2550	0.2923	0.3073	0.2924
Pearson with weak filtering	0.0070	0.0835	0.2026	0.2215	0.2224	0.2236
Pearson with strong filtering	0.0083	0.0944	0.2120	0.2468	0.2434	0.2473
CoHiVa without Hurst analysis	0.0407	0.0392	0.1748	0.2083	0.2205	0.2881
CoHiVa	0.0380	0.0172	0.1467	0.1614	0.1711	0.1926

exception of LoCo and Pearson model integrated with filtering, all the other correlation indexes are quite unreliable in highly variable contexts because they reach CoV values around 0.7. These results confirm that they cannot be used to capture non-linear relationships among highly variable time series for clustering purposes.

Our analyses confirm that the most popular correlation indexes are affected by scarce accuracy and robustness when data exhibit high variabilities and/or non-linear dependency. The main result is that the proposed CoHiVa index is able to guarantee good performance for any considered scenario and represents a good choice as the similarity measure to be used in clustering algorithms working on highly variable datasets.

5. Experimental results

We evaluate the quality of the proposed clustering algorithm by referring to two network-related datasets characterized by high variability: Abilene network traffic, and measurements collected at the border router of our university.

- Abilene network.
  The publicly available Abilene dataset contains aggregate data based on measurements of origin–destination (OD) flows on the Abilene network [31]. We consider sampled data from every router over a 7-day period, starting December 12, 2003.[1] At sampling period of 5 min, each link produces 2016 samples a week.
- University network.
  The dataset is obtained from a monitor attached to a border router of the university and contains flows characterized by different metrics, such as total number of packets (excluding ack packets), packet size statistics (mean, minimum, maximum, quartiles), number of bytes transferred in each direction, number of active connections and number of active clients. These network metrics are aggregated every 10 s. The presented results refer to 1 day characterized by 8640 samples for each network metric.

The clustering algorithms applied to these two datasets are used for two different purposes: for traffic clustering of Abilene network data (Section 5.1), and for server clustering of university network data (Section 5.2).

