Abstract—Occupational fraud affects many companies causing them economic loss and liability issues towards their clients and other entities. Detecting internal fraud requires significant effort since a huge amount of data produced by diverse systems (which are mostly in textual form) has to be processed with little automated support. In this paper, we exploit the advantages of information visualization and present a system that aims to detect occupational fraud in systems which involve a pair of entities (e.g., an employee and a client). The main visualization is a spiral on which the events are drawn according to their time-stamp. Suspicious events are considered those which appear along the same radius or on close radii. The system ranks both entities according to the specifications of the auditor and a video file of their activity is generated such that events with strong evidence of fraud appear first. The system is equipped with several visualizations that facilitate the detection procedure.

I. INTRODUCTION

Occupational fraud represents a serious threat for companies worldwide that causes severe damages to the operation of a company (economic loss and liability issues towards clients, employees, and other entities such as financial institutions, etc.). In this paper, we present a system that uses visualization techniques to detect periodic patterns that may conceal occupational fraud. The proposed system was developed based on feedback provided by internal auditors of a Greek company and visualizes serial data produced by business control systems in which a pair of entities (e.g., employee-client) is involved.

Examining occupational fraud schemes in such systems reveals that events occurring in regular time basis may be indications of fraud e.g., in a billing system, a periodic monthly activity from an employee towards an account of a client is suspicious since it may suggest that the employee possibly falsifies client’s invoices (especially if it occurs before the billing date of the invoice). The main visualization of the system (refer to Fig. 1) is a spiral on which the time-stamp of each event is appropriately represented. Spiral visualizations facilitate the identification of potential periodic patterns (periodic events appear along a radius or on close radii). The system consists of several coordinated views, each dedicated to a particular aspect of data. Its innovation is that it aggregates the activity of employees and clients and ranks both of them according to the specifications of the auditor. Based on this ranking the system produces a video file, in which frames are ordered such that those with strong evidence of fraud appear first.

Much research effort has been focused on fraud detection and several approaches have been proposed including data-mining techniques [1]–[4], pattern matching and graph-pattern matching [5]–[11] and visualization approaches [12]–[16]. Regarding the detection of periodic patterns, spiral visualizations or systems of concentric circles have been used in [17]–[21]. Due to space constraints we will concentrate our description on the main visualization of the system. For a detailed description of related work, the ranking procedure and the supplementary visualizations the reader is referred to [22].

II. OVERVIEW OF THE DETECTION PROCEDURE

The input data consist of log files or database records generated by systems which involve pairs of entities. Each record may have been generated by a call between an employee and a client, a transaction involving both entities, etc. Hence, a record consists of a time-stamp, an employee, a client and an action taken by the employee. Since the log files are generated by different control systems with diverse log mechanisms, the data are appropriately parsed and stored in a database.

The ranking of a client is performed based on the following factors: (i) the number of events close to client’s billing date, (ii) the periodicity of the corresponding event-series, (iii) the events occurring outside working hours, (iv) the number of related employees, (v) the actions taken by the employees and, (vi) client’s current status. For each factor, we define three classes of clients according to the severity (low, medium, high) of the corresponding event-series. The performance of a client on a factor, say f, equals to (i) 2, if the event-series belongs to High Severity Class for factor f, or (ii) 1, if the event-series belongs to Medium Severity Class for factor f, or (iii) 0, if the event-series belongs to Low Severity Class for factor f. The auditor specifies an ordering among the factors that best fits to what she is seeking for. Hence, each factor is assigned a weight of importance based on a formula proposed by Stillwell et al. [23]. For instance, if she is interested in events that occur outside working hours, the corresponding factor should be ranked first, and thus, its weight would be greater than the weights of the other factors. Regarding the employee ranking, by default, the system assigns to the employee the value that corresponds to the maximum ranking of the clients related to the specific employee.

III. SYSTEM DESCRIPTION

The system operates in two modes, either off-line or semi-online. In the off-line mode, static data concerning a period of time are parsed and the corresponding visualizations are generated. The semi-online mode can be used on a daily basis to visualize the daily activity. Visualizations of large data-sets may not be useful in certain cases. To cope with this problem, the auditor is able to specify a time-window and then, the system visualizes events whose time-stamp belongs to the query window. However, ranking can be estimated either on the whole data-set or on data occurring in the specific time window, according to the specifications of the auditor.
The system consists of multiple coordinated views, each dedicated to a different aspect of the audit-data. A snapshot of the system in off-line mode is illustrated in Fig. 1. In the spiral visualization, each spiral branch visualizes a period of one month, while the number of branches is related to the first and last time-stamp of the input data (if not alternatively selected by the auditor), starting from the first month that coincides with the inner branch of the spiral. Each spiral branch is split by a number of lines according to the periodicity value that is examined (e.g., 30 days) and each line corresponds to a day of a month. The default value is 30 days, which implies that the administrator seeks for monthly suspicious activity.

