Visual Exploration of Journal Entries to Detect Accounting Irregularities and Fraud

Andrada Tatu, Marco Schreyer, Jan Hagelauer, and Jixuan Wang

Abstract—Nowadays, companies and organizations register millions of accounting transactions each year. Although most of these journal entries are legitimate, auditors face legal and financial obligations to discover transactions that are fraudulent. In this work, we present a visual analytics workflow to quickly identify unusual transactions in accounting data. In a first step features are derived from journal entries and are clustered to identify transactional patterns. In a second step the data is visualized to support the identification and investigation of unusual transactions. Following this workflow auditors are given the chance to identify new suspicious transactions that might correspond to fraud. We evaluated the proposed approach in a real world scenario by analyzing accounting data and discussed the results with auditors.

Index Terms—Visual Analytics, Forensic Accounting, Financial Statement Fraud, Journal Entry Testing

1 Introduction

According to a recent survey among over five thousand companies, 37% of the companies reported that their organization had experienced economic crime in the period 2011 - 2014 [1]. Accounting fraud received a significant response (22%), making it one of the most frequently reported types of fraud experienced. Similarly, the Association of Certified Fraud Examiners (ACFE) shows in its 2014 survey report that the annual median loss per company from financial statement fraud is more than $1 million [2].

The majority of audit techniques for detecting fraud operate at the financial statement level, a highly aggregated summary of a company's financial activity, e. g. by applying ratio analytics. In general, these techniques offer limited guidance to an auditor beyond a broad indicator of risk at a certain company. Therefore, there is an urgent need for analytics that operate at a much more detailed level, and more specifically on the transactions recorded in a company's general ledger. Such transactions are generally referred to as *journal entries* and exhibit information like account, amount, document type, debit or credit sign, etc.

At the same time, in today's digital age, companies usually register millions of journal entries each year in Enterprise Resource Planning (ERP) systems. Such systems provide a high degree of automation of the various business processes within an organization and are designed to process and record accounting relevant information.

- *A. Tatu, M. Schreyer and J. Hagelauer are with the PricewaterhouseCoopers AG WPG Forensic Services Group, Friedrichstrasse 14, 70104 Stuttgart, Germany (Contact: andrada.tatu@de.pwc.com).*
- *J. Wang is with Technical University (TU) Munich, Arcisstrasse 21, 80333 Munich, Germany.*

The exhaustive volume of journal entries makes it increasingly difficult for auditors to detect fraud and suspicious behavior. This holds in particular given the strict time and budget constraints faced at year-end audits. Applying advanced visual analytics [3] is one promising direction to overcome these challenges and to improve audit procedures.

In this work, we present a visual analytics approach to support the auditor's work in financial statement audit and to detect anomalous as well as potential fraudulent journal entries. To achieve this, we combine an (1) algorithmic and (2) visualization step to visualize and analyze a large number of journal entries.

In the algorithmic step, similar transactions are grouped and anomalous transactions in the data are identified. In the visualization step, the auditor is given a holistic perspective on all the transactions to be audited. This perspective empowers auditors to visually investigate different transaction patterns, trends and correlations that are hidden in the data. Furthermore, it allows to spot possible anomalous transactions that can be examined in more detail with process owners, or internal auditors.

The proposed approach is evaluated investigating anonymous transactions of a case company. The set of sample transactions covers all outgoing vendor payments recorded in the company's ERP system for two fiscal years. We believe that the case study results demonstrate the promising potential of visual analytics in the context of financial statement audits.

2 Related Work

2.1 ERP Data Forensics

The topic of fraud detection in financial accounting has received growing attention in research and

academia [4]. However, the forensic analysis of transactional records from ERP systems is relatively new. Recent publications show promising and interesting methods to detect fraudulent activities based on transactional data.

In [5] Kahn et al. generated transaction profiles of ERP system users based on an analysis of the transaction audit log for each user. Anomalous transactions profiles were identified and investigated in detail afterwards. Islam et al. [6] followed a different approach by defining fraud scenarios and their corresponding audit log signatures.

