A Framework to Monitor Clusters' Evolution Applied to Economy and Finance Problems

Márcia Oliveira*and João Gama

LIAAD, FEP, University of Porto,
Rua de Ceuta 118, 4050-190 Porto, Portugal
marcia@liaad.up.pt
jgama@liaad.up.pt
http://www.liaad.up.pt

Abstract

The study of evolution has become an important research issue, especially in the last decade, due to our ability to collect and store high detailed and time-stamped data. The need for describing and understanding the behavior of a given phenomena over time led to the emergence of new frameworks and methods focused in temporal evolution of data and models. In this paper we address the problem of monitoring the evolution of clusters over time and propose the MEC framework. MEC traces evolution through the detection and categorization of clusters' transitions, such as *births*, *deaths* and *merges*, and enables its visualization through bipartite graphs. It includes a taxonomy of transitions, a tracking method based in the computation of conditional probabilities, and a transition detection algorithm. We use MEC with two main goals: to determine the general evolution trends and to detect abnormal behavior or rare events. To demonstrate the applicability of our framework we present real world economic and financial case studies, using datasets extracted from Banco de Portugal's Central Balance-Sheet Database and the *The Data Page* of New York University - Leonard N. Stern School of Business. The results allow us to draw interesting conclusions about the evolution of activity sectors and European companies.

Keywords: Bipartite Graphs, Change Mining, Clustering Evolution, Monitoring, Transitions

*Corresponding author

1 Introduction

The celerity at which evolution takes place, typically characterized by breaks and shifts, has increased exponentially in last decades. The rapid progress made in science and technology has contributed to the emergence of a volatile and fast pace evolving world, which demands new perspectives in knowledge discovery upon data, such as time-oriented perspectives. The paradigm of *change mining* arises as a consequence of this evolution and encompasses data mining mechanisms that monitor models and patterns over time, compare them, detect and describe changes, and quantify them on their interestingness (Bottcher et al., 2008). Therefore, the challenge of *change mining* lays not only in the adaptation of models to changes in data distribution but also in the understanding of changes themselves. Bearing this in mind some researchers and practitioners of data mining developed methods and techniques to study evolution that provide insights about the behavior of data, or models, over time (see, for example Spiliopoulou et al. (2006), Baron and Spiliopoulou (2004), Aggarwal (2005), Asur et al. (2009)).

Clustering is one of the most popular and useful tasks of data mining, and is broadly used in business fields. However, clustering has been mainly applied to static data, which results in clusters' structures that can only describe well a specific time moment. Nevertheless, companies make medium and long-term decisions based on these results, which are highly susceptible to change. Therefore, monitoring the dynamics of clusters' structures is very important in many real world applications, since it fosters the creation of sustainable knowledge about the evolution of phenomena and, consequently, the adoption of pro-active attitudes. Besides, it may correlate to some important or critical events in the real applications or unveil the emergence of new ones. For these reasons, this study can benefit several areas, such as Marketing, Fraud Detection, Economy and Finance. For instance, the study of the evolution of customers' segments allows the detection of shifts in preferences and consumer's habits, which can help the forecast of trends and consequent redefinition of Marketing's strategies and policies. This can also be useful to improve customer relationship management (CRM). The domain knowledge acquired by these means can act as a powerful differentiating factor in the market and strongly contributes to the creation, or reinforcement, of the company's competitive advantages. Regarding Economics, it may be of interest to study the evolution of economic activity sectors of a given country in order to discover emerging business areas, with high growth potential, or activity sectors that are losing strength. Traditional data mining is not able to help companies or practitioners achieve these goals since it relies on static data and does not take into account its evolving nature. Therefore, the study of the dynamics of clusters contributes to the achievement of a greater understanding of cluster's evolutionary processes, broadening horizons and opening new paths in the way we think about problems.

To help decision makers and practitioners of several areas make better and more reasoned decisions, in this paper we present a generic framework to study the dynamics of clusters, which can provide *solid* knowledge about the studied problem. The `MEC` framework was designed to monitor clusters' transitions over time (our basic idea is depicted in Figure 1), through the identification of temporal relationships among them. `MEC` encompasses a taxonomy of various types of clusters' transitions, a tracking method based in the computation of conditional probabilities and a transition detection algorithm. It also incorporates a visualization technique that explores the concept of bipartite graphs. The input of `MEC` is previously discovered clusters, so the focus of this work is on the evolution process and not in the clustering process. Therefore, our monitoring system is not restricted to numerical attributes or features and can be employed to clusters' structures returned by any clustering algorithm.

Contrary to earlier research, our proposal takes into account the dynamic nature of real world phenomena, and is focused on the modeling, characterization and further understanding of changes in groups of entities (e.g. activity sectors, companies, costumers and countries). The relationship between changes can then be explored to give rise to two different types of analyses: one focused in the determination of the general evolution trends of clusters and another based in the detection of abnormal or rare events in data. Both analyses can be very useful and interesting in the context of Economics and Management problems. To illustrate it we conduct two real world case studies using datasets from Banco de Portugal's Central Balance-Sheet Database (from now on, we will refer to these datasets as *Banco de Portugal's datasets*) and the *The Data Page* of New York University - Leonard N. Stern School of Business.

In this paper we focus on the Business and Economic domain. Yet, the proposed approach may prove to be of importance to several other domains, such as Biology (e.g. to study the evolution of the species), Sociology (e.g. to help understand the evolutionary trends of social groups), Climate (e.g. for studying

changes in spatial clusters of climate stations) and Text Mining (e.g. to analyze the evolution of trends in news topics), just to name a few.

The rest of the paper is organized as follows: Section 2 provides a brief overview of the current state of the art. In Section 3 we formally introduce our approach to monitor clusters' transitions. In this section we present the adopted scheme to define clusters, introduce our taxonomy for clusters' transitions and explain the foundations of our tracking method. In Section 4 we show and discuss two real world case studies using economic and financial datasets. Section 5 concludes the paper with a summary and pointing out directions for future work.

2 Related Work

Despite the extensive study of the clustering problem (surveys can be found in Jain et al. (1999), Berkhin (2006), Jain (2010)), there is not much work conducted in the monitoring of clusters' transitions. In context of evolution, the research endeavor has been mainly directed to the adaptation of clusters to changed populations. However, the dynamic nature of most problems encouraged new directions in research. This effort is clearly present in the areas of machine learning and data analysis.

Currently, there are several algorithms that directly or indirectly aim to capture and understand the dynamic nature of these datasets, particularly susceptible to the occurrence of changes in the underlying structure. The last decade has been especially profuse in the design of transition detection algorithms. Based on recent literature, it was possible to deduce a preliminary classification for algorithms built in this context: there are algorithms more appropriate for static data (snapshots of data) and algorithms designed to operate in a data streams environment. The first may be focused on transitions experienced by generic patterns (Ganti et al., 1999, Bartolini et al., 2009, Chawathe and Garcia-Molina, 1997), clusters (Spiliopoulou et al., 2006, Falkowski et al., 2006, Yang et al., 2005) or association rules (Baron and Spiliopoulou, 2001, 2004). Within the scope of static data, relevant work for generic patterns include the FOCUS framework and the algorithm MH-DIFF. FOCUS framework (Ganti et al., 1999) describes a general model (e.g. decision trees, frequent item sets and clusters) by a structural component and a measure component, and computes the deviation between two different datasets through the comparison of these two components. The significance of the difference, or deviation, is assessed using standard statistical techniques. Chawathe and Garcia-Molina (1997) designed an efficient algorithm, `MH-DIFF`, to detect meaningful changes in hierarchically structured data, such as nested-object data (e.g. trees). The problem of identifying change between two snapshots of data is regarded by the authors as the problem of finding the best way to edit a tree created in t_i, to obtain another tree, created in t_{i+1}. Basically, they compare the distance between two models using a set of editing operations.