5.1. Traffic clustering

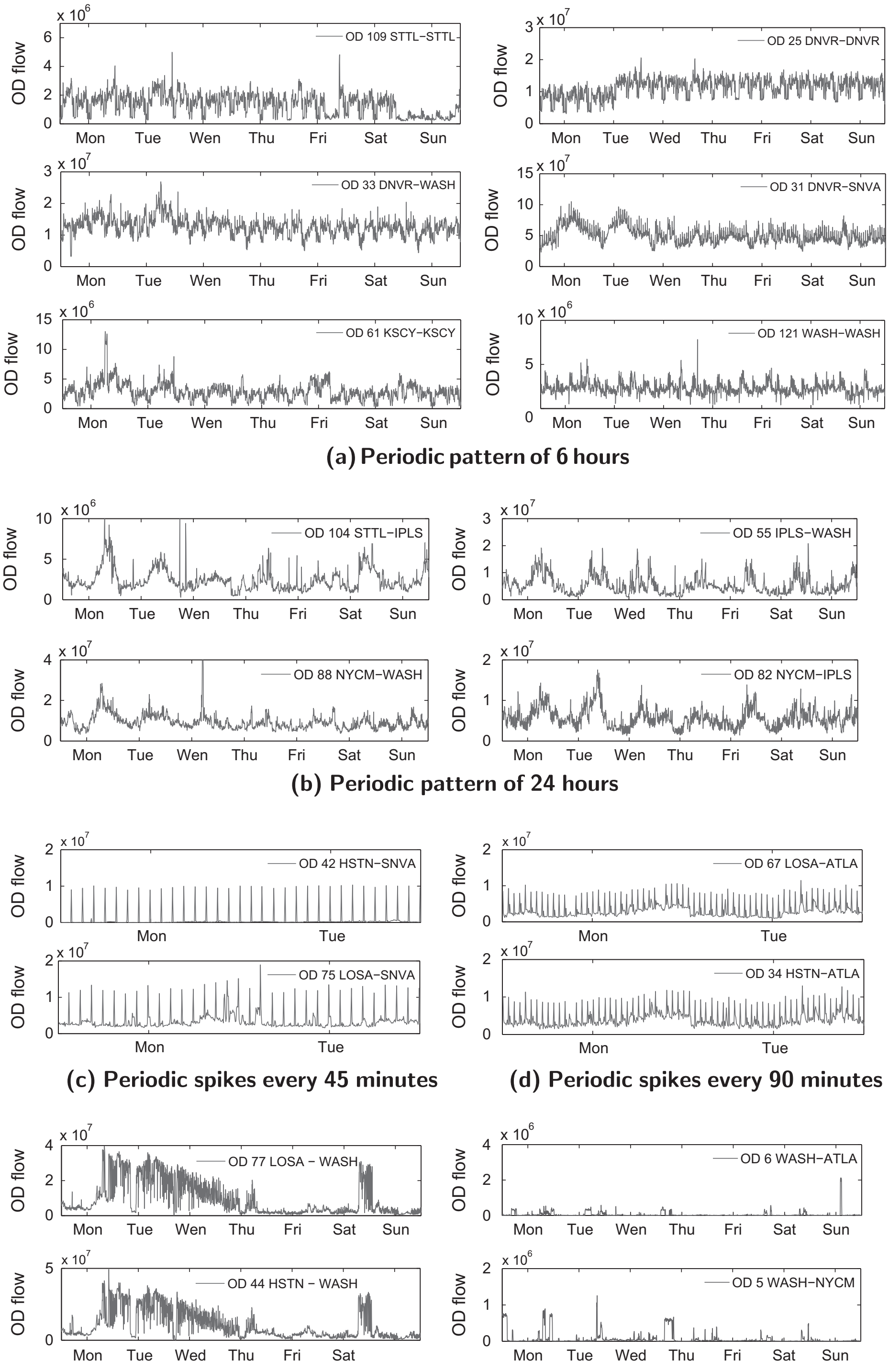
The identification of the main statistical properties of traffic flows and the clustering of flows based on such properties are crucial to many network management tasks and network engineering problems [2]. Due to the high variability of OD traffic flows [16], clustering solutions must be able to group data even in highly variable contexts. Our algorithm represents a good solution for traffic clustering because it is able to identify correlation between OD flow traffic patterns and to cluster together flows presenting similar statistical properties.

We carried out a preliminary analysis of the Abilene dataset with the goal of extracting the main statistical properties shared by the OD traffic flows that we expect

[1] Data available at: http://math.bu.edu/people/kolaczyk/datasets.html.

to be shared by the time series clustered together. This analysis evidences the presence of the following six main statistical properties:

1. *Periodic pattern of 6 h*: some OD flows have fluctuations of request volume over a time period of 6 h, typically related to hourly human behaviors. Some intra-daily periodic patterns are shown in Fig. 8(a).
2. *Periodic pattern of 24 h*: some OD flows present the typical increase of user requests during working hours and a decrease during the night. Some examples of daily patterns are reported in Fig. 8(b).



(a) Periodic pattern of 6 hours

(b) Periodic pattern of 24 hours

(c) Periodic spikes every 45 minutes

(d) Periodic spikes every 90 minutes

(e) Aperiodic trends

(f) Stochastic patterns

Fig. 8. Characteristics of Abilene traffic flows.

3. *Repeated spikes every 45 min*: some OD flows manifest bursts of requests repeated every 45 min. Fig. 8(c) shows this behavior during a time interval of 2 days.
4. *Repeated spikes every 90 min*: some OD flows exhibit regular spikes every hour and a half as in Fig. 8(d).
5. *Aperiodic trend*: some OD flows are characterized by increasing or decreasing trends even thought they do not manifest any periodicity. Some flows with these characteristics are presented in Fig. 8(e).
6. *Stochastic pattern*: some OD flows are characterized by minor bursts and irregular behavior as those in Fig. 8(f).

This preliminary analysis gives us the ground truth. In other words, we can expect that the most effective clustering algorithm is able to find out 6 clusters, each one including all and only the OD flows sharing just one of the identified statistical properties. Thanks to this term of comparison, we can compute the *recall* and the *precision* [32] of the other clustering algorithms.