Bay et al. [7] defined a set of discriminative features to model fraudulent general ledger account behavior based on ERP transactions. Afterwards, classifiers were trained to identify suspicious accounts. McGlohon et al. [8] enhanced this approach by modeling the interrelation of general ledger accounts utilizing network propagation techniques.

In [9], business process mining as well as uni- and multivariate latent clustering algorithms are applied to event logs of ERP purchase orders obtained from an SAP system of a financial institution. The detected process anomalies were audited afterwards in a joint effort with internal audit experts.

2.2 Categorical Data Visualization and Clustering

Categorical data needs particular attention when visualized or analyzed automatically using clustering techniques. Therefore, a series of specific techniques have been developed for this type of data.

A commonly used approach is to compute frequencies of each category and represent the categories by a visual entity scaled according to their frequency. Mosaic Displays [10] use such frequency-scaled rectangles to recursively divide the space and display the attributes. Hofmann [11] extends this technique to interactive mosaic plots and includes domain knowledge for a better visual exploration. Bargrams [12] display each attribute in a row and their categories with rectangles scaled by their frequency. All of these methods however lack the ability to compare multidimensional relations between attributes and do not scale well with a higher number of dimensions.

Parallel Sets [13] are a well known Parallel Coordinates [14] technique for categorical data. Each attribute is plotted as a separate axis which is subdivided into sections for each of the attribute values. Parallelograms connect the axes to show the relations between different attributes. Color can be used to differentiate the attribute values of one axis facilitating the tracing of their distribution among other attributes. Interaction is provided to better understand the data, as we will show later in Section 4.

Due to the size of our data and the relationships among multiple attributes that are relevant for our task, we decided to use Parallel Sets for our approach.

Clustering categorical data also brings up the need for specialized algorithms compared to numerical data. One distinction is the definition of similarity among objects. A complete survey of currently available techniques can be found in [15]. One prominent approach is K-modes [16], a modification of the K-means paradigm. To deal with categorical data, K-modes uses the simple-matching dissimilarity measure. Agglomerative hierarchical clustering [17] groups a dataset by creating a cluster tree. For categorical data, the similarity is adapted, either by transforming the categorical variables into bit vectors, or simple matching.

The ROCK [18] algorithm, also a agglomerative hierarchical clustering algorithm, is specifically developed to cluster a dataset with solely categorical attributes. It introduces the concept of a link, defined as the number of common neighbors between two objects and uses this to form clusters.

3 Proposed Analytical Workflow

We propose a visual analytics approach to support the exploration of journal entries making use of algorithmic grouping in combination with visual-interactive representations for user-based filtering and exploration (see Figure 1).

Fig. 1: Analytical workflow: ERP data is (1) visually analyzed as a whole, (2) grouped, and (3) visualized to interactively support auditors to analyze and explore anomalies, possible fraudulent activities.

As a first step the ERP data is visualized by creating a Parallel Sets [13] view. We believe this technique is one of the most promising methods currently available to represent multidimensional relations among categorical data. By interacting with the visualization the auditor is given the chance to gain an initial understanding of the data and derive an early hypothesis. Afterwards, categorical clustering techniques are applied to identify groups that correspond to typical accounting patterns. We observed that agglomerative hierarchical clustering techniques using single linkage are best suited for this analytical task. Finally, the interactive visualization is enriched by the clustering results. This can, for example, be achieved via adding an additional axis to the Parallel Sets view. Using the enriched interactive view the auditor can browse through the different clusters and verify or falsify the initial hypothesizes. Furthermore, new insights can be gained by linking the information across multiple attributes.