An interesting work regarding association rules was proposed by Baron and Spiliopoulou (2004). These authors devised PAM (Automated Pattern Monitor) to efficiently monitor patterns and detect

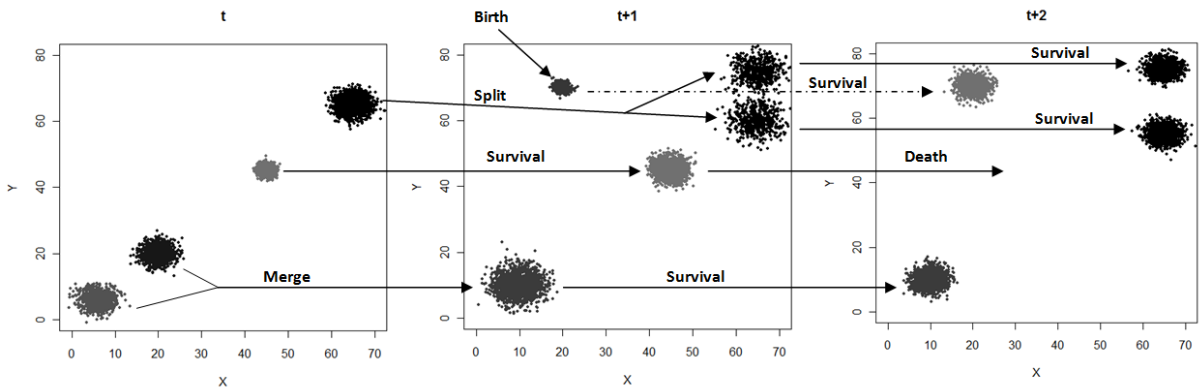


Figure 1: Our proposal to study clusters' evolution

meaningful changes in the evolution of association rules. The framework is subdivided into two phases: in the first one, association rules are discovered applying a data mining algorithm to raw data; then these rules are imported to the monitor and stored according to the GRM (Generic Rule Model), which models both the content and the statistics of a rule as a temporal object. In the second stage, the statistics of the more representative rules are monitored over time, and PAM detects a change when these statistics are above the threshold. The detection of this kind of changes implies the subsequent reapplying of the data mining algorithm to understand the underlying causes.

Regarding clusters, the majority of research is related to the study of social networks (Falkowski et al., 2006, Asur et al., 2009). Falkowski et al. (2006) address the problem of monitoring the dynamics of communities. They consider two types of communities, on the level of subgroups: stable communities and high fluctuating communities; and propose two different methods to tackle each one of the problems. These methods are based in visualizations of the network and their subgroups, and in the analysis of different statistics to detect transitions such as *growing*, *declining*, *merging* and *splitting* of communities. Basically, the main goal of this framework is the detection of changes and shifts in the evolution of patterns, in the context of social networks. A more detailed framework can be found in Asur et al. (2009). In this work, an event-based framework for characterizing the evolution of interaction networks is proposed. The main goal is to model evolution and understand the dynamic behavior of these graphs, which can be characterized through the identification of critical events, such as *formation*, *dissolution*, *split* and *merge* of clusters (communities), or *join*, *leave*, *appearance* or *disappearance* of individuals (entities). Although the evolution is monitored through the comparison of static snapshots of data, obtained in consecutive time periods, the devised event detection algorithm is incremental and can be easily extended to a data streams environment.

Concerning data streams, approaches focusing unclassified data are quite common (here we include clustering algorithms for mobile objects or spatio-temporal objects). These approaches (Aggarwal, 2005, Chen and Liu, 2006, Aggarwal, 2003, O'Callaghan et al., 2002, Elnekave et al., 2007, Kalnis et al., 2005, Li et al., 2004) elect clusters as its main data structure and are concerned with the efficiency and scalability of algorithms. Important research includes the work of Aggarwal (2005), Chen and Liu (2006) and Elnekave et al. (2007). Aggarwal (2005) studied the problem of data evolution and developed a generic tool to understand, visualize and diagnose different types of changes in evolving data streams. The proposed techniques explore the concept of velocity density estimation in order to create both *temporal* and *spatial* velocity profiles of data at different time points. These profiles provide different perspectives on the nature of the underlying change: while the *temporal velocity profile* gives an overview of the density rate of change, over a given period of time, and in a fixed spatial location; the *spatial velocity profile* gives information about the spatial movement of data in the feature space, in a fixed time point. The analysis of these profiles enables the detection of changes and further categorization according to the following taxonomy: data *coagulation*, data *dissolution* and *shift* of data. This framework can also be extended to high-dimensional data streams.

Chen and Liu (2006) proposed an online entropy-based method to automatically validate the clustering and detect change in categorical data streams. In this context, change must be interpreted as the modification of the optimal number of clusters, since this can be an indicator of changes in the underlying structure of data. The proposed framework extends the work done in determining the best k for static datasets - Method BkPlot - to categorical data streams, with the help of a summarization structure called HE-Tree (Hierarchical Entropy Tree).

Elnekave et al. (2007) addressed the problem of clustering the trajectories of moving objects and developed an incremental clustering algorithm to find groups of similar objects that evolve over space and time.

Although these researches study the evolution of clusters, they use quite distinct approaches to tackle the problem. Based on our knowledge, presently there are only two research works that address a problem closely related to ours (Spiliopoulou et al., 2006, 2009). MONIC framework (Spiliopoulou et al., 2006) also proposes a cluster transition model to track cluster changes, supporting cluster comparisons across the time axis. MONIC uses a data aging function that assigns lower weights to older observations and computes the overlap of clusters to capture its evolution. An extension to this framework, called MONIC+, was later proposed by the same authors (Spiliopoulou et al., 2009). MONIC+ is less generic since it defines specific transitions indicators for arbitrary types of clusters (e.g. clusters defined in a metric space, clusters extensionally defined and clusters intentionally defined) to detect changes in clustering.

The introduction of proper heuristics to capture transitions for different kinds of clusters broadens the application of the framework and turns it independent of the clusters discovery process. However, if we compare both systems (MONIC and MONIC+) to our framework, we conclude that they use different metrics to detect change and do not provide techniques to visualize the monitoring process. Our earlier work (Oliveira and Gama, 2010) also proposes a framework to model the evolution of clusters through weighted bipartite graphs, and is built upon related work systems, namely, MONIC and MONIC+. The work presented in this paper is a substantial extension of our previous research (Oliveira and Gama, 2010).

Somewhat related is the work on novelty detection that can identify new concepts (in our terminology, the birth of new clusters), from unlabeled data (see for example Spinosa et al. (2007)). These methods are restricted to cluster's births, contrary to our framework, which is able to capture a much wider variety of concepts, such as splits, merges, deaths and survivals.

There are also classic statistical methods, such as Panel Data Analysis (PDA) (Urga, 1992), focusing on the study of dynamics. Panel data can be seen as a set of individuals (or objects) which features are repeatedly collected at two or more points in time. The analysis of this data aims to model the heterogeneity, or differences, between individuals, in order to capture its dynamics. This is usually done using regression that is applied over the two dimensions of data (cross-section and time-series). Despite the fact that PDA and our framework share the same goal, they use quite different ways to achieve it. Moreover, PDA is focused on raw data instead of clusters' structures, and it uses regression which does not provide any information about the nature of data change. The interpretability of the results of PDA also differs from `MEC`, since the evolution is not as easy to understand using regression models.