The *recall* measures the ability of an algorithm to cluster a flow presenting a statistical property together with other flows having that statistical property (e.g., a flow with a 24-h periodic pattern is clustered together with the flows having a daily period). To achieve a recall of 100%, the clustering algorithm must insert each flow in the cluster representing the corresponding statistical property.

The *precision* gives information about the ability of the clustering algorithm to limit the number of flows that present a statistical property but are clustered together with flows characterized by a different statistical property (e.g., a flow with a 24-h periodic pattern is clustered together with 6-h periodic flows). A precision of 100% means that the algorithm inserts into a cluster only the flows with the corresponding statistical property.

We compare the results of the complete-linkage clustering algorithm using CoHiVa against those obtained through the Pearson product moment and the LoCo score, that achieved the best trade-off between accuracy and robustness on synthetic settings. Table 3 reports the recall and precision values for the three considered algorithms. This table evidences three main results.

- The ability of CoHiVa in disclosing the presence of all 6 clusters, since it obtains high recall and precision values over all the clusters. The null values obtained by Pearson in both precision and recall over the cluster with daily periodic flows mean that this index does not reveal the presence of correlated flows having daily pat-

Table 3
Recall and precision.

Cluster		CoHiVa (%)	Pearson (%)	LoCo (%)
Periodic pattern of 6 h	*Recall*	100	100	100
	Precision	89	56	89
Periodic pattern of 24 h	*Recall*	91	0	81
	Precision	83	0	78
Repeated spikes every 45 min	*Recall*	75	100	0
	Precision	60	40	0
Repeated spikes every 90 min	*Recall*	100	100	0
	Precision	100	100	0
Aperiodic trend	*Recall*	100	100	0
	Precision	100	67	0
Stochastic pattern	*Recall*	81	49	67
	Precision	91	100	83

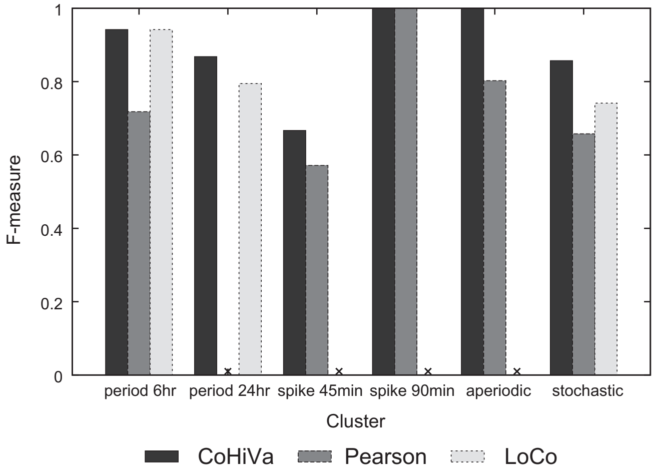


Fig. 9. F-measure.

terns. LoCo is unable to identify spiky and aperiodic flows because it is characterized by null recall and precision values over these three clusters. These results are in accord to those achieved on synthetic settings in the previous section: Pearson is unable to identify correlation among highly variable data (e.g., flows with a 24-h period), while LoCo fails in finding spiky and aperiodic flows because it considers just the first principal component not including information of those irregular patterns.

- The ability of CoHiVa in discriminating between correlated and not correlated flows. This algorithm guarantees high recall and precision values in the identification of correlated flows belonging to the first 5 clusters and the best performance for the identification of uncorrelated stochastic flows. On the other hand, Pearson and LoCo insert in the same cluster many flows that are correlated with other flows, as their lower recall values over the stochastic cluster demonstrate.
- The ability of CoHiVa in guaranteeing the best compromise between the capacity of clustering together flows having similar statistical properties and the capacity of limiting the number of flows that are wrongly classified.

This last result can be appreciated by introducing a further performance measure. Since a trade-off between recall and precision values exists, these two metrics can be combined into one measure, namely the *F-measure* [32], that gives a global estimation of the quality of the clustering algorithm through the weighted harmonic mean of precision and recall, that is:

$$F\text{-}measure = 2\frac{precision * recall}{precision + recall} \qquad (13)$$

The closest the F-measure to 1, the highest the quality of the clustering algorithm.