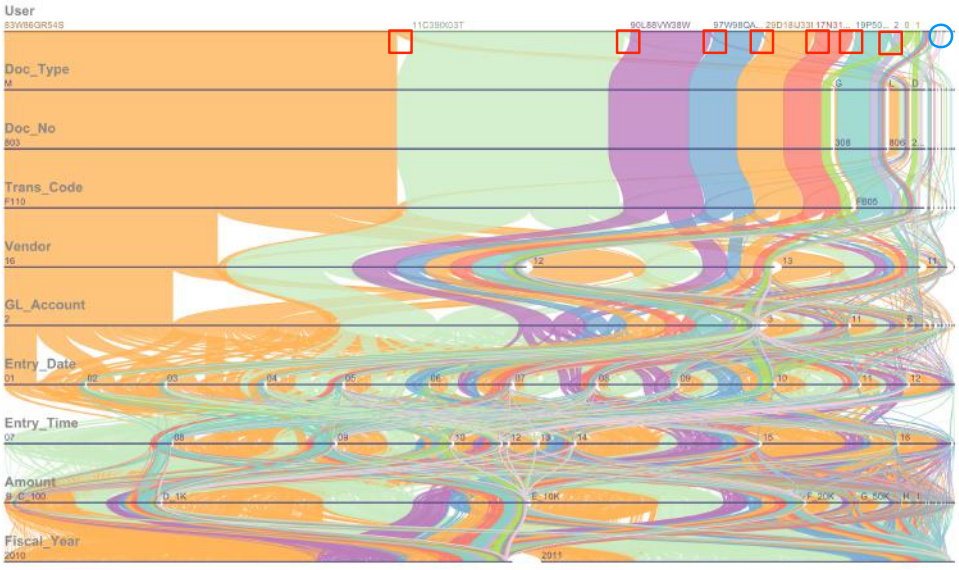

Fig. 2: Visual data overview with Parallel Sets, where each attribute is a horizontal axis and the color represents the journal entry creator (User). The width of each category instance gives the proportion of occurrences of the category's value. *MBPs* are visible by wide connection areas, *SBPs* are marked with red rectangles, and ABTs with a blue circle.

4 Case Study: Vendor Payments

4.1 Analyzed Data

The dataset analyzed in this work consists of accounting data corresponding to outgoing vendor payments of an ERP system. These transactions are especially relevant for assessing fraud risks, as many fraud schemes involve illegitimate payments directly or indirectly benefiting the fraudster.

After the data preprocessing phase encompassing anonymization and binning, the final dataset consists of approximately 25 thousand transactions covering the most relevant characteristics: *amount, transaction code, fiscal year, document number (range), document type, user, G/L account, vendor (range), entry date (month),* and *entry time (hour).* Through binning, we were able to group attribute values based on domain knowledge and obtain a lower number of categories for each attribute.

4.2 Visual Exploration

Traditional audit methods involve an initial analytical step to reduce the high number of transactions. A small subset is then inspected manually in more detail and plausibilized. Our proposed visual analytics approach enables the auditor to gain a holistic understanding of the data, that leads him to a subset of transactions useful for in depth auditing.

Figure 2 represents a Parallel Sets view of all transactions. Each attribute is represented as a horizontal axis and the creator of the journal entry (the user) is distinguished by the different coloring. The width of each category instance gives the proportion of occurrences of the category's value.

Starting from this view, the auditor can perform an in-depth analysis of the entire dataset, to identify suspicious transactions that may require further investigation. Visual exploration helps in understanding (1) main business processes, (2) secondary business processes, and (3) anomalous transactions in the data. Automatic transaction grouping facilitates this task.

As **Main Business Processes** (MBPs), we describe processes that correspond to an usual business transaction. In the Parallel Sets visualization (Figure 2) this is visible by wide paths across multiple axes. In this case, two MBPs are visible: automatic vendor payments represented by transaction code (Trans_Code) *F110*, and manual payments represented by Trans_Code *FB05*. These MBPs are likely to be known by the auditor and will therefore not require in depth analyses.

We define **Secondary Business Processes** (SBPs) as somewhat unusual transactions which still occur too frequently than to be considered as anomalies. In Figure 2, we can see narrow paths across multiple axis (marked by red rectangles), e.g. cheque payments, cash payments, or vendor offsetting. These observations might be of interest for the auditor, since they reflect significant deviations from the main business process.

Anomalous Business Transactions (ABTs) are patterns that don't correspond to any MBPs or SBPs because of the limited number of occurrences of the specific attribute combinations. Such entries (marked in Fig. 2 with a blue circle) could lead the auditor to suspect anomalies. To investigate this further, auditors can apply (1) filtering, excluding the primary pro-

cesses that are not considered relevant for the auditing process, or (2) clustering to group MBPs and SBPs automatically.