3 `MEC` Framework

We developed the `MEC` framework in order to monitor the evolution of clusters' structures (also referred to as *clustering* - Definition 1) obtained at different static snapshots. In this context, the concept of evolution refers to transitions undergone by clusters during the time interval under observation $[t_i, t_{i+\Delta t}]$ ($\Delta t = 1, ..., T - i$). The main goal is to identify hidden behavioral patterns and create useful and actionable knowledge based upon these findings.

In the next subsections we present the adopted scheme to represent clusters, our taxonomy of transitions and the devised method to monitor clusters' evolution.

Definition 1 - Clustering:
A Clustering ξ *is a specific partitioning of a dataset* D *into* k *partitions, usually denoted as* clusters $\xi = \{C_1, ..., C_i, ..., C_k\}$, *such that:*

1. $C_i \cap C_j = \emptyset$, $\forall_{i \neq j}$ - clusters *are disjoint sets (or mutually exclusive);*
2. $\cup_{i=1}^{K} C_i = D$ - clusters *are collectively exhaustive;*
3. *Observations assigned to a given* cluster *are more similar to each other than to observations assigned to other* clusters *belonging to* Clustering ξ.

3.1 Cluster Characterization

The `MEC` framework assumes that clusters are represented by enumeration (also known as *extensional definition of clusters*), which is the most used and straightforward way to define clusters. In this kind of representation a cluster is defined by its members, i.e., by the observations that were assigned to it by a given clustering algorithm (see Definition 2).

Definition 2 - Representation by Enumeration:
Let $\vec{x_i}$, $(i = 1, ..., N)$ *be the ith observation defined as a vector of real numbers in a d-dimensional space* $\vec{x_i} = (x_{i,1}, x_{i,2}, x_{i,3}, ..., x_{i,d})$. *A possible temporal representation of a cluster is defined as follows:*

$$C_j(t) = \{\vec{x_1}, ..., \vec{x_m}\}$$

where m is the number of observations assigned to cluster $C_j(t)$, $j = (1, ..., k)$, k is the number of clusters and $t = (1, ..., T)$, where T corresponds to the last analyzed time stamp.

This type of representation does not involve information loss and enables the monitoring of each object over time. Though, it is not always possible to define clusters in this way, e.g. due to storage demands or privacy issues.

3.2 Taxonomy of Transitions

There are at least eight taxonomic schemes for the classification of transitions in clusters, patterns or graphs that evolve over time, in the literature (Spiliopoulou et al., 2006, Falkowski et al., 2006, Yang et al., 2005, Aggarwal, 2005, Chen and Liu, 2006, Aggarwal, 2003, Li et al., 2004, Kaur et al., 2009, Asur et al., 2009). To capture the changes likely to occur in clusters' structures we considered the following taxonomy:

- **Birth** - a new cluster emerges
- **Death** - a previous discovered cluster disappears
- **Split** - one cluster is separated into two or more clusters
- **Merge** - two or more clusters fuse, or merge, into one cluster
- **Survival** - a cluster that does not suffer any of the above transitions

These transitions are external, as they relate to changes in the whole *Clustering*, and represent five basic types of changes that clusters may undergo between two consecutive time points. The adopted terminology is the one introduced by Spiliopoulou et al. (2006) and can be found in Table 1. The formal definition of these transitions are summarized in Table 2.

Table 1: Terminology for clusters' transitions

Mathematical Notation	**Description**
$\emptyset \rightarrow C_u(t_j)$	Cluster's Birth
$C_m(t_i) \rightarrow \emptyset$	Cluster's Death
$C_m(t_i) \overset{\subseteq}{\rightarrow} \{C_1(t_j), ..., C_r(t_j)\}$	Split of a cluster into r clusters
$\{C_1(t_i), ..., C_p(t_i)\} \overset{\subseteq}{\rightarrow} C_u(t_j)$	Merge of p clusters into one cluster
$C_m(t_i) \rightarrow C_u(t_j)$	Cluster's Survival

The key concept for the detection and evaluation of these transitions is the concept of *mapping*[1], which can be defined as the process of discovering the matches between clusters obtained at time point t_i and clusters obtained at a later time point $t_{i+\Delta t}$, in case they still exist.

3.3 Tracking Method

Here we expose the method we devised to track the evolution of clusters and which allows the detection of the aforementioned transitions. To be able to detect clusters' transitions, first is necessary to find out which clusters of time point t correspond to clusters previously found in $t_{i-\Delta t}$. That is, it is essential to perform the *mapping* of clusters in a given time interval, to discover overlapping regions in the feature space or, in other words, the matches of clusters. In our method, the *mapping* process explores the concept of conditional probability and is restricted by a predefined threshold - **survival threshold** τ -, which assumes the minimum of $\tau = 0.5$. These conditional probabilities are computed for every pair of possible connections between clusters obtained at consecutive time points and they represent the edges' weights in a bipartite graph (see Figure 2). We resort to conditional probabilities to compute these weights since this is a sound and well known mathematical concept which can provide, in this specific

[1] *Mapping* is the process of discovering the matches of clusters between two distinct time points

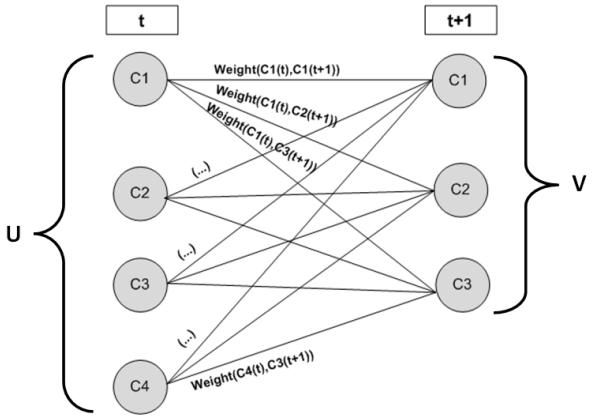


Figure 2: Illustration of a bipartite graph: nodes represent clusters and edges' weights are conditional probabilities

context, important information about the degree of permanency of objects in a given cluster, and can be used to evaluate the similarity of two specific clusters. On the other hand, we use bipartite graphs with two main purposes: it helps the formal definition of transitions and it works as a visualization tool that enables users to gain greater understanding of the monitoring process. Moreover, bipartite graphs are well suited to the modeling of matching problems, which is one of its main applications. The foundations of our transition detection algorithm are based on this idea, which can be defined as follows (Definition 3).

Definition 3 - Weighted Bipartite Graphs:
Given the clusterings ξ_i, $\xi_{i+\Delta t}$, obtained at t_i, $t_{i+\Delta t}$, a graph $G = (U, V, E)$ can be constructed, where U represents the first subset of vertices (clusters of t_i), V represents the second subset of vertices (clusters of $t_{i+\Delta t}$), and E denotes a set of weighted edges between any pair of clusters belonging to ξ_i and $\xi_{i+\Delta t}$. Formally, the weight assigned to the edge connecting clusters $C_m(t_i)$ and $C_u(t_{i+\Delta t})$ ($m = (1, ..., k_{t_i})$ and $u = (1, ..., k_{t_{i+\Delta t}})$, where k_{t_i} and $k_{t_{i+\Delta t}}$ are the number of clusters returned by a given clustering algorithm in time points t_i and $t_{i+\Delta t}$, respectively) are estimated in accordance with the conditional probability (Equation 1):

$$weight(C_m(t_i), C_u(t_{i+\Delta t})) = P(X \in C_u(t_{i+\Delta t}) | X \in C_m(t_i)) = \frac{\sum P(x \in C_m(t_i) \cap C_u(t_{i+\Delta t}))}{\sum P(x \in C_m(t_i))} \quad (1)$$

where X is the set of observations assigned to cluster $C_m(t_i)$ and $P(X \in C_u(t_{i+\Delta t}) | X \in C_m(t_i))$ represents the probability of X belonging to cluster C_u from $t_{i+\Delta t}$ knowing that X belongs to cluster C_m obtained at a previous time stamp t_i.