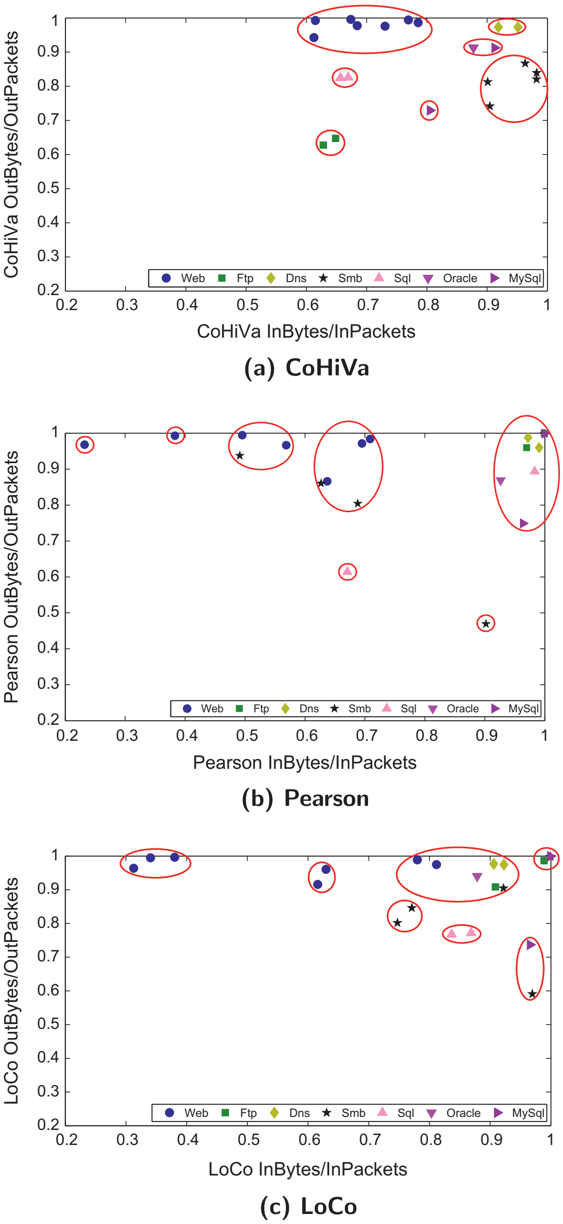
The combined effect of recall and precision can be appreciated in Fig. 9 showing that CoHiVa achieves the best F-measure in all cluster identification. This is an important result about the robustness of CoHiVa: independently of the flow statistical properties, CoHiVa guarantees the best compromise between the number of correctly and wrongly clustered flows. Pearson and LoCo achieve some good results but they show also unacceptable performance in other cases. In particular, the null F-measure values of Pearson and LoCo indexes in the identification of flows with highly variable, spiky and aperiodic flows limit their applicability as similarity measures for correlation-based clustering algorithms.

5.2. Server clustering

As a further test case, we apply the clustering algorithms to the identification of network-based servers having similar statistical behavior. The idea is to group servers performing similar tasks through the analysis of the network traffic flows they generate. Even these scenarios are characterized by highly variable measures gathered by a network monitor connected to the border router of our university. We select 21 destination servers by considering the destination address and the port number of the traffic flows. Table 4 reports the different classes of servers and the number of monitored servers for each class. We expect that a good clustering algorithm is able to group together hosts supporting same server processes. In other words, we expect that Web servers are grouped in the same clus-

Table 4
Server classes.

	Web	DNS	FTP	Samba	Oracle	MySql	Microsoft Sql
# Servers	7	2	2	5	1	2	2



(a) CoHiVa

(b) Pearson

(c) LoCo

Fig. 10. Server clustering results.

ter, and that this cluster is separated from the cluster containing FTP servers, as well as from that containing DNS servers, and so on.

We apply a K-means clustering model [6] to the 21 datasets referring to the monitored network metrics of each considered server. We set $K = 7$ because we have seven classes of servers. We consider clustering in a 2-dimensional space based on the correlation among input/output bytes and input/output packets of the monitored traffic flows, supported by previous studies on correlation of flow characteristics (e.g., [33]). By considering these two pairs of metrics, for each dataset related to one of the 21 servers we compute three similarity matrices containing the pair-wise similarity measures computed through CoHiVa, Pearson, and Loco. These measures allow us to place each server in the spaces generated through the indexes. In each space, the K-means model groups servers so as to minimize the within-cluster distance [6].