4.3 Visual Analytics of Anomalous Transactions

Figure 3 represents the entries colored by the transaction amount ordered from high to low. The auditor could spot through interactive exploration two transactions which represent exceptionally high payments to the same vendor over a six month period. These are highlighted by dark purple in the view. Such unusual patterns can be picked up by the auditor to be discussed in more detail with the responsible process owners, internal auditors and the audit team.

In the next step, clustering is used to identify accounting irregularities. Our examples use the single linkage method. The first axis of Figure 4 represents the clustering result; labeled with the total transaction amount of all cluster entries. It is visible that the main business processes were grouped together in a large orange cluster with a high volume. After excluding this from the view to use the screen space more efficiently, the auditor can see (Figure 5) that the second large cluster (purple) is again covering most of the remaining data. Additionally, highlighted on the second axis, doc. type F reveals two smaller yellow clusters next to the main purple cluster. After switching to the equidistant mode (first two axes in Figure 6), these small clusters can be analyzed in more detail. It is suspicious that the bright yellow cluster shows two entries with round amounts. One can also see that five of the eleven two- and one-entry clusters show round amounts. Auditors can follow up such suspicious observations in detail.

5 Discussion and Conclusion

As fraudsters are getting more creative, analysts also need to embrace more creative ways to identify new fraud patterns.

Our approach exemplifies how visual analytics of financial accounting data can enable internal and external auditors to efficiently assess the plausibility of material subsets of the transactions. We have demonstrated that visual analytics can help to develop a more attractive and powerful way of analyzing journal entries beyond the use of spreadsheets. This new way can lead to building a mental model about the data to be analyzed, support the analyst to conquer the difficult, massive data, and provide a new way to detect accounting anomalies and fraud.

As next steps, we envision to develop a special investigation application to support fraud investigation in ERP data with linked views, providing different perspectives on the transactional patterns. These may enable the auditor to detect even more complex fraud patterns in the future.

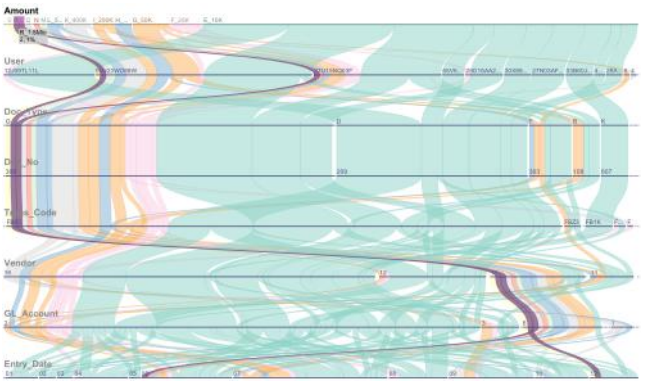

Fig. 3: Two entries representing manual payments to one vendor exhibiting a high amount (highlighted in dark purple).

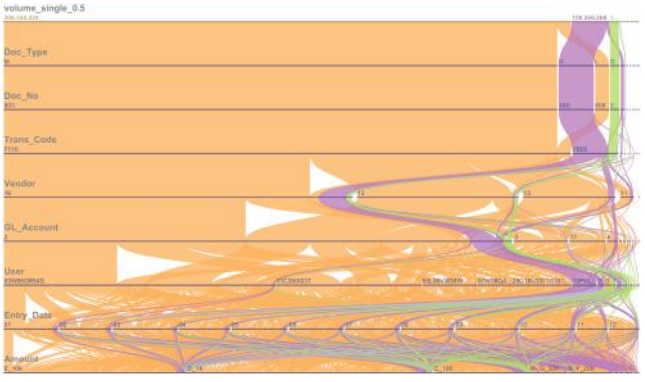

Fig. 4: Enriched view by colored clusters as main axis.