Although in this paper we only explore crisp clustering, the previous formula can be extended in order to apply `MEC` to fuzzy clustering. In fuzzy clustering each object can belong to multiple clusters, presenting different membership degrees to each cluster. This requires some modifications in the metric that can be introduced using fuzzy logic. In such cases, the intersection/union of clusters A and B must be computed using fuzzy set operations like: $A \cap B(x) = \min[A(x), B(x)]$ for the standard intersection operation, and $A \cup B(x) = \max[A(x), B(x)]$ for the standard union operation.

It is worth mentioning that the use of conditional probabilities requires structurally identical datasets, i.e., datasets composed by observations of the same objects, in each time point under analysis. This aspect can be seen as a drawback but also as an advantage since it allows the tracking of each object's trajectory over time. In real world applications this can be of major interest, especially in the Economics and Management field (for instance, it allows the definition of a company's life cycle).

To detect changes, we formally define the transitions that a cluster $C \in \xi_i$ can experience, with respect to $\xi_{i+\Delta t}$. It was introduced a new threshold to help the definition of these transitions: the **split**

threshold λ. This formal design is based on MONIC's framework's external transitions (Spiliopoulou et al., 2006) and is depicted in Table 2.

Table 2: Formal definition of the external transitions of a cluster represented by enumeration

Transitions' Taxonomy	Notation	Formal Definition
Cluster's Birth	$\emptyset \rightarrow C_u(t_{i+\Delta t})$	$0 < weight(C_m(t_i), C_u(t_{i+\Delta t})) < \tau \forall_m$
Cluster's Death	$C_m(t_i) \rightarrow \emptyset$	$weight(C_m(t_i), C_u(t_{i+\Delta t})) < \lambda \forall_u$
Cluster's Split	$C_m(t_i) \overset{\subseteq}{\rightarrow} \{C_1(t_{i+\Delta t}), ..., C_r(t_{i+\Delta t})\}$	$(\exists_u \exists_v : weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \lambda \wedge$ $weight(C_m(t_i), C_v(t_{i+\Delta t})) \geq \lambda) \wedge$ $\sum_{u=1}^{r} weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \tau$
Cluster's Merge	$\{C_1(t_i), ..., C_p(t_i)\} \overset{\subseteq}{\rightarrow} C_u(t_{i+\Delta t})$	$(weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \tau) \wedge$ $\exists C_p \in \xi_i \setminus \{C_m\} : weight(C_p(t_i), C_u(t_{i+\Delta t})) \geq \tau$
Cluster's Survival	$C_m(t_i) \rightarrow C_u(t_{i+\Delta t})$	$(weight(C_m(t_i), C_u(t_{i+\Delta t})) \geq \tau) \wedge$ $\nexists C_p \in \xi_i \setminus \{C_m\} : weight(C_p(t_i), C_u(t_{i+\Delta t})) \geq \tau$

Table 2 presents the formal definition for each one of the transitions. A cluster is **born** when the weights of the edges that connect this cluster to all previous clusters are below the survival threshold τ or, in other words, when is not possible to find, in the clustering of the previous time point, a single cluster that can be considered an approximate match. Similarly, a cluster **dies** when all links that connect this cluster to clusters of the following time point are below the split threshold. Exceptions are made when there is, at least, one edge above the threshold, but the situation cannot be considered as a clusters' split. Regarding clusters' **split**, it is assumed that a cluster of t_i is divided into, at least, two clusters in $t_{i+\Delta t}$ if exists, at least, two clusters $C_u(t_{i+\Delta t})$ and $C_v(t_{i+\Delta t})$ whose edges' weights are equal, or exceed, the split threshold λ and its sum is equal, or exceed, the survival threshold. On the other hand, a clusters' **merge** occurs when there, are least, two different clusters from t_i, whose connections to a given cluster of $t_{i+\Delta t}$ are equal or above the survival threshold. Finally, a cluster **survives** when there is one and only one link between a pair of clusters $C_m(t_i)$ and $C_u(t_{i+\Delta t})$ which weight is equal or above the survival threshold.

Our transition detection algorithm covers the introduced elements of the framework and it was implemented in R 2.10.0. This implementation was supported by the definitions presented in Table 2.

4 Real World Case Studies

In order to show the potential of the application of our framework to economic problems and glean insights about the evolution of clusters, we conducted two real world case studies, using datasets extracted from Banco de Portugal's Central Balance-Sheet Database and *The Data Page* of New York University - Leonard N. Stern School of Business [2]. In the first case study, our main aim is to capture and understand the general evolution trends of the Portuguese activity sectors. In the second case study we focus on the detection of rare events in European companies. Since the experiments require predefined thresholds, we begin this section with a threshold's sensitivity analysis, to help us choose its values.

4.1 Threshold's Sensitivity Analysis

In order to help the decision of choosing reasonable values for the thresholds we studied the impact of small variations in the thresholds values - τ and λ - in the output of the transition detection algorithm, to see if transitions results are stable. We only present the results for the datasets of Banco de Portugal (using the partition returned by the agglomerative hierarchical algorithm, using Ward's method), since the obtained conclusions are very similar to the ones returned using the financial dataset. We made experiences for all values of survival threshold within the range $[0.5, 1]$ [3], and for the values of the split threshold falling into $[0, 0.4]$. The analysis was performed separately by varying the values of one of the thresholds and keeping everything else constant (*ceteris paribus*), e.g. when observing the behavior of τ,

[2] `http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html`

[3] It is assumed a minimum of 0.5 which, in other words, means that the probability of the objects of a given cluster belong to the cluster of the next time point has to be, at least, half of the maximum

we assume a constant value of $\lambda = 0$, and when analyzing λ, is considered a fixed value of survival of $\tau = 0.5$.

4.1.1 Survival Threshold

Table 3 shows the relationship between values of τ and the number of occurrences of each transition. Logically, one would expect that the higher the value of the survival threshold τ, the smaller the number of survivals and merges detected. It would also be expected that the increase of τ leads to a greater number of births, deaths and splits. Table 3 is consistent with the expected results, since the number of merges decreases and the number of splits increases to more demanding values of τ. However, the number of survivals increases when τ changes from 0.6 to 0.7, which seems a contradiction. In fact, the explanation is very simple: due to the change of the threshold value a merge was replaced by one survival, since the weight assigned to one of the edges became lower than the threshold. Concerning the other transitions, it is easy to understand why it was not detected any birth, or death, of clusters, and also why the number of births and deaths remain constant for different threshold values. The reason behind this situation is related to the low number of detected clusters and consequent absence of very low edges' weights.

Table 3: Impact on the number of transitions varying the survival threshold τ, for time interval $[2005, 2006]$

	Values of τ					
Transitions	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Survival	1	1	2	2	2	2
Death	0	0	0	0	0	0
Birth	0	0	0	0	0	0
Merge	1	1	0	0	0	0
Split	0	0	1	1	1	1

In the following time interval $[2006, 2007]$ - Table 4 -, the initial hypothesis is confirmed for the number of survivals, that decreases, and for the number of splits, that increases. The remaining transitions do not suffer modifications.