Fig. 10 shows the results related to the CoHiVa index (Fig. 10(a)), the Pearson product moment (Fig. 10(b)), and the LoCo score (Fig. 10(c)).

For this dataset, CoHiVa is able to group all the Web servers in a cluster, that is separated from the cluster containing the two DNS servers, from that referring to the two FTP servers, from the one for all the Samba servers, and so on. We should note that the chosen metrics and similarity measure do not allow us to discriminate between the Oracle and MySql database servers, that consequently are grouped together.

The results related to Pearson and LoCo are much poorer: they cluster different types of servers in the same cluster, and some servers are in singleton clusters despite their expected correlation. These results are caused by the high variability of some monitored metrics and to the low mean value of others. As a consequence, Pearson and LoCo correlation indexes see as uncorrelated time series that actually present some dependency, while they see high correlations between time series that are actually independent.

The reported results show that CoHiVa represents an effective solution for clustering data in highly variable contexts where state-of-the-art similarity measures are affected by poor results.

6. Conclusion

We propose a novel similarity measure that can be applied to correlation-based clustering algorithms specifically tailored to the analysis of time series characterized by high variability. This paper is motivated by the observation that existing clustering algorithms are affected by poor results when data are highly variable as in most datasets obtained by network and system measurements. Experimental evaluations carried on synthetic and real datasets demonstrate that our solution improves the state of the art in clustering traffic flows and server behaviors. These promising results open the possibility of using the proposed model as a support for several applications including traffic and network management, and capacity planning of networks and systems.

References

[1] Z. Liu, M. Squillante, C. Xia, S. Yu, L. Zhang, Profile-based traffic characterization of commercial web sites, in: Proc. of the 18th International Teletraffic Congress, Berlin, Germany, 2003.

[2] J. Erman, M. Arlitt, A. Mahanti, Traffic classification using clustering algorithms, in: Proc. of the 2006 SIGCOMM Workshop on Mining network data, New York, USA, 2006.

[3] H. Wang, W. Wang, J. Yang, P.S. Yu, Clustering by pattern similarity in large data sets, in: Proc. of the 2002 ACM SIGMOD International Conference on Management of Data, Madison, Wisconsin, 2002.

[4] B. Hay, G. Wets, K. Vanhoof, Clustering navigation patterns on a website using a sequence alignment method, in: Proc. of 17th International Joint Conference on Artificial Intelligence, Washington, USA, 2001.

[5] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, in: KDD Workshop on Text Mining, 2000.

[6] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.

[7] J. Cohen, P. Cohen, S.G. West, L.S. Aiken, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, L. Erlbaum Associates, 2003.

[8] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 100 (3–4) (1904) 441–471.

[9] M.G. Kendall, Rank Correlation Methods, Charles Griffin & Company Ltd., 1962.

[10] S. Papadimitriou, J. Sun, P.S. Yu, Local correlation tracking in time series, in: IEEE International Conference on Data Mining, Los Alamitos, USA, 2006.

[11] M. Andreolini, S. Casolari, M. Colajanni, Models and framework for supporting run-time decisions in web-based systems, ACM Transaction on the Web 2 (3) (2008) 17:1–17:43.

[12] S.G. Mallat, A theory of multiresolution signal decomposition: the wavelet decomposition, IEEE Transaction on Pattern Analysis and Machine Intelligence 11 (7) (1989) 674–693.

[13] P. Barford, M. Crovella, Generating representative Web workloads for network and server performance evaluation, SIGMETRICS Performance Evaluation Review 26 (1) (1998) 151–160.

[14] M. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, IEEE/ACM Transactions on Networking 5 (6) (1997) 835–846.

[15] M.N. Bennani, D.A. Menasce, Assessing the robustness of self-managing computer systems under highly variable workloads, in: Proc. of the First International Conference on Autonomic Computing, Washington, USA, 2004.