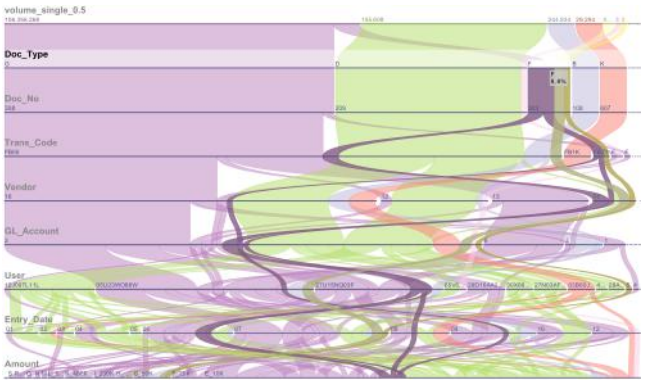

Fig. 5: Largest cluster (automatic payments) filtered out. Doc. type with value F reveals two anomalies (highlighted in yellow and dark yellow).

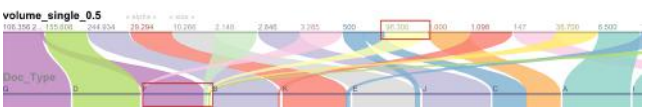

Fig. 6: Entries with round amounts are separated from main purple cluster, visible in yellow.

ACKNOWLEDGMENTS

The authors would like to thank to Robert Kosara for providing Parallel Sets open source for other analysts.

REFERENCES

[1] PricewaterhouseCoopers International Limited, "Economic crime: A threat to business globally," *PwC's 2014 Global Economic Crime Survey*, 2014.

[2] Association of Certified Fraud Examiners (ACFE), "Report to the nations on occupational fraud and abuse," *2014 Global Fraud Study*, 2014.

[3] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual Analytics: Scope and Challenges," in *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Springer, 2008, lNCS.

[4] A. Sharma and P. K. Panigrahi, "A review of financial accounting fraud detection based on data mining techniques," *Int. Journal of Computer Applications*, vol. 39, no. 1, pp. 37–47, 2012.

[5] R. Q. Khan, M. W. Corney, A. J. Clark, and G. M. Mohay, "Transaction mining for fraud detection in erp systems," *Industrial Engineering and Management Systems*, vol. 9, no. 2, pp. 141–156, 2010.

[6] A. Islam, M. Corney, G. Mohay, A. Clark, S. Bracher, T. Raub, and U. Flegel, "Detecting collusive fraud in enterprise resource planning systems," in *Advances in Digital Forensics VII*, ser. IFIP Advances in Information and Communication Technology, G. Peterson and S. Shenoi, Eds. Springer, 2011, vol. 361, pp. 143–153.

[7] S. Bay, K. Kumaraswamy, M. G. Anderle, R. Kumar, and D. M. Steier, "Large scale detection of irregularities in accounting data," in *6th Int. Conf. on Data Mining*. IEEE, 2006, pp. 75–86.

[8] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos, "Snare: a link analytic system for graph labeling and risk detection," in *Proc. of the 15th ACM SIGKDD Int. Conf. on KDD*. ACM, 2009, pp. 1265–1274.

[9] M. Jans, N. Lybaert, and K. Vanhoof, "A framework for internal fraud risk reduction at it integrating business processes: the ifr^2 framework," *The Int. journal of digital accounting research*, vol. 9, p. 7, 2010.

[10] J. A. Hartigan and B. Kleiner, "Mosaics for contingency tables." *Computer Science and Statistics: Proc. of the 13th Symp. on the Interface*, pp. 268—273, 1981.

[11] H. Hofmann, "Exploring categorical data: interactive mosaic plots," *Metrika*, vol. 51, no. 1, pp. 11–26, 2000.

[12] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton, "Parallel bargrams for consumer-based information exploration and choice," in *Proc. of the 14th Annual ACM Symp. on User Interface Software and Technology*. ACM, 2001, pp. 51–60.

[13] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *Transactions on Visualization and Computer Graphics*, vol. 12, pp. 558–568, 2006.

[14] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proc. of the 1st Conf. on Visualization '90*. IEEE CS Press, 1990, pp. 361–378.

[15] S. A. Elavarasi and J. Akilandeswari, "Survey on clustering algorithm and similarity measure for categorical data," *ICTACT Journal on Soft Computing*, vol. 04, p. 8, 2014.

[16] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[17] J. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[18] S. G. Rajeev, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," in *Information Systems*, 1999, pp. 512–521.