Table 4: Impact on the number of transitions varying the survival threshold τ, for time interval $[2006, 2007]$

	Values of τ					
Transitions	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Survival	2	2	0	0	0	0
Death	0	0	0	0	0	0
Birth	0	0	0	0	0	0
Merge	0	0	0	0	0	0
Split	0	0	2	2	2	2

The survival threshold analysis, using the datasets from Banco de Portugal, led to the following main conclusions: the more challenging the survival threshold, the lower the number of survivals and merges and the higher the number of splits. Births and deaths remain constant. Using the European companies' datasets, we also verify that the number of deaths increases for higher values of the threshold.

4.1.2 Split Threshold

Regarding the split threshold λ, it is expected that the higher its value, the greater the number of deaths and births, and lower the number of splits. One can also deduce that the number of merges and survivals is not affected by the variation of λ, based on the formal definition of these transitions. The analysis of the threshold's variation for both time intervals, proved only one of the hypothesis, ie, that λ has not

influence in the number of survivals and merges, which remain unchanged for every values of λ. However, the same analysis for the European companies' datasets confirmed the second hypothesis: the higher the value of λ the higher the number of deaths and the lower the number of splits.

4.1.3 Discussion

The threshold's sensitivity analysis led to some interesting conclusions. The most important is that the tuning of the split threshold values is not critical, since it has not a significant impact in the number of transitions, but the modification of the survival threshold values may lead to completely different results, once the number of transitions are more volatile and sensitive to small variations of τ.

Based on the information acquired by this analysis we chose the following values for the thresholds: $\lambda = 0.2$ and $\tau = 0.6$, which we assume to be reasonable values. We tried to be less demanding to capture a wider variety of transitions.

4.2 First Case Study - Portuguese Activity Sectors

The aim of this case study is to investigate the existence of relevant changes in the Portuguese economic structure over the years, e.g. growth/decline of activity sectors' clusters or appearance of clusters that represent emergent economic areas. We intend to detect the general trends in the economic sectors' context and try to understand what explains it.

To conduct this case study, we extracted three datasets from Banco de Portugal. Each dataset corresponds to a year (2005, 2006 and 2007) and each one consists of 12 observations characterized by 9 continuous attributes. The objects represent activity sectors, and the attributes are financial and economic aggregated indicators (number of companies, net income, investment rate, return on equity, net asset turnover, value added rates, debt ratio, equipment's productivity and labor productivity). These attributes were standardized, using Z-scores, so all attributes could have the same importance in the computation of the clustering results. The adopted scheme to classify the activity sectors was CAE (Portuguese Classification of the Economic Activities), which is a hierarchical nomenclature, with increasing degree of specificity. To simplify the interpretation of the results, we chose the most general class of the nomenclature - the *principal* activity sectors -, which corresponds to the top of CAE's hierarchy and is composed by 17 sectors, namely:

- A - Agriculture, hunting and forestry
- B - Fishery
- C - Mining and quarrying
- D - Manufacturing
- E - Production and distribution of electricity, gas and water
- F - Construction
- G - Wholesale and retail trade, repair of motor vehicles, motorcycles and personal goods
- H - Hotels and restaurants (restaurants and similar)
- I - Transport, storage and communications
- J - Financial Activities
- K - Real estate, renting and business services
- L - Public administration, defense and "mandatory" social security
- M - Education
- N - Health and social work

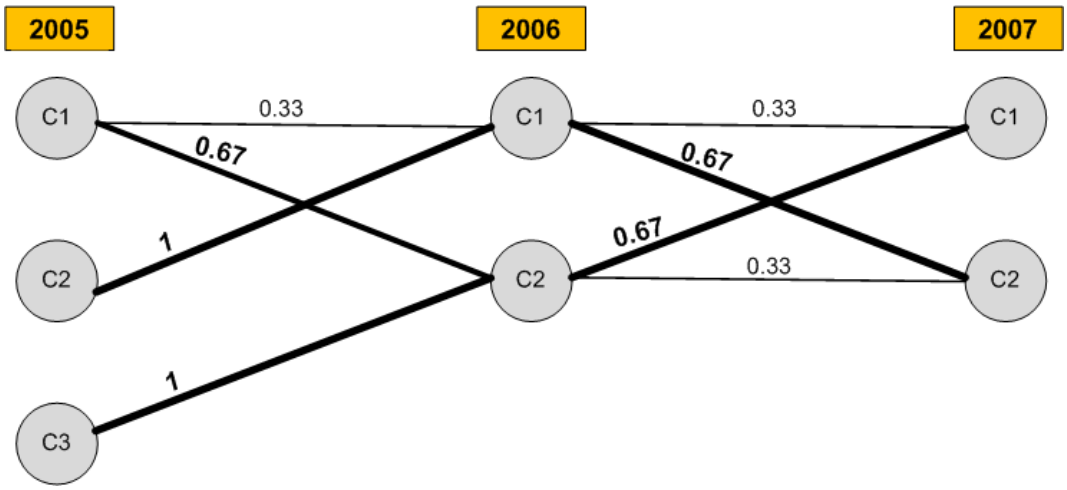


Figure 3: Bipartite graphs for Banco de Portugal's datasets, corresponding to timestamps 2005, 2006 and 2007. The thicker lines indicate weights that are equal or greater to survival threshold ($\tau \geq 0.6$).

- O - Other community, social and personal activities
- P - Activities of households with employed persons and production activities of households for own use
- Q - Organizations and extraterritorial institutions

However, we only have information available for 12 sectors. It should be noted that activity sectors and performance indicators are exactly the same, for all periods under analysis, which is a requirement for the application of the `MEC` framework. In order to discover the clusters (input of our study) from these datasets, we conducted experiments using different algorithms for clustering, but we only present results for the agglomerative hierarchical algorithm, using Ward's method. The determination of the critical clustering structure, through the identification of the best number of clusters - best k -, was supported by the analysis of an internal validation measure: the average silhouette width criterion (Rousseeuw, 1987), which is an useful and popular quality index for crisp clustering evaluation.

Some of the reasons behind the choice of the average silhouette width are the following: is one of the most referred validity measures in the literature (Bolshakova and Azuaje, 2003, Qiao and Edwards, 2009, Cardoso and Ponce de Leon Ferreira de Carvalho, 2009, Albatineh et al., 2006, Dudoit and Fridlyand, 2002, Hruschka et al., 2006); is a measure that can be more accurate than, for instance, the Davies-Bouldin index (Petrovic, 2006); and also because there are some guidelines about how to interpret its thresholds (as mentioned in page 88 of Kaufman (2005)), which is quite rare in this specific area of clustering validation.

Afterwards, we applied our transition detection algorithm to the returned clusterings, setting a fixed $\lambda = 0.2$ and $\tau = 0.6$, whose choice was guided by the previously presented thresholds' sensitivity analysis. These thresholds values are more relaxed and allow the detection of a wider variety of transitions. The resulting bipartite graphs (one bipartite graph for each time interval), showing the transitions between clusters, are depicted in Figure 3.