[16] W. Willinger, D. Alderson, L. Li, A pragmatic approach to dealing with high-variability in network measurements, in: Proc. of the 4th ACM SIGCOMM Conference on Internet Measurement, Taormina, Italy, 2004.

[17] A. Buda, A. Jarynowski, Life-time of correlations and its applications, Wydawnictwo Niezalezne (2010).

[18] D.C. Montgomery, Introduction to Statistical Quality Control, John Wiley and Sons, 2008.

[19] T. Sørensen, A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content, Biologiske Skrifter, E. Munksgaard, 1948.

[20] S. Papadimitriou, P.S. Yu, Optimal multi-scale patterns in time series streams, in: Proc. of the 2006 ACM SIGMOD International Conference on Management of Data, Chicago, USA, 2006.

[21] S. Papadimitriou, S. Jimeng, C. Faloutsos, Streaming pattern discovery in multiple time-series, in: Proc. of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, 2005.

[22] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Inc., 1988.

[23] R. Khattree, D.N. Naik, Multivariate Data Reduction and Discrimination with SAS Software, SAS Institute Inc., 2000.

[24] B.R. Bakshi, Multiscale PCA with application to multivariate statistical process monitoring, AIChE Journal 44 (7) (1998) 1596–1610.

[25] B. Abrahao, A. Zhang, Characterizing Application Workloads on CPU Utilization in Utility Computing, Technical Report HPL-2004-157, Hewlett-Packard Labs, 2004.

[26] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk, N. Taft, Structural analysis of network traffic flows, in: Proc. of the Joint

International Conference on Measurement and Modeling of Computer Systems, New York, USA, 2004.

[27] H.E. Hurst, Long-term storage capacity of reservoirs, Transaction of the American Society of Civil Engineers 116 (1951) 770–799.

[28] R. Weron, Estimating long range dependence: finite sample properties and confidence intervals, Physica A 312 (1–2) (2002) 285–299.

[29] M. Dobber, R. Van det Mei, G. Koole, A prediction method for job runtimes in shared processors: survey, statistical analysis and new avenues, Performance Evaluation 64 (7–8) (2007) 755–781.

[30] B.L. Brockwell, R.A. Davis, Time Series: Theory and Methods, Springer-Verlag, 1987.

[31] Abilene network. <http://abilene.internet2.edu/>.

[32] D.L. Olson, D. Delen, Advanced Data Mining Techniques, Springer, 2008.

[33] K. Lan, J. Heidemann, On the Correlation of Internet Flow Characteristics, Technical Report ISI-TR-574, USC/Information Sciences Institute, 2003.

Stefania Tosi is a Ph.D. student in Information and Communication Technologies at the University of Modena and Reggio Emilia, Italy. She received her master degree (summa cum laude) in Computer Science from the University of Modena and Reggio Emilia in July 2010. Her research interests include performance evaluation of modern data centers and statistical models for data management. She has three publications in international journals and several proceedings of key international conferences. She received a best paper award at WWW/Internet 2010. Home page: http://weblab.ing.unimo.it/people/stefy.

Sara Casolari is a researcher assistant at the Department of Information Engineering of the University of Modena and Reggio Emilia, Italy. She received her master degree (summa cum laude) and the Ph.D. in Computer Engineering form the University of Modena and Reggio Emilia in information engineering in 2004 and 2008, respectively.Her research interests include stochastic models and performance evaluation of distributed systems, and modelling algorithms for supporting large systems and Internet-based application. She received a best paper award at the International Conference on Autonomic and Autonomous Systems (ICAS 2007) and at WWW/Internet 2010. Home page: http://weblab.ing.unimo.it/people/sara.

Michele Colajanni is full professor in computer engineering at the University of Modena and Reggio Emilia since 2000. He received the master degree in computer science from the University of Pisa, and the Ph.D. degree in computer engineering from the University of Roma in 1992. He manages the Interdepartment Research Center on Security and Safety (CRIS), and he is the Director of the postgraduate master course in "Information Security: Technology and Law". He is the author or co-authors of more than 150 papers on performance and prediction models, information security, management of large scale systems. Home page: http://weblab.ing.unimo.it/people/colajanni.