In the spiral we place nodes, where each node represents an event related to a client and its position is determined based on the corresponding time-stamp. Nodes of different colors represent events related to different clients. Also, nodes of different shape correspond to different systems. In Fig. 1, two different systems are visualized using rectangular and circular nodes. To produce the spiral visualization, the system ignores multiple appearances of events that correspond to the same pair of employee-client at the same date. According to the spiral structure, events related to the same client and appearing along a radius of a spiral are considered suspicious and need to be further examined. However, examining cases of fraud has revealed that suspicious events may not always appear on the same date from month to month, and thus, they may appear on close radii.

Employees and clients are ranked according to the specifications of the auditor and the system generates a video file in which each frame depicts the activity of a client within the specified time interval, giving priority to the ones with the higher ranking. The ordering of the frames guarantees that clients that are considered to be suspicious will not be skipped during processing (even in cases of large data sets). Optionally, the nodes of the visualization can be distributed on the frames according to a predefined ordering specified by the auditor (based on predefined knowledge about a client or on a list of marked clients by a previous investigation). The video can also be paused in order to further investigate the activity of a client. By default, on each frame the billing and the due date of the corresponding client are depicted on the visualization (see Fig. 1). The light-gray colored region the main visualization corresponds to the “dangerous” interval of a week before the billing date, while the red-colored region illustrate events from month to month that differ by less than 3 days. The ranking factors can also be used as filters while the results can be exported in separate log files (e.g., from a single frame, the auditor can select only the nodes representing events occurred outside the employee’s working hours). The auditor can also perform custom queries to the database. Optionally, a produced visualization can be saved in a file for post-processing. The system also maintains records about the employees/clients activities and their ranking.

The main visualization of the semi-online mode is again a spiral and all visualization features are similar to the ones of the off-line mode. In contrast to the off-line, in semi-online mode, the inner branch of the spiral corresponds to the events of current day. The other branches coincide to the first and last month of the input data. The spiral is split by lines representing the days of a month. To produce the visualization, the system re-ranks both the clients and employees based on the whole data-set (i.e., including both previous and new records) and defines their ordering in the video. In semi-online mode, each client related to an event is visualized according to its time-stamp on the inner branch along with all its previous related events. Again, the auditor seeks for events that appear along a radius or on close radii of the spiral. This visualization aims to detect of a periodic activity that may have just begun.

**IV. CASE STUDY - FUTURE WORK**

In this section, we present a case study on real data from a major Greek company (the company name cannot be revealed due to a nondisclosure agreement). For confidentiality reasons, the data were preprocessed and made anonymous. The dataset corresponds to a time interval of six months and consists of approximately 35,000 entries involving 7200 distinct clients.

![Fig. 1. A snapshot of the interface of the system. Dates, usernames and actions are made anonymous for confidentiality reasons.](image-url)
and 14 employees. The case study was performed while the system operated in off-line mode and without taking into consideration any prior client or employee ranking.

The main question raised by the auditors was to identify pairs of employee-client involved in more than 10 events during the last six months. To do so, we first ranked the clients based on the number of their events. The system identified 430 clients whose number of events is larger than 10 (about 6% of the number of clients in the initial dataset) and distributed them accordingly in the video frames. Even though this number is smaller w.r.t the total number of clients in the data-set, the investigation was still a hard task (due to the number of clients to be examined). Then, we performed a second ranking on those 430 clients based on two factors: (i) the number of actions that are highly unlikely to appear and are indications of possible fraudulent activity, and (ii) the number of distinct employees serving the client. Details on the configuration values of each factor are omitted due to a nondisclosure agreement. Also, since we were not communicated the information about the billing dates and the employees’ shifts we ignored these ranking factors.

The ranking procedure identified 52 out of 430 clients (about 0.7% of the number of clients) that were highly-ranked in the above factors and presented them in the first frames of the video. In the next frames, the system presented 62 clients (about 0.9% of the number of clients) that were medium-ranked in the above factors. The results were further investigated by the auditors to detect periodic patterns (daily or monthly). They also suggested to apply the periodicity factor to detect events that appear within a time interval of (i) 28 days and (ii) 5 days. For the first case, the system identified 3 out of 52 clients that were highly-ranked and 6 out of 62 clients that were medium-ranked. For the second case, 11 out of 52 highly-ranked clients and 12 out of 62 medium-ranked clients were identified. In the final step, the auditors used the supplementary visualizations and the log viewer of the system along with their experience and supplementary data that were not communicated to us to evaluate the severity of these events. It should be emphasized that the real-time investigation of the data performed by the auditors did not identify any of these reoccurring activity and the results were further investigated.

Our work is ongoing and opens several directions. More ranking factors have to be taken into consideration to produce a more accurate value. It would be better for the auditor to add custom factors and define the appropriate functions. Functionalities such as statistic analysis of the activity of each entity, plots, etc, may be useful.

Acknowledgements

The work of E. N. Argyriou has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heralceitus II. Investing in knowledge society through the European Social Fund.

References

[8] T. E. Senator, “Ongoing management and application of discovered knowledge in a large regulatory organization: a case study of the use and impact of nasd regulation’s advanced detection system (rads),” in Proc. 6th ACM Int. Conf. on Knowledge discovery and data mining.