In Figure 3 we observe two bipartite graphs, one corresponding to transitions experienced during time interval $[2005, 2006]$ and another one corresponding to transitions occurred during $[2006, 2007]$. Each set of vertices represents the clustering returned by the algorithm for a specific time point/year and, therefore, we can deduce that, in 2005, there were three clusters ($k = 3$) and in 2006 and 2007 the best number of clusters was two ($k = 2$). The low k suggested by the analysis of the silhouette width is easily explained by the low number of observations. The composition of each cluster, in each year, is shown in Table 5, Table 6 and Table 7. The corresponding centroids, that are usually used to characterize clusters, can be consulted in Table 8, Table 9 and Table 10. These centroids are standardized so if they are negative it means that the value is below the average of the attribute, and if they are positive then the value is above the average of the attribute. Concerning the transitions, in the former there were:

Table 5: Principal Activity Sectors assigned to each cluster returned by the agglomerative hierarchical algorithm for year 2005

Cluster 1
Fishing
Mining and quarrying
Manufacturing
Wholesale and retail trade, repair of motor vehicles, motorcycles and personal goods and household
Hotels and restaurants (restaurants and similar)
Transport, storage and communications
Cluster 2
Production and distribution of electricity, gas and water
Cluster 3
Construction
Real estate, renting and business services
Education
Health and social work
Other community, social and personal activities

Table 6: Principal Activity Sectors assigned to each cluster returned by the agglomerative hierarchical algorithm for year 2006

Cluster 1
Fishing
Mining and quarrying
Production and distribution of electricity, gas and water
Cluster 2
Manufacturing
Construction
Wholesale and retail trade, repair of motor vehicles, motorcycles and personal goods
Hotels and restaurants (restaurants and similar)
Transport, storage and communications
Real estate, renting and business services
Education
Health and social work
Other community, social and personal activities

- Survival (1) [4] - $C_2(2005) \rightarrow C_1(2006)$
- Merge (1) - $\{C_1(2005), C_3(2005)\} \overset{\subseteq}{\rightarrow} C_2(2006)$

In the latter, we observe:

- Survival (2) - $C_1(2006) \rightarrow C_2(2007)$ and $C_2(2006) \rightarrow C_1(2007)$

To better understand the meaning of the detected changes, and since we have access to whole data, we translated the transitions in terms of the principal activity sectors that were transferred from one cluster to another during each considered time interval.

Survival of cluster $C2(2005)$**:** In $[2005, 2006]$ cluster C2 was composed by only one activity sector (E - Production and distribution of electricity, gas and water), as can be verified in Table 5. This activity sector was then transferred to cluster C1 in year 2006, along with two "new" activity sectors (B - Fishing and C - Mining and quarrying). If we analyze the economic and financial indicators of this group we

[4] Numbers in parentheses indicate the number of occurrences of that specific transition

Table 7: Principal Activity Sectors assigned to each cluster returned by the agglomerative hierarchical algorithm for year 2007

Cluster 1
Fishing
Manufacturing
Wholesale and retail trade, repair of motor vehicles, motorcycles and personal goods
Hotels and restaurants (restaurants and similar)
Transport, storage and communications
Education
Other community, social and personal activities
Cluster 2
Mining and quarrying
Production and distribution of electricity, gas and water
Construction
Real estate, renting and business services
Health and social work

conclude that they encompass the activity sectors with better financial and economic performance in the Portuguese market place, in 2006.

Merge of clusters $C1(2005)$ **and** $C3(2005)$ **into cluster** $C2(2006)$**:** In 2006, activity sectors D - Manufacturing, G - Wholesale and retail trade, repair of motor vehicles, motorcycles and personal goods, H - Hotels and restaurants (restaurants and similar) and I - Transport, storage and communication, from cluster $C1(2005)$, merged with activity sectors F - Construction, K - Real estate, renting and business services, M - Education, N - Health and social work and O - Other community, social and personal activities, that come from cluster $C3(2005)$. This merge was motivated by the increase of the investment rate of these activity sectors which, in turn, was reflected in the increase of the debt ratio and in a remarkable improvement of the net income.

Survival of cluster $C1(2006)$**:** In 2006, cluster C1 was composed by activity sectors B, C and E. Sectors C and E are then transferred to cluster C2 in 2007. Only the activity sector B (Fishing) is assigned to a different cluster in the following year. This comes from the fact that activity sectors C and E have improved their economic and financial performance, contrary to sector B, whose indicators remained unchanged.

Survival of cluster $C2(2006)$**:** During $[2006, 2007]$, some activity sectors leave cluster $C2(2006)$, namely, sector F - Construction, K - Real estate, renting and business services and N - Health and social work, and a "new" sector joins this cluster (sector B - Fishing), which explains the composition of cluster $C1(2007)$. The allocation of sectors F, K and N to another cluster (cluster $C2(2007)$) is related to the improvement of the net income of these sectors.

In general, and bearing in mind the initial goal of this case study, we conclude that there were not remarkable changes in the economic and financial structure of the Portuguese activity sectors, along the considered years. However, we observe that there are some modifications in this structure during time interval $[2005, 2006]$, since there is a merge of clusters. In the following years (2006 and 2007), the main classification of the activity sectors becomes more interpretable and easy to understand, since the clustering algorithm divides the sectors with good economic and financial performance, from the sectors with medium and bad ones. The information acquired by this means can be used to perform a deeper economic analysis, once it uncovers and highlights the important issues of the studied problem.

4.2.1 Second Case Study - European Companies:

The second case study was designed using financial datasets, extracted from *The Data Page* of New York University - Leonard N. Stern School of Business, which contains financial information on individual

Table 8: Centroids of the clusters obtained in 2005 - standardized values

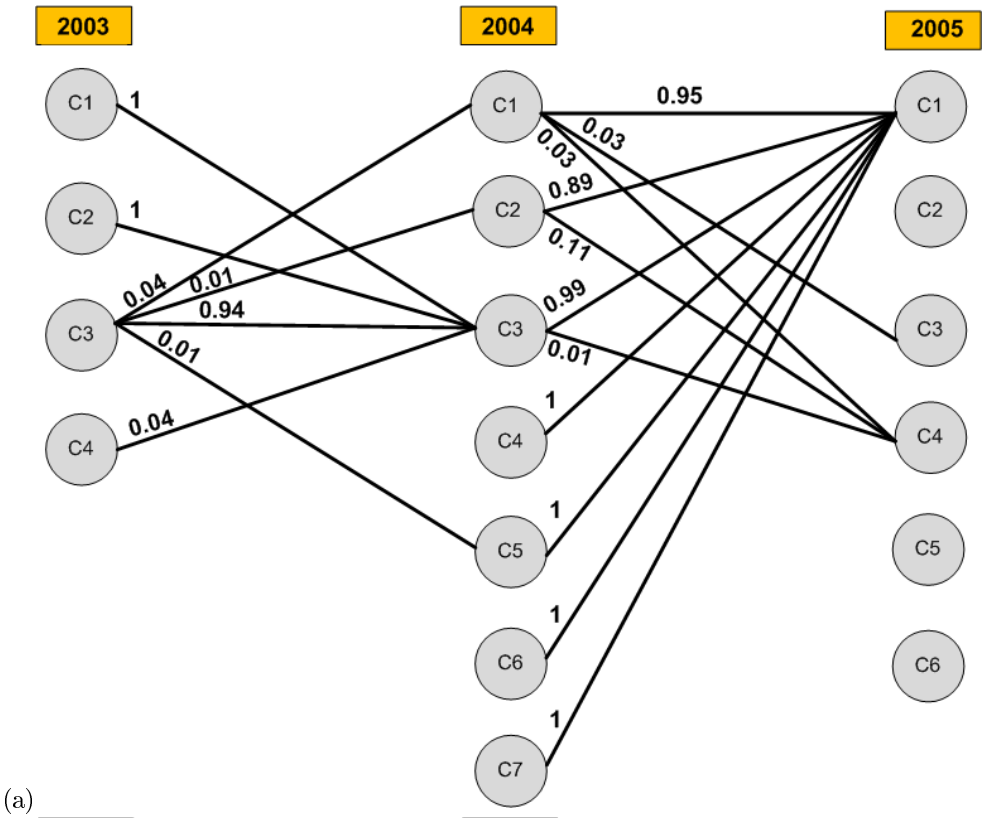
	Cluster 1 (C1)	**Cluster 2 (C2)**	**Cluster 3 (C3)**
Number of companies	0,3482	-0,6265	-0,2925
Net income	-0,2371	2,6952	-0,2546
Investment rate	-0,0742	2,4007	-0,3911
Return on equity	-0,7481	1,0385	0,6901
Net asset turnover	0,2659	-1,4118	-0,0367
Value added rates	-0,5858	0,3321	0,6365
Debt ratio	-0,6793	1,8095	0,4533
Equipment's productivity	-0,5174	-1,2935	0,8796
Labor productivity	-0,4257	2,8089	-0,0509

Table 9: Centroids of the clusters obtained in 2006 - standardized values

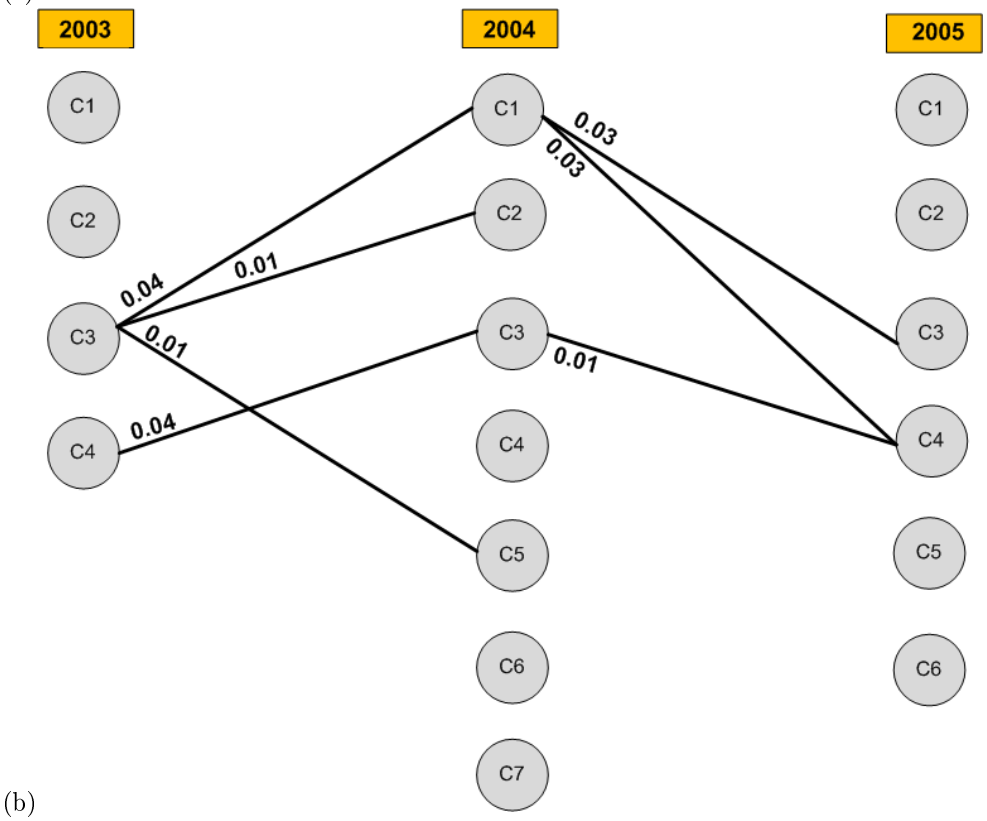
	Cluster 1 (C1)	**Cluster 2 (C2)**
Number of companies	-0,917	0,3057
Net income	1,106	-0,3687
Investment rate	1,3736	-0,4579
Return on equity	-0,3171	0,1057
Net asset turnover	-1,2571	0,419
Value added rates	0,2544	-0,0848
Debt ratio	0,4813	-0,1604
Equipment's productivity	-1,2418	0,4139
Labor productivity	1,2741	-0,4247

European companies. We chose datasets available for years 2003, 2004 and 2005. The data was preprocessed in order to remove missing values, remove outliers and select the best features to describe each company. After this preprocessing stage, each dataset consisted of 1071 observations, corresponding to European companies, and 9 continuous attributes (free cash flow to firm, firm value, invested capital, total debt, revenues, net capital expenditures, EBITDA, net income and tax rate), that describe the financial situation of these companies in each year. Similarly to the first case study, the objects and attributes are exactly the same for 2003, 2004 and 2005. Experimental conditions are also the same, with the exception that we use k-means algorithm, instead of the agglomerative hierarchical algorithm, to obtain the clusters. Nonetheless, the purpose of this analysis is quite different from the previous case study, since the main goal is to track the small groups of companies that don't follow the general trend. The idea is based in the fact that clusters, or groups, who exhibit the most different behavior may merit closer or further investigation. In general, the aim is to monitor cluster's behavior over time, find clusters (or smaller clusters within the main clusters) which evolution can be seen as anomalous, and focus on the analysis of its characteristics. The unusual behavior may not necessarily be related to fraudulent activities, still it can provide clues about which way to go in this kind of investigation or, rather, allow the discovery of interesting facts about the studied problem.

The datasets used in this case study seemed appropriate to test this approach, since there is more variety of links' weights in the bipartite graphs, compared to the previous case study, which increases the probability of finding rare events. In Figure 4-(a) are depicted the resulting bipartite graphs, with information of the weights in all links. As mentioned before, each graph's vertex, or node, represents a cluster, so we can verify that in 2003 there are four clusters, in 2004 there are seven clusters and in 2005 there are six clusters. In Figure 4-(b) we prune the links, removing the ones that are irrelevant for the rare events analysis, and we only keep the edges whose weights are considered low, since high weights are related to general trends. The concept of "low weight" is relative so we adopted the following process to define it: we computed the expected edge's weight as if the conditional probabilities for each cluster were equiprobable (e.g. for $[2003, 2004]$, and for each cluster of 2003, the expected edges' weight - EEW - is $\frac{1}{n_k(2004)=7} \approx 0.14$; for $[2004, 2005]$ the EEW is $\frac{1}{n_k(2005)=6} \approx 0.17$) and then we compare it with real edges' weights; if the real weight is equal, or lower, than $0.25 \times EEW$, we assume that the



(a)



(b)

Figure 4: Bipartite graphs for the financial datasets, corresponding to timestamps 2003, 2004 and 2005. In (a) are represented the bipartite graphs containing all links. In (b) are represented the bipartite graphs after being subjected to the pruning of the edges, according to the definition of low weights.

Table 10: Centroids of the clusters obtained in 2007 - standardized values

	Cluster 1 (C1)	**Cluster 2 (C2)**
Number of companies	-0,9215	0,3072
Net income	1,1024	-0,3675
Investment rate	1,1057	-0,3686
Return on equity	-0,3728	0,1243
Net asset turnover	-1,2386	0,4129
Value added rates	0,3611	-0,1204
Debt ratio	0,4355	-0,1452
Equipment's productivity	-1,2153	0,4051
Labor productivity	1,3285	-0,4428

weight is *low* (in this particular case, the threshold is $0.25 \times 0.14 \approx 0.04$, for time interval $[2003, 2004]$, and $0.25 \times 0.17 \approx 0.04$, for time interval $[2004, 2005]$) and, therefore, the corresponding edge requires a deeper analysis; otherwise, we ignore the link. The bipartite graphs' links that satisfy these conditions are highlighted in Figure 4-(b), and are the following:

Time interval $[2003, 2004]$:

- $\{C3(2003), C1(2004)\}$
- $\{C3(2003), C2(2004)\}$
- $\{C3(2003), C5(2004)\}$
- $\{C4(2003), C3(2004)\}$

Time interval $[2004, 2005]$:

- $\{C1(2004), C3(2005)\}$
- $\{C1(2004), C4(2005)\}$
- $\{C3(2004), C4(2005)\}$

After detecting these links we identified, for each time interval (or bipartite graph), the companies, and the corresponding industries, that migrate from one cluster to another. Then, we ordered the industries based on the frequency of appearance in these clusters and assumed that these industries are those who show rare behavior, i.e., who strongly deviate from the general trends of evolution. Following this procedure we discovered that some companies who operate in industries such as Transport Marine, Retail-Apparel/Shoe, Electric Integrated, Paper & Related products and Medical-Hospital, for time interval $[2003, 2004]$, Apparel Manufacturers, Brewery, Real Estate Management and Services and Finance-Investment Bnkr/Brkr, for time interval $[2004, 2005]$, exhibited special financial behavior in the considered time horizon, that may be related to possible merges or demerges of companies, bankrupts or the occurrence of other significant events. However, the reasons behind the nature of this behavior cannot be inferred based only on these datasets. It would be necessary more information, for example, provided by domain experts, newspapers or economic websites, to achieve more accurate conclusions about the distinct behavior of these industries.

5 Conclusions and Future Work

In this paper we introduced a framework for addressing the problem of monitoring the evolution of clusters' structures over time. The evolution is traced through the detection and categorization of clusters' transitions between snapshots of data. The process of detecting and characterizing these changes explores the sound concepts of conditional probabilities and bipartite graphs. We present real-world case studies using economic and financial datasets to illustrate the applicability and feasibility of our proposal. Each

case study seeks to illustrate the potential of the application of MEC in two types of analysis: one based on the general evolution trends of clusters and another based in the detection of rare events. This kind of analysis can be used as means of gaining insights about important occurred events, such as splits, merges, deaths and births that reflect the evolution of data, going beyond classical analysis techniques.

As future work we are extending and evaluating Equation 1 to fuzzy clusters' structures. A step ahead, is the development of similar methods to monitor the evolution of social networks, which we are starting now.

Acknowledgments.

Thanks to the support of the project Knowledge Discovery from Ubiquitous Data Streams (PTDC/EIA-EIA/098355/2008).

References

Aggarwal, C. C. (2003). A framework for diagnosing changes in evolving data streams. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, SIGMOD '03, pages 575–586, New York, NY, USA. ACM.

Aggarwal, C. C. (2005). On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, 17:587–600.

Albatineh, A. N., Niewiadomska-Bugaj, M., and Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, 23:301–313.

Asur, S., Parthasarathy, S., and Ucar, D. (2009). An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3:1–36.

Baron, S. and Spiliopoulou, M. (2001). Monitoring change in mining results. In Kambayashi, Y., W. W. A. M., editor, *Data Warehousing and Knowledge Discovery*, volume 2114 of *Lecture Notes in Computer Science*, pages 51–60. Springer Berlin / Heidelberg.

Baron, S. and Spiliopoulou, M. (2004). Monitoring the evolution of web usage patterns. In Berendt, B., Hotho, A., Mladenic, D., van Someren, M., Spiliopoulou, M., and Stumme, G., editors, *Web Mining: From Web to Semantic Web*, volume 3209 of *Lecture Notes in Computer Science*, pages 181–200. Springer Berlin / Heidelberg.

Bartolini, I., Ciaccia, P., Ntoutsi, I., Patella, M., and Theodoridis, Y. (2009). The panda framework for comparing patterns. *Data and Knowledge Engineering*, 68(2):244–260.

Berkhin, P. (2006). A survey of clustering data mining techniques. In Kogan, J., N. C. T. M., editor, *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg.

Bolshakova, N. and Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825–833.

Bottcher, M., Hoppner, F., and Spiliopoulou, M. (2008). On exploiting the power of time in data mining. *SIGKDD Explorations*, 10:3–11.

Cardoso, M. G. M. S. and Ponce de Leon Ferreira de Carvalho, A. (2009). Quality indices for (practical) clustering evaluation. *Intelligent Data Analysis*, 13(5):725–740.

Chawathe, S. S. and Garcia-Molina, H. (1997). Meaningful change detection in structured data. *SIGMOD Record*, 26:26–37.

Chen, K. and Liu, L. (2006). Detecting the change of clustering structure in categorical data streams. In Ghosh, J., Lambert, D., Skillicorn, D. B., and Srivastava, J., editors, *Proceedings of the 6th SIAM International Conference on Data Mining*, USA. SIAM.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3:301–313.

Elnekave, S., Last, M., and Maimon, O. (2007). In *Proceedings of the 23th International Conference on Data Engineering*, Data Engineering Workshops. IEEE Computer Society.

Falkowski, T., Bartelheimer, J., and Spiliopoulou, M. (2006). Mining and visualizing the evolution of subgroups in social networks. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 52–58, Washington, DC, USA. IEEE Computer Society.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). A framework for measuring changes in data characteristics. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '99, pages 126–137, New York, NY, USA. ACM.

Hruschka, E. R., Campello, R. J. G. B., and Castro, L. N. (2006). Evolving clusters in gene-expression data. *Information Sciences*, 176:1898–1927.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31:264–323.

Kalnis, P., Mamoulis, N., and Bakiras, S. (2005). On discovering moving clusters in spatio-temporal data. In Bauzer Medeiros, C., Egenhofer, M., and Bertino, E., editors, *Advances in Spatial and Temporal Databases*, volume 3633 of *Lecture Notes in Computer Science*, pages 364–381. Springer Berlin / Heidelberg.

Kaufman, L. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Kaur, S., Bhatnagar, V., Mehta, S., and Kapoor, S. (2009). Concept drift in unlabeled data stream. Technical report, University of Delhi.

Li, T., Ma, S., and Ogihara, M. (2004). Entropy-based criterion in categorical clustering. In *Proceedings of the 21th International Conference on Machine Learning*, ICML '04, pages 68–, New York, NY, USA. ACM.

O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., and Motwani, R. (2002). Streaming-data algorithms for high-quality clustering. In *Proceedings of the 18th International Conference on Data Engineering*, ICDE '02, pages 685–694.

Oliveira, M. and Gama, J. (2010). Bipartite graphs for monitoring clusters transitions. In Cohen, P., Adams, N., and Berthold, M., editors, *Advances in Intelligent Data Analysis IX*, volume 6065 of *Lecture Notes in Computer Science*, pages 114–124. Springer Berlin / Heidelberg.

Petrovic, S. (2006). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In *Proceedings of the 11th Nordic Workshop on Secure IT-systems*, NordSec 2006, pages 53–64.

Qiao, H. and Edwards, B. (2009). A data clustering tool with cluster validity indices. In *International Conference on Computing, Engineering and Information*, ICC '09, pages 303–309.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Spiliopoulou, M., Ntoutsi, I., and Theodoridis, Y. (2009). Tracing cluster transitions for different cluster types. *Control and Cybernetics*, 38:239–259.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., and Schult, R. (2006). Monic: Modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 706–711, New York, NY, USA. ACM.

Spinosa, E. J., Ponce de Leon Ferreira de Carvalho, A., and Gama, J. (2007). Olindda: a cluster-based approach for detecting novelty and concept drift in data streams. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, SAC '07, pages 448–452, New York, NY, USA. ACM.

Urga, G. (1992). The econometrics of panel data: A selective introduction. Economics series working papers, University of Oxford, Department of Economics.

Yang, H., Parthasarathy, S., and Mehta, S. (2005). A generalized framework for mining spatio-temporal patterns in scientific data. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 716–721, New York, NY, USA. ACM.